Fixed Effects Topic Model

Dan Biderman[1], David Blei[2], Wei Cai[2], Andrea Ciccarone[3], Amir Feder[3], and Andrea Prat[2]

[1]Stanford University
[2]Columbia University
[3]Hebrew University

This version: October 13, 2025

Abstract

Social scientists wish to perform topic modeling on documents that are created by different authors in different contexts. However, the same broad topic may be expressed in different ways depending on the environment where the author operates. For example, one may wish to use employee reviews to identify broad corporate culture topics, but the language of reviews is influenced by industry-specific jargon. Existing methods attempt to control for these biases ex post, such as with traditional fixed effect regressions. But these methods cannot fully separate global themes from category-specific language within them. In this paper, we introduce the Fixed Effects Topic Model (FETM), a novel approach to disentangling broad topics from contextual influences by incorporating fixed effects directly into the generative process of language. We use the FETM to identify themes in a large corpus of Glassdoor job reviews. We show that it outperforms conventional topic models, both in interpretability and predictive accuracy.

1 Introduction

In many areas of the social sciences, unsupervised analysis of large corpora has become a standard tool and it is yielding important insights (see eg. Grimmer and Stewart, 2013 and Gentzkow et al., 2019). Topic models, in particular, provide a way to uncover latent thematic structure in text that is not directly observable, and have been applied to study political agendas, legal briefs, managerial behavior, and central bank deliberations (Quinn et al., 2010; Sim et al., 2015; Hansen

Table 1: Example of Employee Reviews

Company	Glassdoor Review
Amazon	*"Ran like the military, very cutthroat environment. You are given rates you need to pick and pack that are very hard to make if you don't have the work in front of you. Thirty minute lunches and 15 minute breaks however you must walk about 5–10 minutes to get your lunch (if it's not stolen) so even that you don't look forward to."*
Aflac	*"Being on the phone all day is kind of like jail. Pre-scheduled breaks and lunch times. An adult should be able to choose when they eat lunch. Sometimes they scheduled your lunch after you'd only been at your shift for an hour or two. Then you are stuck at your desk the rest of the day. Getting time off can be a hassle. It's determined by a computer, not your supervisor. Very little to no advancement opportunities."*

et al., 2018; Bandiera et al., 2020). By transforming unstructured documents into interpretable dimensions of variation, they allow researchers to recover constructs—such as ideologies, narratives, or organizational practices—that are otherwise difficult to measure at scale.

Because of these advantages, one area where topic modeling may be particularly useful – and will be explored later in this paper – is the study of corporate culture. For decades, both practitioners and scholars have noted that different organizations can have very different cultures (Schein 1985) and that these differences may explain why some firms perform much better than other firms even controlling for observable inputs (Syverson 2004, Gibbons-Henderson 2013). Scholars have identified some potential typologies of corporate cultures. For instance, Kotter and Heskett (2011) distinguish between Clan, Adhocracy, Market, and Hierarchy, while Handy proposes Power Culture, Role Culture, Task Culture, Person Culture. To identify different types of cultures we could turn to corpora that cover most medium-large firms in the US and other market economies, like employee reviews websites. However, we face a methodological challenge.

Corpora that cover multiple organizations contain text that is written by a myriad of individual contributors. These contributors describe the culture of the organization they work for, but they do using their words, which tend to be affected by their environment. Consider for instance the two employee reviews in Table 1: the first comes from a Warehouse Associate at Amazon, while the second comes from a Costumer Service Specialist at the insurance company Aflac. Both reviews appear to be addressing the same broad topic, rigidity of the workplace schedule, and are expressing a similar sentiment: frustration with the heavy regulation of the work routine. They even use similar metaphors. However, they do so with different words and expressions that are typical of the industry they operate in.

A standard topic model, taking as input the raw reviews, may struggle to recognize that the

Table 2: Comparison of Topic Distributions Across Models

Review	Model	Topic Name	Intensity
Amazon	**VTM**	Industry Products, Safety, & Communication	0.24
		Directors & Negative Management Experiences	0.15
		Flexible Environment, Client Work & Daily Operations	0.08
		Scheduling, Family/Bonus Concerns & Process/Metric Issues	0.08
	FETM	Workplace Conditions, Lunch/Breaks & Marketing/Commission	0.38
		Collaboration Across Teams & Organizational Focus	0.12
		Managerial Oversight, Performance Reviews & Trust in Leadership	0.12
		Customer-Facing Roles, Hiring Practices & Employee Treatment	0.08
Aflac	**VTM**	Stress, Turnover, & Performance/Training Challenges	0.16
		Worker Conditions, Shift Work, & Insurance/Tech	0.12
		Directors & Negative Management Experiences	0.08
		High-Pressure, Tech/Services & Employment Processes	0.07
	FETM	Workplace Conditions, Lunch/Breaks & Marketing/Commission	0.29
		Customer-Facing Roles, Hiring Practices & Employee Treatment	0.14
		Career Advancement, Professional Development & Technology	0.10
		Learning Culture, Diversity, & Development Initiatives	0.07

two reviews address the same fundamental issue. Because the model does not take into account that language was generated in different contexts, the estimated topics may be characterized by industry or occupation-specific jargon. In our example, phrases like "given rates" and "pick and pack" are strongly associated with warehouse work, while terms like "phone", "computer" and "desk" are more common in office or call-center settings. As a result, the estimated topic intensities on the reviews may focus on the environment-specific language rather than the universal topic of workplace rigidity.

Table 2 illustrates this phenomenon. A vanilla topic model (VTM) fails to assign the two reviews to the same broad topic . Instead, it tends to capture both the industry-specific components (industry product and safety for Amazon, stress and insurance/tech for Aflac) as well as the sentiment-specific component (negative management).

The problem can be posed in more general terms. Suppose we face a multi-source corpus (text data produced by multiple sources). We believe there are some universal themes and we wish to identify them. However, suppose that those sources operate in different environments and the language used by those sources is affected by those environments. A standard topic model approach may struggle to identify the universal topics because they will be expressed in different ways in different environments, something we refer to as an environmental language fixed effect.

The objective of this paper is to introduce a methodology to address environmental language

fixed effects in multi-source corpora. We propose a novel Fixed Effects Topic Model (FETM) designed to disentangle environment specific shifts from the underlying, general managerial topics. By introducing fixed effects into the generative process, our model allows us to estimate latent managerial themes, while adjusting for factors such as industry, review sentiment, or any other environment related attributes. The key advantage is that these fixed effects capture vocabulary and context unique to a given environment, ensuring that general managerial styles emerge more cleanly. As shown in Table 2, FETM accounts for the environmental language fixed effects and identifies, in both reviews, the broad topic associated to workplace conditions and schedule.

We apply this model to a large dataset of Glassdoor job reviews. Our approach automatically identifies latent managerial types in an unsupervised manner, avoiding the need for strong prior assumptions or labeled training data. Methodologically, we demonstrate how incorporating fixed effects in a topic model can improve both interpretability and predictive performance. Substantively, we show how this helps uncover the deeper, cross-industry dimensions of managerial style, a crucial step for understanding how managers shape organizational outcomes and how these styles might vary within and across different contexts.

In a synthetic experiment where we artificially introduce a known effect to test causal inference, the FETM recovers the true effect more accurately, underscoring its robustness and reliability. These exercises demonstrate that estimating topics without controlling for fixed effects can lead to biased inferences, whenever those fixed effects are correlated with the outcome of interest.

The paper is structured as follows. Section 2 develops the theoretical foundation and technical formulation of the FETM. Section 3 describes our empirical application, including data sources, sample construction, and estimation details. Section 4 presents our main empirical findings, comparing of the performance and characteristics of FETM to those of a "vanilla" topic model (VTM). Section 5 extends the analysis by testing FETM on additional datasets. Section 6 concludes.

2 Fixed Effects Topic Model

2.1 The "Fixed-Effects Problem"

When using topic proportions for causal analysis, one of the most popular applications of topic analysis in social science, there may be a significant issue related to how the topic is estimated according to different categories (e.g., industries or sentiment labels). As a result, the estimated topic proportions themselves can be biased.

Take, for instance, the "Workplace Conditions" industry reviews on the "Salary/Benefits" topic. If there are sufficiently many reviews within this industry, a "naive" topic model will likely pick up industry-specific language as a separate topic. If we also consider negative versus positive reviews, this issue becomes even more pronounced: the model may create two different topics for what is

actually the same broad theme. In cases where our topic proportion measure needs to capture the "what" rather than the "how", this is evidently problematic (e.g., we want to know what employees are discussing when reviewing their company, not the rhetorical style or worse the industry specific language).

When using this biased topic proportion as a regressor for causal inference, simply adding fixed effects in the regression specification will not fix the underlying bias in the topic estimates. That is, by including fixed effects, one cannot fully correct ex-post for the topic model potentially lumping specific and global language together, or splitting similar content into different categories.

Formally, denote the true topic proportions as θ_i. We also encode the categories (e.g., the industry of the review or whether it is pro/negative) in the indicator vector $\mathbf{x}_i$. A naive topic model trained on the pooled data may produce an estimated topic proportion

$$\widehat{\theta}_i \;=\; f\big(\theta_i, \mathbf{x}_i\big),$$

where the function f reflects how or category-specific language distorts the naive topic estimate.

Suppose we then use $\widehat{\theta}_i$ in a downstream regression to explain some outcome y_i, including fixed effects for the categories:

$$y_i \;=\; \alpha \;+\; \gamma\,\widehat{\theta}_i \;+\; \delta^\top \mathbf{x}_i \;+\; \varepsilon_i.$$

Substituting $\widehat{\theta}_i = f(\theta_i, \mathbf{x}_i)$ directly into the regression, we obtain:

$$y_i \;=\; \alpha \;+\; \gamma\, f(\theta_i, \mathbf{x}_i) \;+\; \delta^\top \mathbf{x}_i \;+\; \varepsilon_i.$$

Because the estimated topic proportions, $\widehat{\theta}_i \;=\; f(\theta_i, \mathbf{x}_i)$, may incorporate category-specific idiosyncrasies, including fixed effects for the category indicators, $\mathbf{x}_i$, in a regression framework does not necessarily separate global content from environment-specific variation. To formalize this, consider the following stylized representation:

$$\widehat{\theta}_i = \theta_i + g(\mathbf{x}_i) + \nu_i,$$

where θ_i represents the true topic proportions, $g(\mathbf{x}_i)$ captures systematic distortions due to environment-specific language, and ν_i is an error term.

Suppose we are interested in estimating the causal effect of the true topic proportions, θ_i, on an outcome y_i, using the regression:

$$y_i = \alpha + \gamma\widehat{\theta}_i + \delta^\top \mathbf{x}_i + \varepsilon_i,$$

where $\delta^\top \mathbf{x}_i$ represents the fixed effects for the categories. Substituting $\widehat{\theta}_i$ into the regression yields:

$$y_i = \alpha + \gamma\big(\theta_i + g(\mathbf{x}_i) + \nu_i\big) + \delta^\top \mathbf{x}_i + \varepsilon_i.$$

Rewriting, we have:

$$y_i = \alpha + \gamma\theta_i + \gamma g(\mathbf{x}_i) + \gamma\nu_i + \delta^\top \mathbf{x}_i + \varepsilon_i.$$

The term $\gamma g(\mathbf{x}_i)$ reflects the bias introduced by the category-specific distortions in the topic estimates. Because $g(\mathbf{x}_i)$ is correlated with the fixed effects $\mathbf{x}_i$, the inclusion of $\delta^\top \mathbf{x}_i$ cannot fully absorb the bias. Consequently, the coefficient γ on $\widehat{\theta}_i$ will remain biased. Such distortions need to be addressed during the topic modeling stage rather than relying solely on ex-post adjustments in the regression specification, which leads us to our Fixed Effects Topic Model.

2.2 Model Specification

Consider a corpus of n job reviews with the corresponding fixed-effects information represented as $\mathscr{D} = \{(\mathbf{w}_1, \mathbf{x}_1), \ldots, (\mathbf{w}_n, \mathbf{x}_n)\}$, where each document $\mathbf{w}_i$ is paired with its corresponding fixed-effects vector $\mathbf{x}_i$.

Each text document $\mathbf{w}_i$ is a sequence of m word tokens, given by $\mathbf{w}_i = \{w_{i1}, \ldots, w_{im}\}$, that come from a vocabulary $w_{ij} \in \mathbb{1}^{|V|}$.

The vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$encode indicator information for which categories apply to each document. In our main specification, each entry in $\mathbf{x}_i$ indicates whether the review belongs to a particular industry (according to the NAICS sector classification) or whether it is a "pro" or "con" review. If a review is from the Accommodation and Food Services sector, the corresponding entry in $\mathbf{x}_i$ would be 1, while entries for other industries would be 0. Similarly, if the review is a "pro" review, the fixed-effect vector would have the associated entry set to 1.

We let $\mathbf{x}_i \in \{0, 1\}^{|E|}$, where E is is the set of all possible categories a review might have (eg. pro/con, industry 1, industry 2 etc.). Thus, the dimension of E is the total number of categories for which an indicator can be 1 (e.g., with 2 review types - pro or con - and 2 industries, we would have $|E| = 4$). According to these categories, the actual fixed effects are then captured by the coefficients γ_k described below, which quantify how each category influences the topic-word distribution.

Our goal is to learn global topics while separating out their fixed-effect-specific adjustments. Recall our running example, where pros and cons reviews discuss the same topics in unique ways. We want to capture the unique ways these reviews discuss the same topic while simultaneously extracting common terms shared among all reviews.

Each document is represented as a mixture of topics, with a local latent variable θ_i denoting the per-document topic intensities. Topics are denoted by β, and each β_k is a probability distribution over the vocabulary, $\beta_k \in \mathbb{R}^v$. We introduce a new latent variable, $\gamma_k \in \mathbb{R}^{e \times v}$, where $k \in$

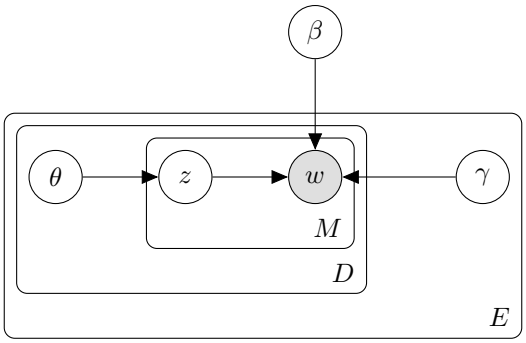


Figure 1: A graphical model for the fixed effect topic model (FETM). M denotes words in a document and D documents. E denotes the fixed-effect environments documents are drawn from (determined by different configurations of $\mathbf{x}$). z denotes topic assignment, β denotes global weights for each word in the vocabulary, and γ denotes environment-specific weights.

$\{1, \ldots, K\}$, that is designed to capture the fixed-effect adjustments, i.e., the actual impact of each category on the topic-word distribution β_k.

We assume the following generative process:

1. Draw $\beta_k \sim \mathcal{N}(\cdot, \cdot)$, $\beta_k \in \mathbb{R}^v$, for $k = 1, \ldots, K$.
2. Draw $\gamma_k \sim p(\gamma)$, $\gamma_k \in \mathbb{R}^{e \times v}$, for $k = 1, \ldots, K$.
3. For each document i:
   (a) Draw topic intensity $\theta_i \sim \mathcal{N}(\cdot, \cdot)$.
   (b) For each word j:
      i. Choose a topic assignment $z_{ij} \sim \text{Cat}(\pi(\theta_i))$.
      ii. Choose a word $w_{ij} \sim \text{Cat}(\pi(\beta_{z_{ij}} + \gamma_{z_{ij}} \cdot x_i))$.

The graphical model for this fixed-effects topic model is represented in Figure 1. Given data, the posterior finds the topic and word distributions that best explain the corpus overall (β_k) and also the distribution of how words shift when particular categories (i.e., fixed effects) are present (γ_k). FETM represents the topics shared by all reviews in β_k, while capturing the particular ways pros, cons, and specific industries talk about those topics via γ_k.

We next specify $p(\gamma)$ using the automatic relevance determination (ARD) prior.

Automatic Relevance Determination (ARD) and Empirical Bayes. FETM is built with the additional assumption that documents are generated based on different configurations of observed categories. The goal is to separate global topics from fixed-effect-specific adjustments on topics. To

do this, we introduce a new latent variable, γ_k. We further posit that these fixed-effect adjustments on the global topic-word distribution β_k should be sparse. Consider again our running example: nearly all words will be shared across job reviews, so we want to ensure γ_k only places high density on terms that are truly shifted for a particular category (e.g., "pro" or "industry 2").

In many real-world tasks, the input data contains a large number of irrelevant features. ARD is a method used to filter them out (MacKay, 1992; Tipping, 2001). Its basis is to assign independent Gaussian priors to the feature weights. Given the feature weights η, the ARD assigns priors as:

$$\sigma_c \sim \text{Gamma}(a, b), \tag{1}$$

$$p(\eta|\alpha) = \prod_c \mathcal{N}(\eta_c \mid 0, \alpha_c^{-1}). \tag{2}$$

The precisions, $\alpha = \{\alpha_c\}$, represent a vector of hyperparameters. Each hyperparameter α_i controls how far its corresponding weight η_c is allowed to deviate from zero. Rather than fixing them a priori, ARD hyperparameters are learned from the data by maximizing the likelihood of the data with empirical Bayes (Carlin and Louis, 2000; Efron, 2012).

In the FETM, ARD places the prior on $\gamma_{e,k,v}$:

$$\sigma_{e,k,v} \sim \text{Gamma}(a, b),$$
$$\gamma_{e,k,v} \sim \mathcal{N}(0, \sigma_{e,k,v}^{-1}).$$

We set the parameters of the Gamma distribution by maximizing the likelihood of the data:

$$\hat{a}, \hat{b} = \arg\max_{a,b} p(\mathcal{D} \mid a, b). \tag{3}$$

This prior encourages the majority of the category-specific deviations to exhibit strong shrinkage, driving them towards zero, while allowing some to possess significant non-zero values. We incorporate it into the FETM to highlight influential fixed effects (γ), while still allowing β to capture most of the variation across documents.

2.3 Inference

With the FETM defined, we now turn our attention to procedures for inference and parameter estimation. FETMs rely on multiple latent variables: topic-word distributions β, document-topic proportion θ, and environment-specific deviations on the topic-word distribution γ. Conditional on the text and document specific features, we perform inference on these latents through the posterior distribution $p(\theta, z, \beta, \gamma|\mathscr{D})$, where $\mathscr{D} = \{(\mathbf{w}_1, \mathbf{x}_1), \ldots, (\mathbf{w}_n, \mathbf{x}_n)\}$.

As calculating this posterior is intractable, we rely on approximate inference. We use black-box variational inference (BBVI) Ranganath et al., 2014. Using the reparameterization trick we marginalize out z_{ij}, leaving us with only continuous variables (Kingma and Welling, 2013).

We rely on mean-field variational inference to approximate the posterior distribution (Jordan et al., 1999; Blei et al., 2017). We set $\phi = (\theta, \beta, \gamma)$ as the variational parameters, and let $q_\phi(\theta, \beta, \gamma)$ be the family of approximate posterior distribution, indexed by the variational parameters. Variational inference aims to find the setting of ϕ that minimizes the KL divergence between q_ϕ and the posterior (Blei et al., 2017). To approximate θ, we use an encoder neural network that takes $\mathbf{w}_i$ as input and consists of one hidden layer with 50 units, ReLU activation, and batch normalization. Minimizing this KL divergence is equivalent to maximizing the evidence lower bound (ELBO):

$$\text{ELBO} = \mathbb{E}_{q_\phi}[\log p(\theta, \beta, \gamma) + \log p(x|\theta, \beta, \gamma) - \log q_\phi(\theta, \beta, \gamma)]. \tag{4}$$

To approximate the posterior, we use the mean-field variational family, which results in our latent variables, θ, β, and γ being mutually independent and each governed by a distinct factor in the variational density. We employ Gaussian factors as our variational densities, thus our objective is to optimize the ELBO with respect to the variational parameters:

$$\phi = \{\mu_\theta, \sigma^2_\theta, \mu_\beta, \sigma^2_\beta, \mu_\gamma, \sigma^2_\gamma\}.$$

The model parameters are optimized using minibatch stochastic gradient descent in PyTorch by minimizing the negative ELBO. To achieve this optimization, we employ the Adam optimizer (Kingma and Ba, 2014). The complete algorithm is described in Algorithm 1.

3 Glassdoor Data

Our study utilizes a substantial corpus of employee-generated content sourced from Glassdoor, an extensive online platform established in 2008. Glassdoor serves as a repository for current and former employees to anonymously share insights regarding their workplace experiences. By early 2022, the platform reported significant engagement, with 55 million unique monthly visitors and reviews covering 1.9 million employers[1]. The platform captures a wide spectrum of the workforce, featuring over 190,000 distinct job titles, with a predominant representation of rank-and-file employees over senior executives Karabarbounis and Pinto, 2018.

For this research, we constructed a firm-year panel by collecting Glassdoor job reviews for S&P 1500 firms spanning the years 2008 to 2020. This dataset, comprising over a million original

[1]As reported by Expanded Ramblings, see `https://expandedramblings.com/index.php/numbers-15-interesting-glassdoor-statistics/` Expanded Ramblings, 2022.

Algorithm 1 Fixed Effects Topic Model (FETM)

Input: Number of topics K, number of words V, number of environments E
Output: Document intensities $\hat{\theta}$, global topics $\hat{\beta}$, environment-specific effects on global topics $\hat{\gamma}$
Initialize: Variational parameters μ_θ, σ^2_θ, μ_β, σ^2_β, μ_γ, σ^2_γ randomly
while the *evidence lower bound* (ELBO) has not converged **do**
    sample a document index $d \in \{1, 2, \ldots, D\}$
    sample z_θ, z_β, and $z_\gamma \sim \mathcal{N}(0, I)$ ▷ Sample noise distribution
    Set $\tilde{\theta} = \exp(z_\theta \odot \sigma_\theta + \mu_\theta)$ ▷ Reparameterize
    Set $\tilde{\beta} = \exp(z_\beta \odot \sigma_\beta + \mu_\beta)$ ▷ Reparameterize
    Set $\tilde{\gamma} = \exp(z_\gamma \odot \sigma_\gamma + \mu_\gamma)$ ▷ Reparameterize
    for $v \in \{1, \ldots, V\}$ **do**
        Set $w_{dv} = \sum_k \tilde{\theta}_{dk}(\tilde{\beta}_{kv} + \tilde{\gamma}_{ekv})$ ▷ Log-likelihood term
    end for
    Set $\log p(w_d|\tilde{\theta}, \tilde{\beta}, \tilde{\gamma}) = \sum_v \log p(w_{dv}|\tilde{\theta}, \tilde{\beta}, \tilde{\gamma})$ ▷ Sum over words
    Compute $\log p(\tilde{\theta}, \tilde{\beta}, \tilde{\gamma})$ and $\log q(\tilde{\theta}, \tilde{\beta}, \tilde{\gamma})$ ▷ Prior and entropy terms
    Set ELBO $= \log p(\tilde{\theta}, \tilde{\beta}, \tilde{\gamma}) + N \cdot \log p(w_d|\tilde{\theta}, \tilde{\beta}, \tilde{\gamma}) - \log q(\tilde{\theta}, \tilde{\beta}, \tilde{\gamma})$
    Compute gradients ∇_ϕELBO using automatic differentiation
    Update parameters ϕ
end while
Return approximate posterior means $\hat{\theta}$, $\hat{\beta}$, $\hat{\gamma}$

reviews, has been leveraged in prior academic research (e.g., Cai et al., 2024; Lee et al., 2021; Green et al., 2019). Glassdoor also addresses data quality concerns like self-selection bias via several key mechanisms: a "give-to-get" policy encourages broad participation Marinescu et al., 2021; user accounts, linked to real names (though reviews are published anonymously), enhance authenticity Green et al., 2019; and content moderation systems bolster review reliability.

The rich, textual nature of Glassdoor reviews makes this dataset particularly well-suited for our study on topic modeling. Specifically, these reviews offer a valuable lens into employee perceptions and discussions of corporate culture, which is a central focus of our application of FETM. The narratives within these reviews contain diverse linguistic expressions related to various workplace themes. Furthermore, the heterogeneity of the data, originating from employees across a multitude of firms and industries, presents an ideal scenario for testing the core capability of our proposed FETM: its ability to disentangle universal thematic content from context-specific language, such as industry jargon. This is crucial, as employees in different sectors may discuss similar underlying cultural aspects using varied terminology.

4 Evaluating FETM

We estimate the Fixed Effects Topic Model (FETM) on the Glassdoor review corpus with two sets of fixed effects: industry and sentiment.

For the industry, we consider one fixed effect per two-digit NAICS industry code. This fixed effects essentially controls for systematic variation in language across industries. For sentiment, recall that entry in our Glassdoor dataset has a text "pro" section, as well as a "con" section. The sentiment fixed effect encodes whether the review is "pro" or "con", and captures difference in positive and negative assessments within the same broad topic. We consider only NAICS code for which we have enough observations, leaving us with 18 industry fixed effects and 2 sentiment fixed effects.

In the remaining sections, we compare the performance of our FETM approach with what we call a traditional approach which we refer to as Vanilla Topic Model (VTM), with the same number of global topics but without fixed effects. This benchmark resembles standard approaches in which all semantic variation is absorbed by the topics themselves. VTM follows exactly the generative and training process of FETM, but without the γ parameter. Throughout the analysis, we maintain this side-by-side comparison between the baseline model and FETM to highlight how introducing fixed effects alters both model performance and substantive conclusions.

Incorporating fixed effects into the generative process changes both the performance and the interpretation of the model. On the performance side, we expect FETM to achieve lower perplexity than a standard variational topic model (VTM). The reason is that VTM must use additional topics to absorb environment-specific variation, whereas FETM partials out this variation through fixed effects. As a result, FETM can reach a given level of fit with fewer topics, since its topics are not forced to double as proxies for industry jargon or polarity.

On the data analysis side, FETM should yield topic intensities that are more stable across industries. This "environment-agnostic" property follows directly from the model's structure: fixed effects absorb the contextual language, leaving global topics to reflect broad underlying themes. In practice, this distinction becomes evident when comparing predictions across models. When FETM identifies a review as centered on a single topic while VTM produces a more diffuse distribution, it indicates that FETM has captured a clean global theme. Conversely, when VTM assigns a sharp topic distribution but FETM remains more dispersed, this typically reflects VTM's tendency to conflate environment-specific terminology with genuine thematic content.

Finally, if FETM is indeed recovering global structure rather than environment-specific correlations, it should generalize better in applications that depend on substantive content rather than linguistic artifacts. Consistent with this reasoning, we expect FETM to recover causal effects more accurately and to predict real-world outcomes—with greater precision than VTM.

For the analysis that follows, when comparing FETM and VTM, we mostly focus on the 30-Topics versions of both models.

4.1 Estimated Topics: an Example

Table 3: Illustration of the "Career Growth/Opportunities" Topic under VTM and FETM

	Top Words
VTM	career, learn, leadership, slow, market, available, extremely, frustrating, technologies
FETM: β	opportunities, career, advancement, growth, development, professional, willing
FETM: $\beta + \gamma_{\textbf{Con}}$	growth, career, slow, opportunities, advancement, development, boring, minimal
FETM: $\beta + \gamma_{\textbf{NAICS 51}}$	career, opportunities, technology, growth, development, advancement, microsoft, software

As an introduction to this section, and to fix ideas about the differences between FETM and VTM, we provide an illustrative example of how the two models estimate a given topic. We focus on what we identify as the "Career Growth" topic, which is representative of the type of semantic themes recovered by the models. The example contrasts the global topic produced by VTM with the general and fixed-effects refinements produced by FETM. The full list of estimated topics is reported in the Appendix.

We evaluate whether accounting for fixed effects improves topic quality by holding a single semantic theme fixed and contrasting how models represent it. The table reports four columns: (i) the VTM topic; (ii) the FETM global (β); (iii) the FETM adjusted "Con" environment ($\beta + \gamma_{\text{Con}}$); and (iv) the FETM adjusted information sector environment ($\beta + \gamma_{\text{NAICS 51}}$). For the adjusted columns, we form environment-specific topics by adding the relevant γ to β. The design of FETM should isolate what is common across environments from what is specific to sentiment (pros/cons) or industry, while keeping the underlying semantic object constant.

The comparison reveals precisely the contamination problem and its remedy. The VTM column blends general career language with environment-specific content and noise (e.g., *slow, frustrating, technologies, extremely*), indicating that, absent fixed effects, the "global" topic absorbs sentiment and indsutry jargon. FETM's β recovers instead a much cleaner backbone, with most of the load going on a more congruent set of words (*career, advancement, growth, development...*). Adding γ_{Con} re-introduces negative framing (*slow, boring, minimal*) without distorting the global theme, while adding $\gamma_{\text{NAICS 51}}$ injects domain-specific terms (*technology, software, Microsoft*).

The illustrative comparison clarifies how fixed effects reshape the semantic structure of topics. Our next step is to compare the models' performance via perplexity scores, which measure how well

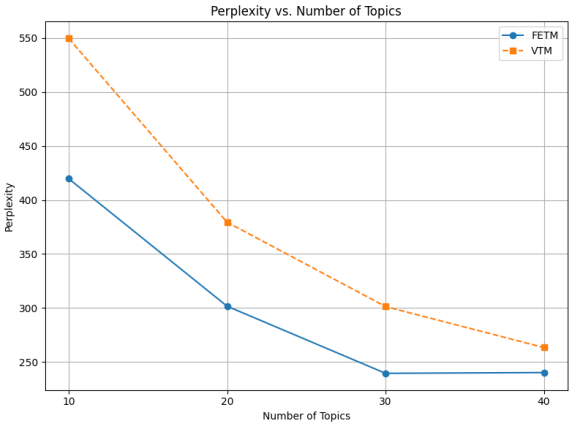


Figure 2: Perplexity across different subsets of the data.

each model explains held-out documents. Lower perplexity indicates better predictive fit, and thus provides a formal criterion to assess whether accounting for fixed effects improves performance beyond the descriptive gains highlighted so far.

4.2 FETM Performs Systematically better than VTM

A direct implication of our discussion is that FETM should fit the data more efficiently: it should achieve lower perplexity than a standard variational topic model (VTM), and it should require fewer topics to reach a given level of performance. We begin by testing these predictions. While perplexity is an imperfect measure of topic models (Chang et al., 2009), it remains useful for assessing topic stability across environments and for evaluating how well models generalize to data from different distributions.

Figure 3 compares perplexity for FETM and VTM as the number of topics increases. The results show a clear pattern: FETM achieves substantially lower perplexity across the board, with the gains most pronounced when the model is restricted to a smaller number of topics. This is precisely the setting where VTM is forced to make a trade-off between capturing global themes and soaking up environment-specific variation. By assigning industry and sentiment effects to fixed effects, FETM frees its topics to represent only broad concepts. As a result, FETM achieves the predictive performance of a 30-topic VTM with only 20 topics. Beyond about 30 topics the gap between the models narrows, but the message is clear: FETM uses its topic capacity more efficiently, reaching high levels of performance with fewer topics.

Figure 2 provides a complementary perspective by evaluating the two models on different slices of the data. Again FETM dominates VTM, with lower perplexity on the full corpus, on pro reviews, on con reviews, and even within a single industry (Accommodation and Food Services). The relative advantage is most pronounced in the subsets where contextual variation is strongest—for instance,

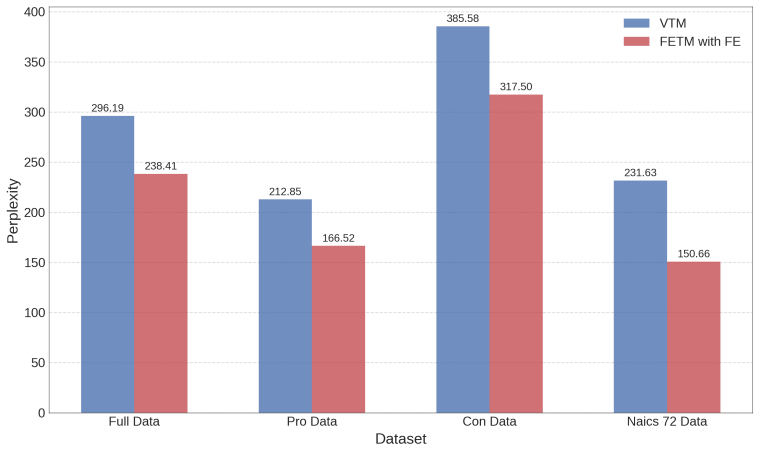


Figure 3: Perplexity as a function of the number of topics. FETM achieves lower perplexity across all values.

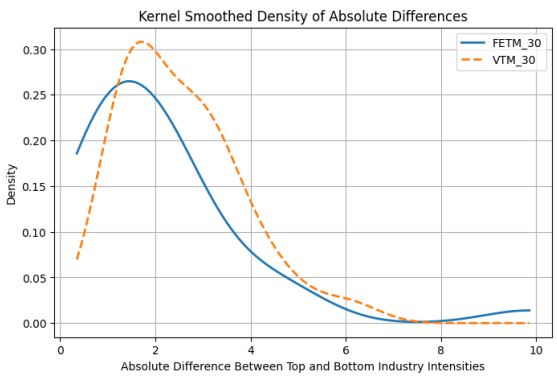
negative reviews, which have a distinctive vocabulary of complaints, or single-industry samples, where sector-specific jargon dominates. In these settings VTM struggles, because it must use its topics to absorb environment-specific language. FETM, by contrast, performs robustly: its fixed effects capture the local variation, leaving topics to reflect the global cultural themes that cut across contexts.

Taken together, the two figures demonstrate that FETM is not simply a more flexible statistical model; it is a more economical one. It delivers lower perplexity overall, and it does so with fewer topics, meaning that the topics themselves are sharper and more interpretable. Moreover, its advantage grows precisely in the places where context matters most, underscoring the value of building fixed effects into the generative process rather than attempting to control for them ex post. These results provide the foundation for our subsequent analysis: if FETM topics are indeed environment-agnostic, then they should not only fit better but also prove more stable across industries and more predictive of meaningful external outcomes.

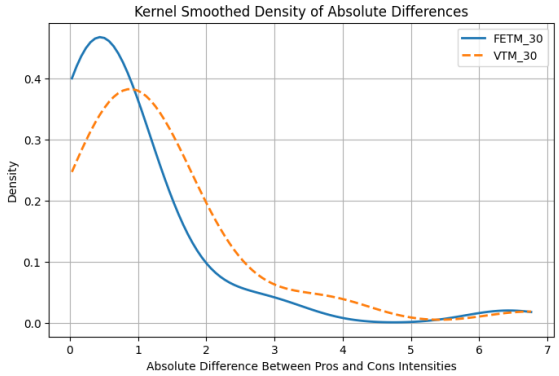
These findings suggest that FETM is successfully disentangling global themes from local linguistic environments. We now turn to examining this property directly. In particular, we assess whether FETM produces topic intensities that are more stable across industries and review types, thereby addressing the fixed-effects problem that motivates our approach.

4.3 The "Environment-Agnostic" Property of FETM

The first exercise we present to argue that FETM is indeed environment-agnostic relies on directly comparing the volatility of topics in FETM against VTM. If the model is working as intended, topic intensities should be relatively stable across environments, reflecting the prevalence of broad cultural features rather than the language used to describe them.



FETM Mean: 2.41 - VTM Mean: 2.54



FETM Mean: 0.87 - VTM Mean: 1.35

Figure 6: Distribution of Absolute Topic Intensity Differences Across Contexts

A defining objective of FETM is to separate portable, general cultural features from contextual language that reflects the environment in which reviews are written. Standard topic models often conflate the two: a topic may appear to capture workplace "management practices," but in practice it is dominated by industry-specific jargon (e.g., "claims" in insurance, "fulfillment" in retail) or by polarity of sentiment (positive in pros, negative in cons). By introducing fixed effects, FETM is designed to absorb systematic contextual variation so that the residual topics reflect patterns that generalize across environments.

To evaluate whether this mechanism works in practice, we compare how volatile topics are across environments in the two models. For each topic, we compute the maximum–minimum difference in estimated topic intensities across environments. This statistic asks, "How much more is this topic used in the environment where it is most prevalent than in the one where it is least prevalent?" If a topic captures a broad cultural trait such as work-life balance, its intensity should be relatively similar across industries and review types, and the max–min difference will be small. If instead a topic reflects contextual language, its intensity will vary sharply across environments, producing a large difference.

We also verify that our conclusions are not sensitive to the specific choice of statistic. While the main analysis uses the maximum–minimum difference across environments, we obtain similar results when we summarize variation using alternative measures such as the interquartile range or the difference between the upper and lower quartile environments. These statistics downweight extreme industries and instead capture typical dispersion across environments. In all cases, the distribution of topic-level variation shifts downward under FETM relative to VTM, reinforcing the conclusion that FETM recovers topics that are less tied to contextual factors.

This measure has two appealing features. First, it is simple and transparent. Second, it requires no alignment across models: we treat each set of topics as its own collection, summarize environment

variation topic-by-topic, and then compare the resulting distributions across VTM and FETM. The test is also deliberately stringent: max–min relies on the two most extreme environments, so it accentuates sensitivity rather than smoothing it away. A systematic reduction in this statistic under FETM would therefore provide strong evidence that the model succeeds in producing environment-agnostic topics.

Figure 6 reports the distribution of max–min differences across topics for both models. Two environments are considered: industries (left panel) and review type (right panel). Under VTM, topics are highly sensitive to industry, with an average max–min difference of 2.54 percentage points. Under FETM, this difference falls to 2.41. Although modest in magnitude, the distribution shifts systematically toward lower values under FETM, consistent with the interpretation that fixed effects absorb industry-specific language and leave behind more portable cultural topics.

The contrast is sharper when comparing positive and negative reviews. VTM topics display an average difference of 1.35 percentage points between pros and cons, indicating that standard models tend to assign entire topics to sentiment polarity. By contrast, FETM reduces this gap substantially, to only 0.87, suggesting that it separates the cultural dimension of a topic (e.g., work-life balance) from whether that dimension is discussed positively or negatively.

As an additional validation, we examine the extent to which topics from each model predict environments directly. We first regress the pro/con indicator on topic proportions:

$$\text{pro}_i \;=\; \alpha + \sum_k \beta_k \, \hat{\theta}^m_{ik} + \varepsilon_i, \qquad m \in \{\text{FETM}, \text{VTM}\}. \tag{5}$$

Results are reported in Table 4. VTM topics explain a substantially larger share of variation in the pro/con outcome ($R^2 = 0.175$) than FETM topics ($R^2 = 0.094$), around 86% more. This gap is informative: VTM appears to capture not only cultural dimensions but also context-specific polarity, whereas FETM strips away part of this variation and recovers topics that are less mechanically tied to sentiment labels.

Table 4: Predicting Sentiment with Topics

	VTM Topics	FETM Topics
N	1,181,688	1,181,688
R^2	0.175	0.094

We then reverse the exercise and ask whether topics predict industry membership. For transparency, we focus on a binary classification between two of the largest industries in the data:

Manufacturing (NAICS2 = 33) versus Information (NAICS2 = 51). Specifically, we estimate

$$\text{manuf}_i \;=\; \alpha + \sum_k \beta_k \, \hat{\theta}^m_{ik} + \varepsilon_i, \qquad m \in \{\text{FETM}, \text{VTM}\}, \tag{6}$$

where $\text{manuf}_i = 1$ if a review belongs to Manufacturing and 0 if it belongs to Information. Table 5 reports the results. VTM topics predict industry with an R^2 of 0.0064, while FETM topics achieve only 0.0048, or a 33% increase. Although these values are small in absolute terms, the consistent pattern is that VTM entangles topics with industry-specific language, while FETM removes much of this predictability by absorbing contextual variation.

Together with the volatility results, these regressions provide a complementary perspective on the environment-agnostic property of FETM. The volatility comparison shows that FETM topics move less across environments, while the regressions demonstrate that they are also less predictive of environments in the first place. In other words, FETM topics are not only smoother across industries and review types, they are also less entangled with environmental identifiers. This is precisely the intended mechanism: contextual variation is absorbed during estimation, leaving topics that reflect substantive cultural content. By contrast, VTM learns topics that appear more powerful in explaining both sentiment and industry, but this explanatory power is spurious—it arises because VTM conflates general themes with the particular vocabularies of different contexts.

This distinction matters for inference. If the goal is to study how broad cultural traits evolve across firms, industries, or time, a model whose topics fluctuate with contextual language will produce misleading comparisons: apparent differences in topic intensity may reflect nothing more than variation in terminology across settings. The environment-agnostic property of FETM mitigates this risk. By design, it recovers topics that are portable and comparable, providing a stable foundation for downstream analyses of cultural dynamics and causal inference.

Table 5: Predicting Industry with Topics: Manufacturing (33) vs Information (51)

	VTM Topics	FETM Topics
N	415,502	415,502
R^2	0.0064	0.0049

4.4 FETM is Better at Recognizing Broad Topics.

We posit that FETM learns general themes—topics that recur across firms and industries—whereas a vanilla topic model (VTM) tends to blend thematic content with industry or firm idiosyncrasies. A simple implication is that, in documents where one model is markedly more confident that a single topic dominates, that model's top topic should align more closely with a human (or human-like)

assignment of a broad, sector-agnostic label. We operationalize this idea by contrasting the two models precisely in the reviews where their inferred topic concentration differs most.

Formally, for each review i, we compute

$$\Gamma_i \;=\; \frac{\max_k \hat{\theta}_{ik}^{\mathrm{FETM}}}{\max_k \hat{\theta}_{ik}^{\mathrm{VTM}}},$$

where $\hat{\theta}_{ik}^{\mathrm{FETM}}$ and $\hat{\theta}_{ik}^{\mathrm{VTM}}$ denote the document–topic share (posterior mean) for topic k under FETM and VTM, respectively. Thus, $\Gamma_i > 1$ indicates that FETM places more mass on its top topic (is more concentrated) than VTM for the same document; $\Gamma_i < 1$ indicates the reverse. The empirical distribution of Γ_i is reported in Figure 11.

Two examples illustrate the contrast. In Appendix Figure 12, a retail-sector review criticizes poor management and a stressful work environment. FETM assigns most probability mass to Topic 10 (*Managerial Oversight, Performance Reviews, and Trust in Leadership*), while VTM disperses mass across several topics. This is a high-Γ case. By contrast, Appendix Figure 13 shows a low-Γ review describing engineers and experimental projects: VTM concentrates on Topic 14 (*Technical Innovation, Senior Leadership & Corporate Practices*), while FETM spreads mass across multiple themes. Intuitively, FETM's concentration spikes when the content cuts across firms (e.g., "management quality"), whereas VTM's concentration often coincides with industry-coded language ("engineers," "technology") that may not translate into a broad, cross-sector theme.

We use a large-language-model (LLM) annotator to provide a common, sector-agnostic reference label. Specifically, we draw two subsets of reviews: the 100 with the highest Γ_i ("High-Γ", where FETM is more concentrated) and the 100 with the lowest Γ_i ("Low-Γ", where VTM is more concentrated). For each review, we prompt the LLM with a standardized instruction to assign a single, broad topic label (e.g., "management quality," "compensation and benefits," "work–life balance," "innovation culture").[2]

To quantify alignment, we compare each model's top topic for a given document to the LLM-assigned label. Let $k_i^{\mathrm{FETM}} = \arg\max_k \hat{\theta}_{ik}^{\mathrm{FETM}}$ and $k_i^{\mathrm{VTM}} = \arg\max_k \hat{\theta}_{ik}^{\mathrm{VTM}}$ be the top topics. We embed the label (a short phrase) into a fixed semantic space and construct an embedding for a topic as the probability-weighted average of the embeddings of its top words, $v_k = \sum_w p(w \mid k)\, e(w)$.[3] For the High-$\Gamma$ set we compare $v_{k_i^{\mathrm{FETM}}}$ to the label embedding; for the Low-Γ set we compare $v_{k_i^{\mathrm{VTM}}}$ to the label embedding. If the more concentrated model has indeed identified the *broad* theme of

[2]The prompt asks for one short, sector-agnostic label describing the main theme of the review. We use the same prompt for all documents: *"Your task is to classify the following Glassdoor Review into 1 broad, comprehensive Topic. Please do that and provide only the name of that topic."*

[3]We compute v_k using the top words in topic k weighted by their topic–word probabilities $p(w \mid k)$. The label is embedded with the same encoder; alignment is measured by cosine similarity. Results are not sensitive to using top-N words (e.g., $N \in \{25, 50, 100\}$).

the document, the cosine similarity should be higher.

Figure 7 reports the average cosine similarity in each set. In High-Γ reviews (where FETM is more concentrated), the alignment between FETM's top topic and the LLM label is markedly higher (mean $= 27.64$) than the corresponding alignment for VTM in Low-Γ reviews (mean $= 23.43$). This pattern is exactly what the fixed-effects logic predicts: by absorbing firm and industry heterogeneity, FETM's topics capture cross-cutting themes that generalize beyond sectoral jargon; when VTM is confident, its dominant topic is more likely to reflect industry-specific language that an annotator naturally treats as narrower than the broad label.

Taken together, these facts support our central claim: when a single topic dominates a document, FETM's dominant topic aligns more closely with broad, sector-agnostic semantics than VTM's. This is consistent with FETM's design—partialing out firm/sector fixed effects before learning topics—and it helps explain why FETM yields more portable themes in downstream analyses.

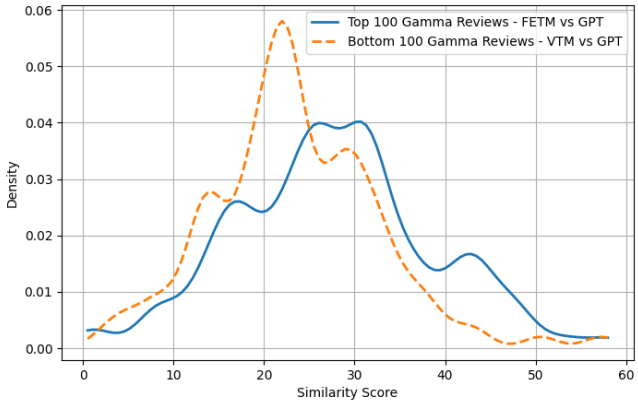


Figure 7: Average similarity between model-predicted topic and LLM topic.
FETM (High Γ) Mean: 27.64 - VTM (Low Γ) Mean: 23.43

4.5 Synthetic Experiment

The motivation for FETM is that estimated topic proportions $\hat{\theta}_i$ can be systematically biased when documents contain context-specific vocabulary. Words tied to particular firms or industries may be absorbed into global topics, thereby distorting both the estimated topic–word distributions β_k and the document–topic proportions. Such misattribution can bias any downstream causal analysis. To illustrate how FETM mitigates this issue, we design a semi-synthetic experiment with a known ground truth.

We begin by selecting a theme. To do so, we prompted an LLM with the VTM and FETM topics and asked it to align topics across models: *"For each of the FETM topics, what is the closest VTM topic?"* We then asked: *"Among these pairs, which one most clearly identifies a broad, well-defined theme, given a set of Glassdoor reviews?"*. The LLM identified the Compensation

and Benefits theme, with FETM Topic 11 and VTM Topic 8 as the clearest pair. FETM Topic 11 includes terms such as "employees, pay, health, insurance, bonus, plan, benefits," while VTM Topic 8 contains related terms such as "insurance, benefit, medical, supervisor." We also verified that other topics—FETM Topics 2, 24, and 28 and VTM Topics 2 and 23—capture additional aspects of the Compensation and Benefits dimension.

On this basis, we define a "true" Compensation and Benefits vocabulary D consisting of forty canonical unigrams and bigrams such as "salary," "health insurance," "pension plan," "bonus structure," and "maternity leave." The full list is provided in the Appendix.

For each document i, we then construct a ground-truth treatment indicator

$$\theta_i(\tau) = \begin{cases} 1 & \text{if document } i \text{ contains at least } \tau \text{ words from } D, \\ 0 & \text{otherwise,} \end{cases}$$

where the threshold τ controls the strictness of the definition. As τ increases, the treated group becomes smaller and more sharply defined.

We generate outcomes with a fixed treatment effect of 0.2. For each document we draw

$$Y_i(\tau) \sim \text{Binomial}\left(p_i = \tfrac{1}{2} + 0.2 \cdot \theta_i(\tau) + \epsilon_i\right), \quad \epsilon_i \sim \mathcal{N}(0, 0.05).$$

This implies $\mathbb{E}[Y_i|\theta_i = 1] = 0.7$ and $\mathbb{E}[Y_i|\theta_i = 0] = 0.5$, so that the true effect equals 0.2. The estimation task for each model is to recover this effect from inferred topic proportions rather than from the known θ_i.

For implementation, we define the model-implied treatment indicators by mapping documents to the set of topics most clearly associated with Compensation and Benefits. Specifically, $\hat{\theta}_i^{FETM} = 1$ if i is assigned to Topics 2, 11, 24, or 28 in FETM, and $\hat{\theta}_i^{VTM} = 1$ if i is assigned to Topics 2, 8, or 23 in VTM. This yields a treated share of approximately 16.8% under FETM and 14.6% under VTM.

For each model m we then estimate

$$Y_i = \beta_0 + \beta_1 \hat{\theta}_i^m + \eta_i,$$

where $\hat{\theta}_i^m$ is the model-implied proportion of the Compensation and Benefits topic in document i. A well-specified model should deliver $\hat{\beta}_1$ close to 0.2. Because the treated group shrinks as τ increases, we report bootstrap confidence intervals for $\hat{\beta}_1$ to account for sampling variability.

Figure 8 shows the results. FETM consistently delivers estimates closer to the true effect across thresholds, while VTM underestimates the effect more severely, particularly when τ is large and the treated group is defined by more specific vocabulary. Increasing τ sharpens the definition of

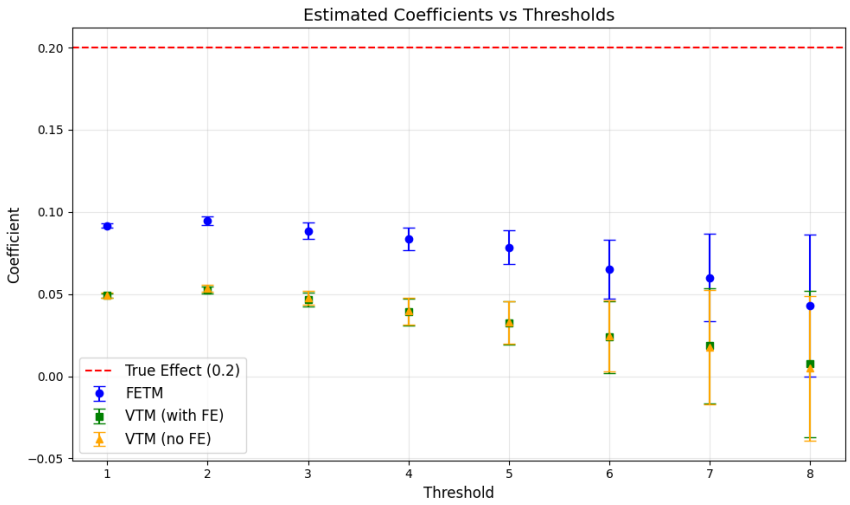


Figure 8: Estimated treatment effect (β_1) across models and thresholds (τ). Bootstrapping confidence intervals are shown. FETM outperforms VTM as τ increases, approaching the true effect (0.2) more reliably.

treatment, so documents classified as treated contain language unambiguously tied to Compensation and Benefits. At the same time, the number of treated documents falls, which makes the estimates noisier and widens the confidence intervals.

Neither model recovers the full treatment effect of 0.2 because the estimated topic proportions are only a noisy proxy for the true treatment indicator. This attenuation is present for both models, with FETM effect around 0.1 while VTM is further away from the true effect, and drifts further downward as τ increases. The experiment demonstrates that although both models become noisier when the treated set is small, FETM consistently isolates the Compensation and Benefits theme more effectively, yielding estimates that remain substantially closer to the true effect.

Finally, we compare to a specification where firm or industry fixed effects are added ex post to regressions based on VTM proportions. The results are nearly unchanged from baseline VTM, underscoring that the improvement arises from incorporating fixed effects in the topic estimation stage itself. Because FETM learns topic–word distributions net of contextual heterogeneity, its topics remain interpretable and transferable in ways that ex post corrections cannot replicate.

The synthetic experiment thus provides a controlled validation: when the ground truth is known, FETM recovers the true treatment effect more accurately and more robustly than VTM. This evidence supports the broader claim that fixed effects are essential for identifying topics that reflect generalizable themes rather than context-specific correlations.

5 Testing FETM on Additional Datasets

To assess scalability and external validity, we test FETM on three multi-environment datasets drawn from distinct political domains. Each dataset defines environments exogenously (e.g., partisan regions, speaker party, or source type) so that we can evaluate whether fixed effects improve topic quality while preserving comparability across settings.

Table 6: A summary of the datasets we construct for testing topic models across multiple environments.

Dataset	**Style**	**Ideology**	**Political advertisements**
Focus of text	US Immigration	Politics	Politics
Environments	{Tweets from US Senators, US Senate speeches, news articles}	{Republican, Democrat} politicians	Channels from {Republican, Democrat} voting regions
Training set size	$4,052$ per environment	$12,941$ per environment	$12,446$ per environment

5.1 Political Ads Dataset

The Political Ads dataset contains 24,892 TV advertisements aired across U.S. media markets. We define two environments by assigning channels from Republican-voting regions to one environment and channels from Democratic-voting regions to the other. We construct a unigram vocabulary including tokens that occur in at least 0.6% and at most 40% of documents, remove stopwords, and drop city, state, and politician names. For FETM, we use an empirical-Bayes specification with gamma prior parameters $a = 3.8$ and $b = 0.13$, trained for 15 epochs; in each epoch we take two gradient steps to update (a, b) for every one step used to update the remaining parameters.

Table 7: Example advertisements. WKRG is licensed to Mobile, AL; KSWB is based in San Diego, CA.

Source	Text
Alabama (WKRG)	*What does Governor Bob Riley call over 70,000 new jobs? A great start. His conservative leadership's given us the lowest unemployment in Alabama history, turning a record deficit into a record surplus. Now Governor Riley has delivered the most significant tax cuts in our history. The people get up every morning and work, they are the ones that allowed us to have the surplus. The only thing I'm saying, they should have some of it back. Governor Bob Riley, honest, conservative leadership.*
California (KSWB)	*State budget cuts are crippling my classroom. So I can't believe the Sacramento politicians cut a backroom deal that will give our state's wealthiest corporations a new billion dollar tax giveaway. . . . Prop 24 repeals the unfair corporate giveaway and puts our priorities first. Vote yes on Prop 24 because it's time to give our schools a break, not the big corporations.*

We illustrate how FETM separates a general backbone from environment-specific refinements using a *U.S. military* topic. The global component assigns high probability to generic military terms across regions; the environment deviations emphasize partisan framing.

Table 8: Top terms for a *U.S. military* topic in political ads.

Source	Top words
Global	*america*, *veterans*, *war*, *proud*, *iraq*, *military*, *troops*
Republican-leaning	*terror*, *liberties*, *isis*, *terrorism*, *freedom*, *terrorists*, *defeat*
Democratic-leaning	*iraq*, *stay*, *guard*, *veterans*, *soldiers*, *port*, *home*

The qualitative comparison clarifies how fixed effects reshape topic semantics. We now turn to a quantitative evaluation using perplexity on held-out documents. As usual, lower perplexity indicates a better predictive fit.

Across both regional test splits, FETM attains substantially lower perplexity than VTM, and adding a region-specific deviation further improves fit where partisan framing is salient. These results, together with the qualitative evidence above, indicate that modeling fixed effects yields cleaner general topics and better predictive performance while accommodating environment-specific language.

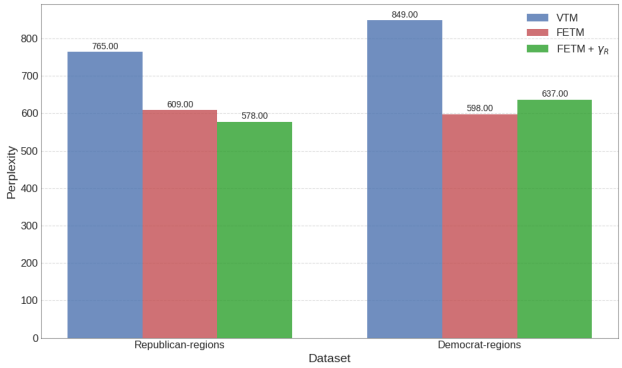


Figure 9: Perplexity on held-out data across models trained on a dataset of political advertisements from **political ads** across different regions of the U.S. The FETM $+\gamma_R$ represents global β with Republican-specific deviations γ_R. FETM outperforms all baselines across all regions.

5.2 Ideology Dataset

The Ideology dataset contains U.S. political advertisements from the past two decades. We define two environments by splitting ads according to the party of the sponsoring politician, yielding balanced samples from Republican and Democratic politicians (12,941 per environment). To evaluate generalization, we test models on three held-out sets: Republican-only ads, Democrat-only ads, and a balanced mixture of both. This setup allows us to assess whether FETM captures both shared themes and partisan-specific deviations.

We construct a unigram vocabulary that includes tokens appearing in at least 0.6% and at most 40% of the corpus. As before, we remove stopwords, locations, and politician names. For FETM, the gamma prior hyperparameters are set to $a = 4.0$ and $b = 0.11$, updated under empirical Bayes by taking two gradient steps for (a, b) for every one step of the remaining parameters, and training for 15 epochs.

Table 9 illustrates the decomposition achieved by FETM on a health-care related topic. The global backbone captures generic terms such as *health*, *insurance*, and *medicare*. The partisan deviations sharpen this theme in opposite directions: the Republican deviation emphasizes language around opposition to the Affordable Care Act (e.g., *obamacare*, *repeal*, *bureaucrats*), while the Democratic deviation stresses protections against insurers and rising prices (e.g., *conditions*, *deny*, *prices*). This structure reflects how FETM recovers a common semantic space while uncovering meaningful partisan-specific refinements.

Table 9 represents the top terms the FETM learns on ideological dataset.

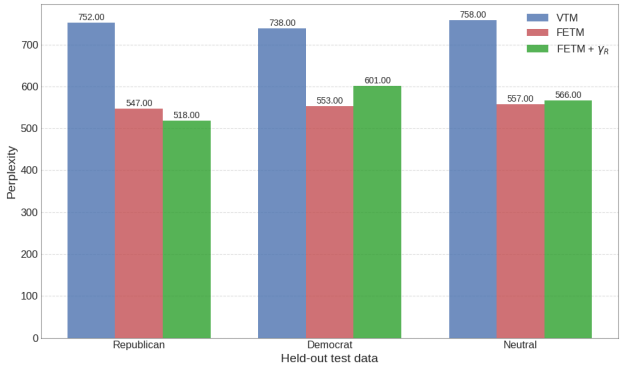


Figure 10: Perplexity on held-out data across models trained on the **ideological** dataset, consisting of political advertisements from Republican and Democrat politicians. The FETM $+\gamma_R$ represents global β with Republican-specific deviations γ_R. FETM outperforms all baselines on all three test sets.

Table 9: When trained on the ideological dataset FETM learns meaningful terms for the Republican and Democrat environments, while simultaneously uncovering meaningful global topics.

Source	**Top Words**
β_k: Global	*health*, *seniors*, *insurance*, *medicare*, *plan*, *costs*, *drug*, *affordable*, *healthcare*, *fix*
γ_k: Republican	*obamacare*, *health*, *takeover*, *bureaucrats*, *replace*, *medicare*, *supports*, *repeal*, *lawsuits*, *choices*
γ_k: Democrat	*health*, *companies*, *protections*, *conditions*, *deny*, *insurance*, *prices*, *voted*, *drug*, *gut*

We next compare predictive performance using perplexity. Figure 10 reports held-out test scores across the three splits. FETM achieves consistently lower perplexity than VTM, and adding Republican fixed effects further improves performance, but exclusively on the Republican-only test set. These results reinforce the qualitative evidence: by separating shared structure from partisan-specific variation, FETM yields more coherent topics and superior predictive fit.

5.3 Style Dataset

The style dataset consists of news articles, senator tweets, and senate speeches related to U.S. immigration. The U.S. immigration articles are gathered from the Media Framing Corpus (Card et al., 2015). We use all $4,052$ articles in the dataset. We augment the dataset used by Vafa et al. (2020), which is based on an open-source set of tweets of U.S. legislators from 2009–2017. We create a list of keywords related to immigration and sample $4,052$ tweets that contain at least one of

the keywords; we repeat the process for Senate speeches from the 111-114th Congress. (Gentzkow et al., 2018). The environments for the style dataset are defined by the distinct writing styles of tweets, speeches, and articles.

We constructed a vocabulary of unigrams that occurred in at least 0.6% and in no more than 50% of the documents. We use the same tokenization scheme for all baselines we compare to. We removed cities, states, and the names of politicians in addition to stopwords. For FETM, we set the hyperparameters of the gamma distribution, a and b, to be 3.7 and 0.34 respectively. These values were determined by training our model for 50 epochs, taking 2 gradient steps for updating a and b in the empirical Bayes method for every 1 step for the rest of the model. This approach helps guarantee that hyperparameter updates are not overshadowed by the updates of the rest of the parameters in the model. We set the number of topics, k, to be 20 for all experiments in this paper.

Table 10: Top words for a particular topic distribution learned by FETM when trained on the style dataset. The words in global topics appear across environments, while the words that receive the top γ values predominantly appear in one environment. We observe distinctive word choices in tweets, articles, and senate speeches, reflecting different communication styles.

Source	**Top Words**
Global	*country*, *law*, *status*, *policy*, *illegal*, *immigrants*, *immigration*, *border*, *citizenship*
News Articles	*immigration*, *primary*, *illegal*, *immigrants*, *legal*, *naturalization*, *states*, *driver*
Senate Speeches	*immigration*, *border*, *security*, *gang*, *secretary*, *everify*, *homeland*, *colleagues*
Tweets	*country*, *discuss*, *policy*, *immigration*, *reform*, *illegal*, *applications*, *check*, *plan*

We further evaluate FETM by testing its ability to generalize across environments with distinct linguistic styles. To this end, we construct an out-of-distribution exercise: models are trained only on Senate speeches and news articles and then evaluated on held-out tweets. This design isolates the extent to which a model trained on formal, relatively long-form text can adapt to the distinct vocabulary and brevity of social media communication. If fixed effects truly disentangle general topical content from environment-specific variation, FETM should yield representations that transfer more effectively to unseen styles.

When training on speeches and articles and testing on tweets the training dataset has 8104 samples. We constructed a vocabulary of unigrams that occurred in at least 0.8% and in no more

Table 11: Performance (held-out perplexity) across environments when training on congressional senate speeches and news articles. The FETM has substantially lower perplexity, especially when tested on the out-of-distribution tweets.

	In-Distribution		OOD
Model	Articles	Speeches	Tweets
VTM	$1,613$	$1,598$	$2,206$
FETM	**1,502**	**1,524**	**1,690**

than 50% of the documents. For FETM, we set the hyperparameters of the gamma distribution, a and b, to be 2.92 and 0.25 respectively. These values were determined by training our model for 50 epochs, taking 2 gradient steps for updating a and b in the empirical Bayes method for every 1 step for the rest of the model.

Table 11 reports perplexity scores for VTM and FETM under this setup. Both models achieve similar performance when evaluated on in-distribution test sets (articles and speeches), but the differences become pronounced once we turn to OOD tweets. While VTM's perplexity rises sharply when faced with the unfamiliar style of tweets, FETM maintains substantially lower perplexity. This result highlights the value of modeling environment-specific deviations: FETM's global backbone captures generalizable topical content, while its γ parameters absorb style-specific noise during training, preventing it from distorting the shared representation.

Fixed effects do more than improve interpretability of topic content; they also enhance out-of-sample predictive performance when text comes from a new environment. The gains on tweets—arguably the most challenging domain due to their short length and idiosyncratic vocabulary—suggest that FETM is especially well-suited for applications where models trained on one corpus must generalize to another, such as cross-platform political communication or media spillovers.

6 Conclusions

This paper develops a simple idea with broad implications: when text comes from multiple environments, we must distinguish what people talk about from how they talk. FETM achieves this by building contextual heterogeneity (for instance, industries, sentiment, or platforms) into the topic model itself. In doing so, it recovers a set of global topics that are comparable across settings and interpretable as genuine dimensions of meaning rather than artifacts of language.

Our results show that incorporating context at the estimation stage substantially improves both interpretability and generalization. Compared with standard topic models, FETM extracts themes that are more stable across environments, less predictive of contextual labels, and more aligned with

the intuitive content of documents. This stability matters empirically: when text is used to measure latent constructs such as managerial style or organizational culture, differences in estimated topic intensities can otherwise reflect vocabulary rather than substance. By correcting for this, FETM produces measures that travel better across firms, industries, and datasets.

Beyond fit and interpretability, the model clarifies a conceptual point about measurement with text. Treating fixed effects as part of the generative process, and not as an afterthought in regressions, changes what we learn from language. It ensures that text-based measures capture underlying constructs rather than artifacts of wording or style. As a result, researchers can use FETM to obtain topic-based measures that are closer to the underlying managerial or cultural constructs they aim to study, and less tied to the words through which those constructs happen to be expressed.

Our framework could be extended to settings with unobserved or evolving environments, hierarchical contexts, or supervised outcomes. But the central message is straightforward: text from multiple sources contains both concepts and context. Disentangling the two is essential for credible inference. FETM offers a practical way to do so, providing cleaner building blocks for the study of organizations, markets, and culture.

References

Bandiera, Oriana, Andrea Prat, Stephen Hansen, and Raffaella Sadun (2020). "Ceo behavior and firm performance". *Journal of Political Economy* 128.4.

Blei, David M, Alp Kucukelbir, and Jon D McAuliffe (2017). "Variational inference: a review for statisticians". *Journal of the American statistical Association* 112.518.

Cai, Wei, Andrea Prat, and Jiehang Yu (2024). "Measuring organizational capital". *Available at SSRN 4870532*.

Card, Dallas, Amber Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith (2015). "The media frames corpus: Annotations of frames across issues". *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*.

Carlin, Bradley P and Thomas A Louis (2000). "Empirical bayes: past, present and future". *Journal of the American Statistical Association* 95.452.

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David M Blei (2009). "Reading tea leaves: How humans interpret topic models". *Advances in neural information processing systems*. Vol. 22.

Efron, Bradley (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*. Vol. 1.

Expanded Ramblings (2022). *Numbers: 15 Interesting Glassdoor Statistics*. https://expandedramblings.com/index.php/numbers-15-interesting-glassdoor-statistics/. Accessed: January 2022 (content may update, original data point from description).

Gentzkow, Matthew, Bryan Kelly, and Matt Taddy (2019). "Text as data". *Journal of Economic Literature* 57.3.

Gentzkow, Matthew, Jesse M Shapiro, and Matt Taddy (2018). "Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts". *URL: https://data. stanford. edu/congress text*.

Green, Thomas C., Russell Jame, and Brandon Lock (2019). "Crowdsourced analyses of corporate culture and firm performance". *Journal of Financial Economics* 133.3.

Grimmer, Justin and Brandon M Stewart (2013). "Text as data: the promise and pitfalls of automatic content analysis methods for political texts". *Political analysis* 21.3.

Hansen, Stephen, Michael McMahon, and Andrea Prat (2018). "Transparency and deliberation within the fomc: a computational linguistics approach". *The Quarterly Journal of Economics* 133.2.

Jordan, Michael I, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul (1999). "An introduction to variational methods for graphical models". *Machine learning* 37.

Karabarbounis, Loukas and Gabriel Pinto (2018). *The Macroeconomic Effects of Firm-Level Uncertainty: A Case Study of an Online Review System*. Tech. rep. w25090. National Bureau of Economic Research.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: a method for stochastic optimization". *arXiv preprint arXiv:1412.6980*.

Kingma, Diederik P and Max Welling (2013). "Auto-encoding variational Bayes". *arXiv preprint arXiv:1312.6114*.

Lee, Jihun, Scott D. Dyreng, Jeffrey L. Hoopes, and Jaron H. Wilde (2021). "The effects of corporate tax avoidance news on employee perceptions of their employers". *Journal of Business Ethics* 171.4.

MacKay, David JC (1992). "Bayesian interpolation". *Neural computation* 4.3.

Marinescu, Ioana, Sanjid Islam, and Ellis J. G. Scharfenaker (2021). "The incentives to review: evidence from glassdoor.com". *Labour Economics* 73.

Quinn, Kevin M, Burt L Monroe, Michael Colaresi, Michael H Crespin, and Dragomir R Radev (2010). "How to analyze political attention with minimal assumptions and costs". *American Journal of Political Science* 54.1.

Ranganath, Rajesh, Sean Gerrish, and David Blei (2014). "Black box variational inference". *Artificial intelligence and statistics*. PMLR.

Sim, Yanchuan, Bryan Routledge, and Noah Smith (2015). “The utility of text: the case of Amicus briefs and the Supreme Court”. *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 29. 1.

Tipping, Michael E (2001). “Sparse bayesian learning and the relevance vector machine”. *Journal of machine learning research* 1.Jun.

Vafa, Keyon, Suresh Naidu, and David M Blei (2020). “Text-Based Ideal Points”. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Appendix

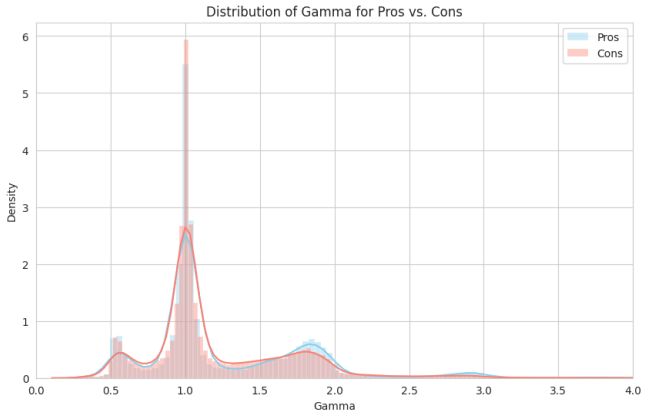


Figure 11: Distribution of concentration ratio Γ across reviews.

$$
D = \left\{ \begin{array}{lll}
\text{"benefits"} & \text{"salary"} & \text{"pay"} \\
\text{"wage"} & \text{"compensation"} & \text{"stock options"} \\
\text{"insurance"} & \text{"health insurance"} & \text{"retirement"} \\
\text{"pension"} & \text{"free meals"} & \text{"holidays"} \\
\text{"paid time off"} & \text{"gym membership"} & \text{"paid leave"} \\
\text{"bonus"} & \text{"equity"} & \text{"401k"} \\
\text{"healthcare"} & \text{"meal vouchers"} & \text{"transportation stipend"} \\
\text{"childcare support"} & \text{"tuition reimbursement"} & \text{"employee discounts"} \\
\text{"paid holidays"} & \text{"profit sharing"} & \text{"performance bonuses"} \\
\text{"paid maternity leave"} & \text{"paid paternity leave"} & \text{"company car"} \\
\text{"health and wellness programs"} & \text{"life insurance"} & \text{"dental insurance"} \\
\text{"vision insurance"} & \text{"employee stock purchase"} & \text{"retirement plans"} \\
\text{"paid volunteer time"} & \text{"educational assistance"} & \text{"relocation assistance"} \\
\text{"sign-on bonuses"} & &
\end{array} \right\}.
$$

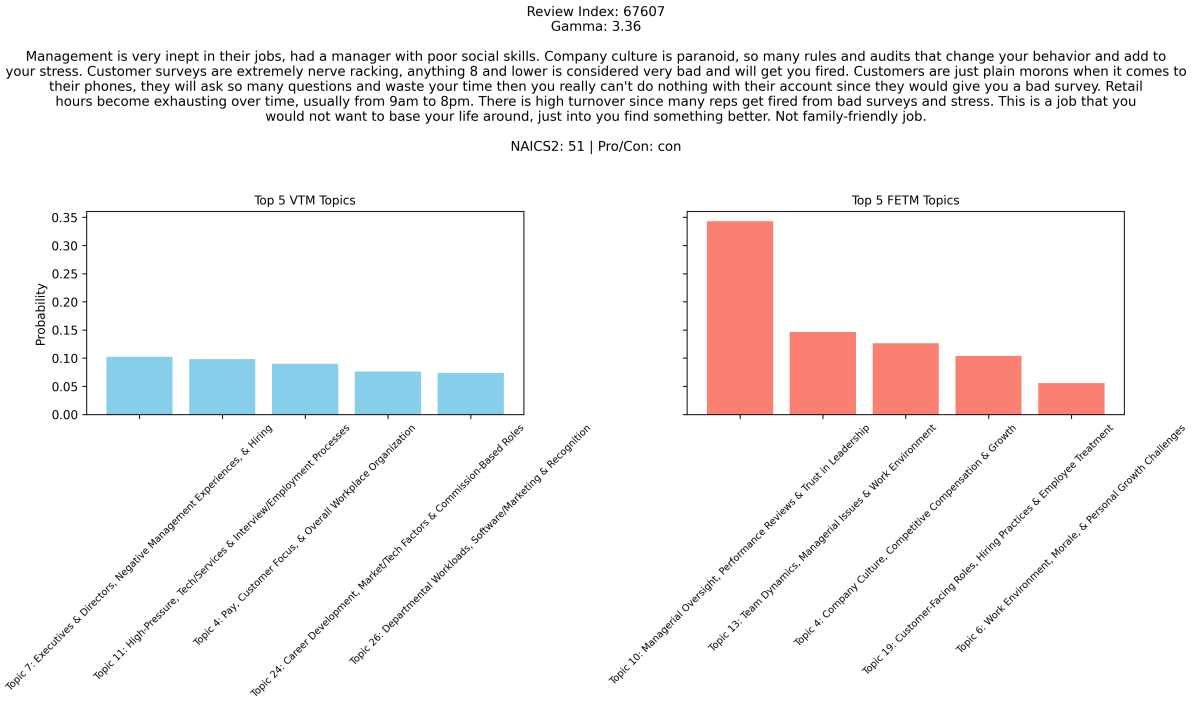


Figure 12: An high Γ review. FETM associates the review with a single topic (Topic 10), while VTM predictions are more distributed across topics.

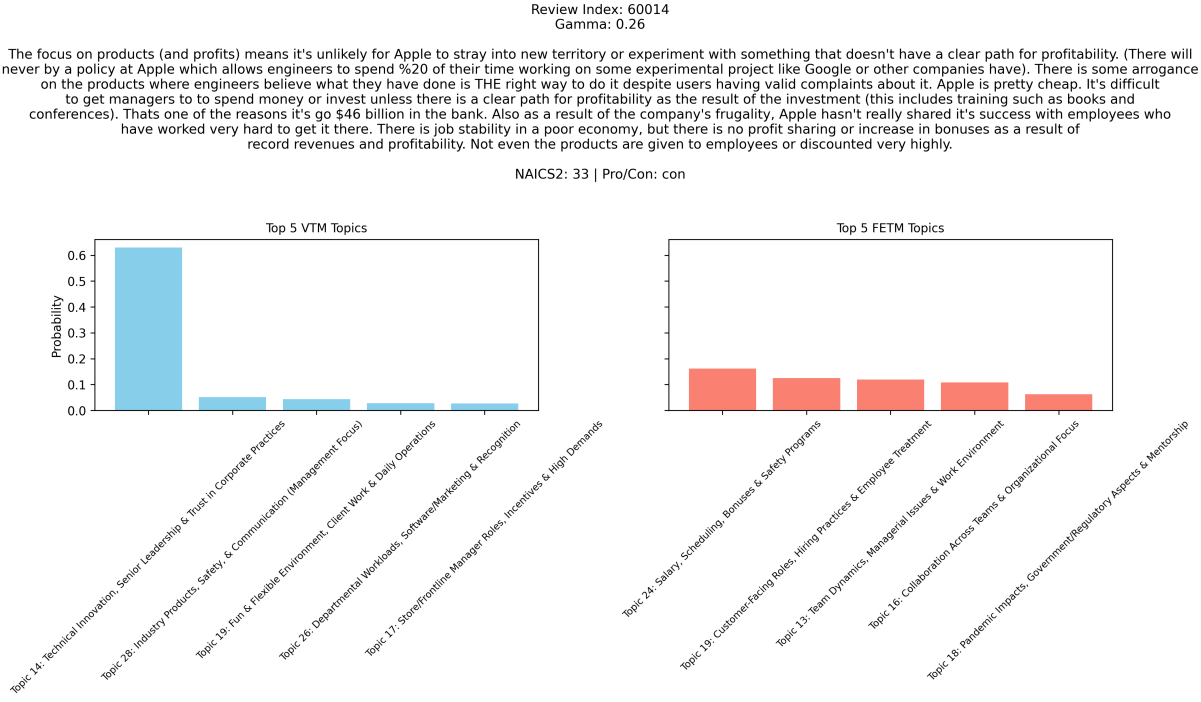


Figure 13: A low Γ review. VTM associates the review with a single topic (Topic 14), while FETM predictions are more distributed across topics.

Table 12: VTM Global Topics

Topic	Top Words
1	friendly, competitive, retail, values, diversity, roles, kind, started, general, major, toxic, door, sucks
2	hours, balance, salary, growth, personal, open, bonuses, depending, plan, package, weekends, breaks, tough
3	employee, sales, discount, food, product, potential, usually, cares, engineers, field, match, community, promote
4	pay, employees, customers, customer, help, jobs, average, goals, departments, holidays, school, value, deal
5	training, high, changes, performance, stressful, constant, turnover, layoffs, lunch, oriented, wonderful, provided, lead
6	job, grow, politics, based, old, area, short, rude, raise, multiple, recognition, numbers, demanding
7	told, generous, firm, executives, timings, left, directors, saying, defined, execution, overly, resulting, water
8	poor, decent, free, workers, insurance, shift, skills, world, helpful, tech, line, benefit, weeks
9	coworkers, paid, promotions, group, scheduling, calls, allowed, impossible, changed, idea, sharing, mentality, games
10	management, difficult, projects, upper, project, term, extra, middle, point, programs, experienced, possible, standards
11	bad, focus, pressure, especially, discounts, ibm, services, phone, pays, information, employment, goes, interview
12	advancement, overtime, constantly, expected, outside, stay, travel, pros, feedback, responsibilities, overworked, size, weekend
13	lack, home, cons, family, bonus, issues, processes, cool, quickly, schedules, past, fairly, super
14	senior, positive, technical, focused, cut, innovation, engineering, terms, impact, corporation, practices, trust, valued
15	company, environment, fast, technology, overall, busy, internal, stock, paced, takes, huge, advance, options
16	corporate, positions, needs, smart, structure, brand, decisions, hired, driven, remote, meetings, sell, fantastic
17	manager, store, money, excellent, leave, given, expectations, certain, exposure, depends, unrealistic, exciting, alot
18	colleagues, person, starting, treated, longer, close, leads, education, reviews, push, pro, challenge, repetitive
19	fun, days, companies, flexibility, associates, plenty, clients, daily, treat, truly, path, events, drive
20	place, health, week, hr, atmosphere, perks, program, supportive, changing, salaries, talent, college, respect
21	process, location, offer, supervisors, negative, night, holiday, meeting, stuff, minimal, awful, living, force
22	position, terrible, shifts, expect, resources, systems, stores, individual, loved, share, stability, teamwork, self
23	benefits, opportunities, nice, lots, compensation, limited, promotion, interesting, willing, knowledge, role, wage, minimum
24	career, learn, leadership, slow, market, available, extremely, professional, commission, technologies, decision, annual, bank
25	culture, learning, promoted, rate, needed, political, tasks, amazon, center, innovative, boss, peers, half
26	department, ability, compared, stress, lower, looking, workload, levels, software, diverse, late, develop, build
27	flexible, schedule, policy, security, areas, required, problem, happy, enjoy, hourly, responsibility, hike, feels
28	managers, industry, products, communication, organization, awesome, highly, cost, provide, heavy, current, break, sick
29	low, level, large, development, amazing, teams, vacation, raises, horrible, policies, chance, taking, ideas
30	team, easy, challenging, pto, global, try, members, favoritism, boring, reason, early, local, issue

Table 13: FETM Global Topics

Topic	Top Words
1	employee, kind, stressful, taking, facility, access, apply, discount, simple, cafeteria, gym, freedom, satisfaction, ton, understaffed
2	companies, paid, challenging, personal, extremely, perks, areas, package, compared, vision, respect, members, generally, helpful
3	leadership, days, week, available, rate, boss, standard, network, executive, option, perfect, senior, kept, mandatory, transparency
4	management, company, family, compensation, culture, competitive, large, position, help, especially, market, resources, travel
5	product, services, products, model, solutions, lines, chain, businesses, economy, portfolio, demand, serve, supply, worry, enterprise
6	low, atmosphere, knowledge, moving, stress, medical, local, changing, pros, path, interesting, fairly, goes, healthcare, supervisor
7	easy, communication, projects, busy, weekends, decisions, enjoy, limited, offers, build, late, night, mind, negative, spend
8	responsibilities, compensated, employ, adequately, csa, overview, doubling, delegated, modernization, overburdened, halfway, sole
9	grow, overtime, organization, salaries, shifts, direction, mobility, lead, problem, retail, wonderful, hires, upward, tough, demanding
10	based, department, job, given, performance, managers, supervisors, truly, help, feedback, hr, level, direct, fired, told
11	employees, pay, health, insurance, bonus, workers, plan, benefits, customers, decent, needs, benefit, excellent, oriented, diverse
12	scheduling, equipment, summer, hot, lay, weather, cycle, offs, trainers, winter, gaining, protection, ending, heat, relax
13	lots, job, managers, coworkers, level, bad, positions, high, structure, team, flexible, environment, teams, departments, security
14	learning, diversity, stock, takes, constantly, options, promote, tools, point, sell, meetings, play, hands, rewarded, price
15	values, individuals, core, practices, ethics, dealing, mission, integrity, ethical, slowly, followed, mentors, unethical
16	focus, major, follow, groups, encouraged, teams, feeling, executives, credit, internally, significant, treatment, seriously, card, items
17	open, policy, door, pick, greatest, setting, american, foot, completed, minded, greater, experts, laptop, orientation, suggestions
18	pandemic, retirement, likely, risk, main, considering, effective, deadlines, background, mgmt, excel, dept, contracts
19	jobs, customer, person, issues, problems, treat, hiring, possible, drive, meeting, half, regular, leads, helping, paying
20	industry, balance, learn, average, depending, hr, short, pressure, policies, roles, early, base, daily, driven, constant
21	corporate, potential, cost, cutting, cut, living, edge, forced, mention, costs, connections, speed, expenses, earning, america
22	overall, looking, free, left, chance, workplace, reason, lunch, food, manage, commission, break, center, smart, marketing
23	place, home, manager, outside, role, stay, world, starting, hire, field, internal, provide, college, line, talent
24	salary, schedule, bonuses, program, safety, difficult, extra, programs, ideas, hourly, consistent, super, recommend, car, cool
25	opportunities, advancement, career, growth, development, professional, willing, technology, systems, plenty, slow, progression
26	training, money, leave, upper, sales, process, raises, try, focused, match, goals, holidays, multiple, provided, site
27	fast, location, friendly, pto, processes, paced, project, pace, successful, lost, production, guess, dependent, operation, orders
28	vacation, weeks, changes, levels, sick, minimum, wage, individual, challenges, alot, incentives, taken, wages, yearly, financial
29	flexibility, brand, activities, ones, handle, knowledgeable, promotes, cab, budgets, thorough, logistics, invests
30	hired, highly, driving, qualified, knowing, occasionally, aware, timing, sound, grown, promise, hang, disconnect, sucked, collaborate