Variational Language Concepts for Interpreting Foundation Language Models

Hengyi Wang* **Shiwei Tan** **Zhqing Hong** **Desheng Zhang** **Hao Wang**

Department of Computer Science, Rutgers University

Abstract

Foundation Language Models (FLMs) such as BERT and its variants have achieved remarkable success in natural language processing. To date, the interpretability of FLMs has primarily relied on the attention weights in their self-attention layers. However, these attention weights only provide word-level interpretations, failing to capture higher-level structures, and are therefore lacking in readability and intuitiveness. To address this challenge, we first provide a formal definition of *conceptual interpretation* and then propose a variational Bayesian framework, dubbed VAriational Language Concept (VALC), to go beyond word-level interpretations and provide concept-level interpretations. Our theoretical analysis shows that our VALC finds the optimal language concepts to interpret FLM predictions. Empirical results on several real-world datasets show that our method can successfully provide conceptual interpretation for FLMs[1].

1 Introduction

Foundation language models (FLMs) such as BERT (Devlin et al., 2018) and its variants (Lan et al., 2019; Liu et al., 2019; He et al., 2021; Portes et al., 2023) have achieved remarkable success in natural language processing. These FLMs are usually large attention-based neural networks that follow a pretrain-finetune paradigm, where models are first pretrained on large datasets and then finetuned for a specific task. As with any machine learning models, interpretability in FLMs has always been a desideratum, especially in decision-critical applications (e.g., healthcare).

To date, FLMs' interpretability has primarily relied on the attention weights in self-attention layers. However, these attention weights only provide raw word-level importance scores as interpretations. Such low-level interpretations fail to capture higher-level semantic structures, and hence lack readability and intuitiveness. For example, low-level interpretations often fail to capture influence of similar words to predictions, leading to unstable or even unreasonable explanations (see Sec. 5.4 for details).

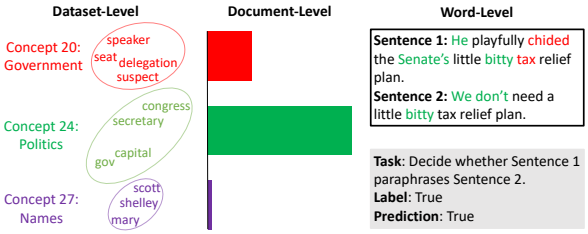

Figure 1: Visualization of VALC's learned concepts. A document consists of two sentences. The task is to decide whether 'Sentence 1' paraphrases 'Sentence 2'. **Left:** Dataset-level concepts for MRPC dataset with 3 concepts and their nearest word embeddings. **Middle:** Document-level concept strength, showing that this document is mostly related to Concept 20 and Concept 24. **Right:** Word-level concepts, where the FLM correctly predicts the label to be 'True', and VALC interprets that this is because the both sentences consist of words with Concept 24, i.e., *Politics*.

In this paper, we aim to go beyond word-level attention and interpret FLM predictions at the concept level. Such higher-level semantic interpretations are complementary to word-level importance scores and often more readable and intuitive. For example, as shown in Fig. 1, VALC interprets the FLM with the following multi-level concepts (details in Appendix F):

- **Dataset-level** concepts are highlighted by the top words and the distribution of their embeddings in the PLM (Fig. 1(left)). For example, *Concept 20 (Government)* corresponds to the red ellipse, encompassing words relevant to government entities and activities, as shown in Fig. 1(left).

*Correspondence to: Hengyi Wang <hengyi.wang@rutgers.edu>

[1]Code will soon be available at https://github.com/Wang-ML-Lab/interpretable-foundation-models

Findings of the Association for Computational Linguistics: EACL 2024, pages 8645–8671

- **Document-level** concepts are demonstrated by each document's topics; for instance, in the 3 bars representing probability distribution over 3 concepts for the document in Fig. 1(middle), VALC identifies *Concept 24*, i.e., 'politics', and *Concept 20*, i.e., 'government', as considerably more relevant concepts compared to *Concept 27*, i.e. 'names'.
- **Word-level** concepts are identified by words in documents. For example, in the box displaying the document in Fig. 1(right), VALC highlights the words 'chided' and 'tax' because they are highly related to *Concept 20*, i.e., 'government'. Terms like 'Senate' and 'bitty' are associated with *Concept 24*, i.e. 'politics', aligning with the document-level concepts.

We start by developing a comprehensive and formal definition of *conceptual interpretation* with four desirable properties: (1) multi-level structure, (2) normalization, (3) additivity, and (4) mutual information maximization. With this definition, we then propose a variational Bayesian framework, dubbed VAriational Language Concept (VALC), to provide *dataset-level*, *document-level*, and *word-level* (the first property) conceptual interpretation for FLM predictions. Our theoretical analysis shows that maximizing our VALC's evidence lower bound is equivalent to inferring the optimal conceptual interpretation with *Properties (1-3)* while maximizing the mutual information between the inferred concepts and the observed embeddings from FLMs, i.e., *Property (4)*.

Drawing inspiration from hierarchical Bayesian deep learning (Wang and Yeung, 2016, 2020; Wang et al., 2016), the core of our idea is to treat a FLM's contextual word embeddings (and their corresponding attention weights) as observed variables and build a probabilistic generative model to automatically infer the higher-level semantic structures (e.g., concepts or topics) from these embeddings and attention weights, thereby interpreting the FLM's predictions at the concept level. Our VALC is compatible with any attention-based FLMs and can work as an conceptual interpreter, which explains the FLM predictions at multiple levels with theoretical guarantees. Our contributions are as follows:

- We identify the problem of multi-level interpretations for FLM predictions, develop a formal definition of *conceptual interpretation*, and propose VALC as the first general method to infer such conceptual interpretation.
- Theoretical analysis shows that learning VALC is equivalent to inferring the optimal conceptual interpretation according to our definition.
- Quantitative and qualitative analysis on real-world datasets show that VALC can infer meaningful language concepts to effectively and intuitively interpret FLM predictions.

2 Related Work

Foundation Language Models. Foundation language models are large attention-based neural networks that follow a pretrain-finetune paradigm. Usually they are first pretrained on large datasets in a self-supervised manner and then finetuned for a specific downstream task. BERT (Devlin et al., 2018) is a pioneering FLM that has shown impressive performance across multiple downstream tasks. Following BERT, there have been variants (He et al., 2021; Clark et al., 2020; Yang et al., 2019; Liu et al., 2019; Lewis et al., 2019) that design different self-supervised learning objectives or training schemes to achieve better performance. While FLMs offer attention weights for interpreting predictions at the word level, these interpretations lack readability and intuitiveness because they fail to capture higher-level semantic structures.

Interpretation Methods for FLMs. Existing conceptual interpretation methods for FLMs typically rely on topic models (Blei et al., 2003; Blei and Lafferty, 2006; Blei, 2012; Wang et al., 2012; Chang and Blei, 2009; Mcauliffe and Blei, 2007; Hoffman et al., 2010) and prototypical part networks (Chen et al., 2019). There has been recent work that employs deep neural networks to learn topic models more efficiently (Card et al., 2017; Xing et al., 2017; Peinelt et al., 2020), using techniques such as amortized variational inference. There is also work that improves upon traditional topic models by either leveraging word similarity as a regularizer for topic-word distributions (Das et al., 2015; Batmanghelich et al., 2016) or including word embeddings into the generative process (Hu et al., 2012; Dieng et al., 2020; Bunk and Krestel, 2018; Duan et al., 2021). There is also work that builds topic models upon embeddings from FLMs (Grootendorst, 2020; Zhang et al., 2022; Wang et al., 2022; Zhao et al., 2020; Meng et al., 2022). However, these methods often rely on a pipeline involving dimensionality reduction and basic clustering, which is not end-to-end, leading to

potential information loss between FLM embeddings and clustering outcomes. This can result in *unfaithful* interpretations for the underlying FLM. Additionally, they typically generate interpretations at a single level (e.g., document level), lacking a multi-level conceptual structure.

Beyond topic models, attribution-based approaches such as LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017) assign importance to input features to explain predictions. Concept bottleneck models (CBMs) (Koh et al., 2020; Yuksekgonul et al., 2023; Yang et al., 2023; Kim et al., 2018; Schulz et al., 2020; Paranjape et al., 2020; Schrouff et al., 2021) offer interpretations by learning conceptual activation and then performing classifications on these concepts, while inherent models (Xie et al., 2023; Ren et al., 2023; Shi et al., 2021) focus on model redesign/re-training for interpretability. However, these approaches often require extra supervision or re-training, making them unsuitable for our setting. In contrast, our method is inherently multi-level and end-to-end, models *concepts* across dataset, document, and word levels, and produces faithful post-hoc interpretations for any models based on FLMs with theoretical guarantees.

3 Methods

In this section, we formalize the definition of *conceptual interpretation*, and describe our proposed VALC for conceptual interpretation of FLMs.

3.1 Problem Setting and Notation

We consider a corpus of M documents, where the m'th document contains J_m words, and a FLM $f(\mathcal{D}_m)$, which takes as input the document m (denoted as $\mathcal{D}_m$) with J_m words and outputs (1) a CLS embedding $\mathbf{c}_m \in \mathbb{R}^d$, (2) J_m contextual word embeddings $\mathbf{e}_m \triangleq [\mathbf{e}_{mj}]_{j=1}^{J_m}$, and (3) the attention weights $\mathbf{a}_m^{(h)} \triangleq [a_{mj}^{(h)}]_{j=1}^{J_m}$ between each word and the last-layer CLS token, where h denotes the h'th attention head. We denote the average attention weight over H heads as $a_{mj} = \frac{1}{H}\sum_{h=1}^{H} a_{mj}^{(h)}$ and correspondingly $\mathbf{a}_m \triangleq [a_{mj}]_{j=1}^{J_m}$ (see the FLM at the bottom of Fig. 2). In FLMs, these last-layer CLS embeddings are used as document-level representations for downstream tasks (e.g., document classification). Furthermore, our VALC assumes K concepts (topics) for the corpus. For document m, our VALC interpreter tries to infer a concept distribution vector $\boldsymbol{\theta}_m \in \mathbb{R}^K$ (also known as the topic proportion in topic models) for the whole document and a concept distribution vector $\boldsymbol{\phi}_{mj} = [\phi_{mjk}]_{k=1}^{K} \in \mathbb{R}^K$ for word j in document m. In our continuous embedding space, the k'th concept is represented by a Gaussian distribution, $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, of contextual word embeddings; we use shorthand $\boldsymbol{\Omega}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ for brevity. The goal is to interpret FLMs' predictions *at the concept level* using the inferred document-level concept vector $\boldsymbol{\theta}_m$, word-level concept vector $\boldsymbol{\phi}_{mj}$, and the learned embedding distributions $\{\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^{K}$ for each concept (see Sec. 5.4 for detailed descriptions and visualizations).

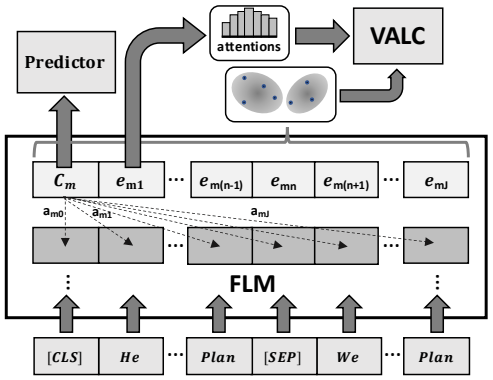

Figure 2: Overview of VALC framework.

3.2 Formal Definition of Language Concepts

Below we formally define 'conceptual interpretation' for FLM predictions (see notations in Sec. 3.1):

Definition 3.1 (Conceptual Interpretation). Assume K concepts and a dataset $\mathcal{D}$ containing M documents, each with J_m words ($1 \le m \le M$). Conceptual interpretation for a document m consists of K *dataset-level* variables $\{\boldsymbol{\Omega}_k\}_{k=1}^{K}$, a *document-level* variable $\boldsymbol{\theta}_m$, and J_m *word-level* variables $\{\boldsymbol{\phi}_{mj}\}_{j=1}^{J_m}$ with the following properties:

(1) **Multi-Level Structure.** Conceptual interpretation has a three-level structure:
   (a) Each *dataset-level* variable $\boldsymbol{\Omega}_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ describes the k'th concept; $\boldsymbol{\mu}_k \in \mathbb{R}^d$ and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{d\times d}$ denote the mean and covariance of the k'th concept in the embedding space (i.e., $\mathbf{e}_{mj} \in \mathbb{R}^d$), respectively.
   (b) Each *document-level* variable $\boldsymbol{\theta}_m \in \mathbb{R}_{\ge 0}^K$ describes document m's relation to the K concepts.
   (c) Each *word-level* variable $\boldsymbol{\phi}_{mj} \in \mathbb{R}_{\ge 0}^K$ describes word j's relation to the K concepts.

(2) **Normalization.** The document- and word-

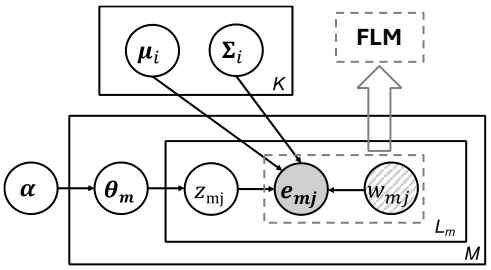

Figure 3: Graphical model of our VALC. The *striped* circle represents *continuous* word counts.

level interpretations, $\boldsymbol{\theta}_m$ and $\boldsymbol{\phi}_{mj}$, are normalized:

(a) $\sum_{k=1}^{K} \theta_{mk} = 1$ for document m.

(b) $\sum_{k=1}^{K} \phi_{mjk} = 1$ for word j in document m.

(3) **Additivity.** We can add/subtract the k's concept from the contextual embeddings $\mathbf{e}_{mj}$ of word j in document m, i.e., $\mathbf{e}_{mj} \leftarrow \mathbf{e}_{mj} \pm x_k \boldsymbol{\mu}_k$ (x_k is the editing weight of concept k).

(4) **Mutual Information Maximization.** The conceptual interpretation achieves maximum mutual information between the observed embeddings $\mathbf{e}_m$ in FLMs and the document-level/word-level interpretation, $\boldsymbol{\theta}_m$ and $\boldsymbol{\phi}_{mj}$.

In Definition 3.1, Property (1) provides comprehensive three-level conceptual interpretation for FLM predictions, Property (2) ensures proper normalization in concept assignment at the document and word levels, Property (3) enables better concept editing (more details in Sec. 5.3) to modify FLM predictions, and Property (4) ensures minimal information loss when interpreting FLM predictions.

3.3 VAriational Language Concepts (VALC)

Method Overview. Drawing inspiration from hierarchical Bayesian deep learning (Wang and Yeung, 2016, 2020; Wang et al., 2016; Mao et al., 2022; Yan and Wang, 2023; Xu et al., 2023; Wang et al., 2024), we propose our model, VAriational Language Concepts (VALC), to infer the optimal conceptual interpretation described in Definition 3.1. Different from *static* word embeddings (Mikolov et al., 2013) and topic models, FLMs produce *contextual* word embeddings with continuous-value entries $[\mathbf{e}_{mj}]_{j=1}^{J_m}$ and more importantly, associate each word embedding with a continuous-value attention weight $[a_{mj}]_{j=1}^{J_m}$; therefore this brings unique challenges.

To effectively discover latent concept structures learned by FLMs at the dataset level and interpret FLM predictions at the data-instance level, our VALC treats both the contextual word embeddings and their associated attention weights as observations to learn a probabilistic generative model of these observations, as shown in Fig. 2. The key idea is to use the attention weights from FLMs to compute a virtual continuous count for each word, and model the contextual word embedding distributions with Gaussian mixtures. The generative process of VALC is as follows (we mark key connection to FLMs in blue and show the corresponding graphical model in Fig. 3):

For each document $m, 1 \leq m \leq M$,

1. Draw the document-level concept distribution vector $\boldsymbol{\theta}_m \sim \text{Dirichlet}(\boldsymbol{\alpha})$.
2. For each word j $(1 \leq j \leq J_m)$,
   (a) Draw the word-level concept index $z_{mj} \sim \text{Categorical}(\boldsymbol{\theta}_m)$.
   (b) With a continuous word count $w_{mj} \in \mathbb{R}$ from the FLM's attention weights, draw the contextual word embedding of the FLM from the corresponding Gaussian component $\mathbf{e}_{mj} \sim \mathcal{N}(\boldsymbol{\mu}_{z_{mj}}, \boldsymbol{\Sigma}_{z_{mj}})$.

Given the generative process above, discovery of latent concept structures in FLMs at the dataset level boils down to learning the parameters $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ for the K concepts. Intuitively the global parameters $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K}$ are shared across different documents, and they define a mixture of K Gaussian distributions. Each Gaussian distribution describes a 'cluster' of words and their contextual word embeddings.

Similarly, interpretations of FLM predictions at the data-instance level is equivalent to inferring the latent variables, i.e., document-level concept distribution vectors $\boldsymbol{\theta}_m$ and word-level concept indices z_{mj}. Below we highlight several important aspects of our VALC designs.

Attention Weights as *Continuous* Word Counts. Different from typical topic models (Blei et al., 2003; Blei, 2012) and word embeddings (Mikolov et al., 2013) that can only handle *discrete* word counts, our VALC can handle *continuous* (virtual) word counts; this better aligns with continuous attention weights in FLMs. Specifically, we denote as $w_{mj} \in \mathbb{R}_{\geq 0}$ the (non-negative real-valued) *continuous word count* for the j'th word in document m. We explore three schemes of computing w_{mj}:

- **Identical Weights:** Use identical weights for different words, i.e., $w_{mj} = 1, \forall m, j$. This is equivalent to typical discrete word counts.
- **Attention-Based Weights with Fixed Length:**

Use $w_{mj} = J' a_{mj}$, where J' is a fixed sequence length shared across all documents.

- **Attention-Based Weights with Variable Length:** Use $w_{mj} = {}^{J_m a_{mj}}/_{\sum_{i=1}^{J_m} a_{mi}}$, where J_m is true sequence length without padding. Note that in practice, $\sum_{i=1}^{J_m} a_{mi} \neq 1$ due to padding tokens in FLMs.

Contextual Continuous Word Representations. Note that different from topic models (Blei et al., 2003) and typical word embeddings (Mikolov et al., 2013; Dieng et al., 2020) where word representations are *static*, word representations in FLMs are *contextual*; specifically, the same word can have different embeddings in different documents (contexts). For example, the word 'soft' can appear as the j_1'th word in document m_1 and as the j_2'th word in document m_2, and therefore have two different embeddings (i.e., $\mathbf{e}_{m_1 j_1} \neq \mathbf{e}_{m_2 j_2}$).

Correspondingly, in our VALC, we do not constrain the same word to have a static embedding; instead we assume that a word embedding is drawn from a Gaussian distribution corresponding to its latent topic. Note that word representations in our VALC is continuous, which is different from typical topic models (Blei et al., 2003) based on (discrete) bag-of-words representations.

3.4 Objective Function

Below we discuss the inference and learning procedure for VALC. We start by introducing the *inference* of document-level and word-level concepts (i.e., z_{mj} and $\boldsymbol{\theta}_m$) given the global concept parameters (i.e., $\{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^K$), and then introduce the *learning* of these global concept parameters.

3.4.1 Inference

Inferring Document-Level and Word-Level Concepts. We formulate the problem of interpreting FLM predictions at the concept level as inferring document-level and word-level concepts. Specifically, given global concept parameters $\{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^K$, the *contextual* word embeddings $\mathbf{e}_m \triangleq [\mathbf{e}_{mj}]_{j=1}^{J_m}$, and the associated attention weights $\mathbf{a}_m \triangleq [a_{mj}]_{j=1}^{J_m}$, a FLM produces for each document m, our VALC infers the posterior distribution of the document-level concept vector $\boldsymbol{\theta}_m$, i.e., $p(\boldsymbol{\theta}_m | \mathbf{e}_m, \mathbf{a}_m, \{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^K)$, and the posterior distribution of the word-level concept index z_{mj}, i.e., $p(z_{mj} | \mathbf{e}_m, \mathbf{a}_m, \{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^K)$.

Variational Distributions. These posterior distributions are intractable; we therefore resort to variational inference (Jordan et al., 1998; Blei et al., 2003) and use variational distributions $q(\boldsymbol{\theta}_m | \boldsymbol{\gamma}_m)$ and $q(z_{mj} | \boldsymbol{\phi}_{mj})$ to approximate them. Here $\boldsymbol{\gamma}_m \in \mathbb{R}^K$ and $\boldsymbol{\phi}_{mj} \triangleq [\phi_{mjk}]_{k=1}^K \in \mathbb{R}^K$ are variational parameters to be estimated during inference. This leads to the following joint variational distribution:

$$\begin{aligned} & q(\boldsymbol{\theta}_m, \{\mathbf{z}_{mj}\}_{j=1}^{J_m} | \boldsymbol{\gamma}_m, \{\boldsymbol{\phi}_{mj}\}_{j=1}^{J_m}) \\ = \; & q(\boldsymbol{\theta}_m | \boldsymbol{\gamma}_m) \cdot \prod_{j=1}^{J_m} q(z_{mj} | \boldsymbol{\phi}_{mj}). \end{aligned} \tag{1}$$

Evidence Lower Bound. For each document m, finding the optimal variational distributions is then equivalent to maximizing the following evidence lower bound (ELBO):

$$\begin{aligned} & \mathcal{L}(\boldsymbol{\gamma}_m, \{\boldsymbol{\phi}_{mj}\}_{j=1}^{J_m}; \boldsymbol{\alpha}, \{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^K) \\ = \; & \mathbb{E}_q[\log p(\boldsymbol{\theta}_m | \boldsymbol{\alpha})] + \sum_{j=1}^{J_m} \mathbb{E}_q[\log p(z_{mj} | \boldsymbol{\theta}_m)] \\ & + \sum_{j=1}^{J_m} \mathbb{E}_q[\log p(\mathbf{e}_{mj} | z_{mj}, \boldsymbol{\mu}_{z_{mj}}, \boldsymbol{\Sigma}_{z_{mj}})] \\ & - \mathbb{E}_q[\log q(\boldsymbol{\theta}_m)] - \sum_{j=1}^{J_m} \mathbb{E}_q[\log q(z_{mj})], \end{aligned} \tag{2}$$

where the expectation is taken over the joint variational distribution in Eq. 1.

Likelihood with *Continuous* Word Counts. One key difference between VALC and typical topic models (Blei et al., 2003; Blei, 2012) is the virtual continuous (real-valued) word counts (discussed in Sec. 3.3). Specifically, we define the likelihood in the third term of Eq. 2 as:

$$p(\mathbf{e}_{mj} | z_{mj}, \boldsymbol{\mu}_{z_{mj}}, \boldsymbol{\Sigma}_{z_{mj}}) = [\mathcal{N}(\mathbf{e}_{mj}; \boldsymbol{\mu}_{mj}, \boldsymbol{\Sigma}_{mj})]^{w_{mj}}. \tag{3}$$

Note that Eq. 3 is the likelihood of w_{mj} (virtual) words, where w_{mj} is a real value derived from the FLM's attention weights (details in Sec. 3.3). Therefore, in the third item of Eq. 2, we have:

$$\begin{aligned} & \mathbb{E}_q[\log p(\mathbf{e}_{mj} | z_{mj}, \boldsymbol{\mu}_{z_{mj}}, \boldsymbol{\Sigma}_{z_{mj}})] \\ = \; & \sum\nolimits_k \phi_{mjk} w_{mj} \log \mathcal{N}(\mathbf{e}_{mj} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\ = \; & \sum\nolimits_k \phi_{mjk} w_{mj} \{ -\frac{1}{2} (\mathbf{e}_{mj} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{e}_{mj} - \boldsymbol{\mu}_k) \\ & - \log[(2\pi)^{d/2} |\boldsymbol{\Sigma}_k|^{1/2}] \}. \end{aligned} \tag{4}$$

Update Rules. Taking the derivative of the ELBO in Eq. 2 w.r.t. ϕ_{mjk} (see Appendix A for details) and setting it to 0 yields the update rule for ϕ_{mjk}:

$$\begin{aligned} \phi_{mjk} \propto \; & \frac{w_{mj}}{|\boldsymbol{\Sigma}_k|^{1/2}} \exp[\Psi(\gamma_{mk}) - \Psi(\sum\nolimits_{k'} \gamma_{mk'}) \\ & - \frac{1}{2} (\mathbf{e}_{mj} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{e}_{mj} - \boldsymbol{\mu}_k)], \end{aligned} \tag{5}$$

with the normalization constraint $\sum_{k=1}^K \phi_{mjk} = 1$.

$$\gamma_{mk} = \alpha_k + \sum\nolimits_{j=1}^{J_m} \phi_{mjk} w_{mj}, \tag{6}$$

where $\boldsymbol{\alpha} \triangleq [\alpha_k]_{k=1}^K$ is the hyperparameter for the Dirichlet prior distribution of $\boldsymbol{\theta}_m$. In summary, the inference algorithm will alternate between updating ϕ_{mjk} for all (m, j, k) tuples and updating $\boldsymbol{\gamma}_{mk}$ for all (m, k) tuples.

3.4.2 Learning

Learning Dataset-Level Concept Parameters. The inference algorithm in Sec. 3.4.1 assumes availability of the dataset-level (global) concept parameters $\{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^K$. To learn these parameters, one needs to iterate between (1) inferring document-level variational parameters $\boldsymbol{\gamma}_m$ as well as word-level variational parameters $\boldsymbol{\phi}_{mj}$ in Sec. 3.4.1 and (2) learning dataset-level concept parameters $\{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^K$.

Update Rules. Similar to Sec. 3.4.1, we expand the ELBO in Eq. 2 (see Appendix A for details) and set its derivative w.r.t. $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ to $\mathbf{0}$, yielding the update rule for learning $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$:

$$\begin{aligned} \boldsymbol{\mu}_k &= \frac{\sum_{m,j} \phi_{mjk} w_{mj} \mathbf{e}_{mj}}{\sum_{m,j} \phi_{mjk} w_{mj}}, \\ \boldsymbol{\Sigma}_k &= \frac{\sum_{m,j} \phi_{mjk} w_{mj} (\mathbf{e}_{mj} - \boldsymbol{\mu}_k)(\mathbf{e}_{mj} - \boldsymbol{\mu}_k)^T}{\sum_{m,j} \phi_{mjk} w_{mj}}. \end{aligned} \tag{7}$$

Algorithm 1: Algorithm for VALC

Input: Initialized $\{\boldsymbol{\gamma}_m\}_{m=1}^M$, $\{\boldsymbol{\phi}_m\}_{m=1}^M$, and $\{\boldsymbol{\Omega}_k\}_{k=1}^K$, documents $\{\mathcal{D}_m\}_{m=1}^M$, number of epochs T.
for $t = 1 : T$ **do**
  for $m = 1 : M$ **do**
    Update $\boldsymbol{\phi}_m$ and $\boldsymbol{\gamma}_m$ using Eq. 5 and Eq. 6, respectively.
  Update $\{\boldsymbol{\Omega}_k\}_{k=1}^K$ using Eq. 7.

Effect of Attention Weights. From Eq. 7, we can observe that the attention weight of the j'th word in document m, i.e., a_{mj}, affects the virtual continuous word count w_{mj} (see Sec. 3.3), thereby affecting the update of the dataset-level concept center $\boldsymbol{\mu}_k$ and covariance $\boldsymbol{\Sigma}_k$. Specifically, if we use attention-based weights with fixed length or variable length in Sec. 3.3, the continuous word count w_{mj} will be proportional to the attention weight a_{mj}. Therefore, when updating the concept center $\boldsymbol{\mu}_k$ as a weighted average of different word embeddings $\mathbf{e}_{mj}$, VALC naturally places more focus on words with higher attention weights a_{mj} from FLMs, thereby making the interpretations sharper (see Sec. 5.4 for detailed results and Appendix I for theoretical analysis).

Algorithm 2: Algorithm for VALC Concept Editing

Input: FLM $f(\cdot)$, classifier $g(\cdot)$, classification loss L, document $\mathcal{D}_m$ with J_m words, labels $\mathbf{y}$, constant factor ω.
for $j = 1 : J_m$ **do** $\quad \mathbf{e}_{mj} = f(\mathcal{D}_{mj})$
  $\mathbf{x}^* = QP(\mathbf{e}_{mj}, \{\boldsymbol{\mu}_k\}_{k=1}^K)$
  $k^* = \arg\min L(g(\mathbf{e}_{mj} - \omega \cdot x_k^* \boldsymbol{\mu}_k), y_m)$
  $\mathbf{e}_{mj} \leftarrow \mathbf{e}_{mj} - \omega \cdot x_{k^*}^* \boldsymbol{\mu}_{k^*}$

4 Theoretical Analysis

In this section, we provide theoretical guarantees of VALC on the four properties in Definition 3.1.

Multi-Level Structure. As shown in Alg. 1, VALC (1) learns the *dataset-level* interpretation $\{\boldsymbol{\Omega}_k\}_{k=1}^K$ describing the K concepts, (2) infers the distribution of *document-level* interpretation $\boldsymbol{\theta}_m$ for document m, i.e., $q(\boldsymbol{\theta}_m|\boldsymbol{\gamma}_m)$ (parameterized by $\boldsymbol{\gamma}_m$), and (3) infers the posterior distribution of *word-level* concept index, i.e., $q(z_{mj}|\boldsymbol{\phi}_{mj})$, parameterized by $\boldsymbol{\phi}_{mj}$. Such three-level interpretations correspond to Property (1) in Definition 3.1.

Normalization. The learned variational distribution $q(\boldsymbol{\theta}_m|\boldsymbol{\gamma}_m)$ (described in Eq. 1) is a Dirichlet distribution; therefore we have $\sum_{k=1}^K \theta_{mk} = 1$. The update of $\boldsymbol{\phi}_{mj}$ (Eq. 5) is naturally constrained by $\sum_{k=1}^K \phi_{mjk} = 1$ since $\boldsymbol{\phi}_{mj}$ parameterizes a Categorical distribution (over z_{mj}).

Additivity. VALC is able to perform ***Concept Editing***, i.e, add/subtract the learned concept activation μ_k from FLMs via the following Quadratic Programming (QP) problem ($\mathbf{x} = [x_k]_{k=1}^K$):

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^K} \quad & \| \sum_{k=1}^K x_k \boldsymbol{\mu}_k - \mathbf{e}_m \|^2, \\ \text{subject to} \quad & \mathbf{x} \geq \mathbf{0} \text{ and } \sum_{k=1}^K x_k = 1. \end{aligned} \tag{8}$$

Given learned concepts $\{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^K$, VALC obtains this QP's optimal solution $\mathbf{x}^* \in \mathbb{R}^K$ and add/subtract any concept k from arbitrary FLM embedding $\mathbf{e}_m$ by: $\mathbf{e}_m \leftarrow \mathbf{e}_m \pm x_k^* \boldsymbol{\mu}_k$. Alg. 2 summarizes this *concept editing* process; one can also replace $\mathbf{e}_{mj}$ with the CLS embedding $\mathbf{c}_m$ for document-level editing (details in Appendix D).

Mutual Information Maximization. Theorem 4.1 below shows that our inferred document-level and word-level interpretation, $\boldsymbol{\theta}_m$ and $\{\boldsymbol{\phi}_{mj}\}_{j=1}^{J_m}$, satisfy Property (4), Mutual Information Maximization, in Definition 3.1.

Theorem 4.1 (Mutual Information Maximization). *In Eq. 2, the ELBO* $\mathcal{L}(\boldsymbol{\gamma}_m, \{\boldsymbol{\phi}_{mj}\}_{j=1}^{J_m}; \boldsymbol{\alpha}, \{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^{K})$ *is upper bounded by the mutual information between contextual embeddings* $\mathbf{e}_m$ *and multi-level interpretation* $\boldsymbol{\theta}_m, \{\boldsymbol{\phi}_{mj}\}_{j=1}^{J_m}$ *in Definition 3.1. Formally, with approximate posteriors* $q(\boldsymbol{\theta}_m|\boldsymbol{\gamma}_m)$ *and* $q(z_{mj}|\boldsymbol{\phi}_{mj})$*, we have*

$$\begin{aligned} &\mathcal{L}(\boldsymbol{\gamma}_m, \{\boldsymbol{\phi}_{mj}\}_{j=1}^{J_m}; \boldsymbol{\alpha}, \{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^{K}) \\ &\leq I(\mathbf{e}_m; \boldsymbol{\theta}_m, \{z_{mj}\}_{j=1}^{J_m}) - H(\mathbf{e}_m), \end{aligned} \tag{9}$$

where the entropy term $H(\mathbf{e}_m)$ *is a constant.*

From Theorem 4.1 we can see that maximizing the ELBO in Eq. 2 is equivalent to maximizing the mutual information between our document-level/word-level concepts and the observed contextual embeddings in FLMs (proof in Appendix H).

In summary, VALC enjoys all four properties in Definition 3.1 and therefore generates the optimal conceptual interpretation for FLMs. In contrast, state-of-the-art methods only satisfy a small part of them (Table 1 and Sec. 5.2). In Appendix I, we provide theoretical guarantees that (1) under mild assumptions our VALC can learn better conceptual interpretations for FLMs for in noisy data and (2) attention-based schemes is superior to the identical scheme (described in Sec. 3.3).

5 Experiments

5.1 Experiment Setup

Datasets. We use three datasets in our experiments, namely 20 Newsgroups, M10 (Lim and Buntine, 2015), and BBC News (Greene and Cunningham, 2006). For preprocessing details, see Appendix C.

Baselines. We compare our method with the following state-of-the-art baselines:

- **SHAP and LIME** (Lundberg and Lee, 2017; Ribeiro et al., 2016) are interpretation methods that attribute importance scores to input features. In this paper, we use embeddings of 'CLS' token as input to SHAP/LIME.
- **BERTopic** (Grootendorst, 2020) is a clustering-based model that uses HDBSCAN (McInnes and Healy, 2017) to cluster sentence embeddings from BERT, performs Uniform Manifold Approximation Projection (UMAP) (McInnes et al., 2018), and then uses class-based TF-IDF (c-TF-IDF) to obtain words for each cluster.

Table 1: Comparing methods on the properties in Definition 3.1 (MIM: Mutual Information Maximization).

Model	Multi-Level	Normalization	Additivity	MIM
SHAP/LIME	No	No	Partial	No
BERTopic	No	Hard	Partial	No
CETopic	No	Hard	Partial	No
VALC	**Yes**	**Soft**	**Full**	**Yes**

- **CETopic** (Zhang et al., 2022) is a clustering-based model that first uses UMAP to perform dimensionality reduction on BERT sentence embeddings, performs K-Means clustering (Lloyd, 1982), and then uses weighted word selection for each cluster.

Evaluation Metric. Inspired by Koh et al. (2020), we perform concept editing experiments to evaluate conceptual interpretation for FLMs; higher *accuracy gain* after editing indicates better interpretation performance. We leverage BERT-base-uncased (Devlin et al., 2018) as the contextual embedding model, and use accuracy on the test set as our metric. For a fair comparison, we adhered to the baseline methodologies (e.g., BERTopic and CETopic) by setting the number of concepts (topics) K to 100 across all datasets. This number was chosen as it strikes an effective balance between capturing adequate detail and avoiding model overfitting. See Appendix D for more details.

We can perform concept editing on either input tokens or contextual embeddings of FLMs. Specifically, we can perform *hard* concept editing for concept k by directly removing tokens that belong concept k (applicable for hard clustering methods such as our baselines); we could also perform *soft* concept editing for concept k by removing concept subspace vectors from contextual embeddings $\mathbf{e}_m$ (applicable for VALC using Alg. 2).

Following (Lyu et al., 2024), we conducted additional experiments to evaluate the faithfulness metric. The faithfulness metric is implemented as the accuracy score of predictions using logistic regression, with the inferred conceptual explanations as inputs.

5.2 Comparison on Four Properties in Definition 3.1

In Sec. 4 we show that VALC satisfies the four properties of conceptual interpretation in Definition 3.1. In contrast, baseline models do not necessarily learn concepts that meet these requirements. Table 1 summarizes the comparison between VALC and the baselines. We can see that VALC is superior

Table 2: **Accuracy gain on 20 Newsgroups (20NG), M10, and BBC News (BBC) (%).** We mark the best results with **bold face** and the second best with underline.

Dataset		Unedited	SHAP /LIME	BERTopic	CETopic	VALC	Finetune (Oracle)
20NG	Acc.	51.26	61.74	60.76	61.93	**62.54**	64.38
	Gain	-	10.48	9.50	10.67	**11.28**	13.12
M10	Acc.	69.74	75.60	76.79	79.18	**80.74**	82.54
	Gain	-	5.86	7.05	9.44	**11.00**	12.80
BBC	Acc.	93.72	95.96	95.52	**96.86**	96.41	97.76
	Gain	-	2.24	1.80	**3.14**	2.69	4.04

to baselines in terms of the following four aspects:

(1) **Multi-Level Structure.** Baselines either apply clustering algorithms directly on the document-level embeddings from FLMs or assign importance scores to input features, and thus can only provide single-level interpretation, necessitating complex post-processing to generate dataset-level concepts. In contrast, VALC adopts an integrated approach, learning concepts at the dataset, document, and word level in a joint, end-to-end manner.

(2) **Normalization.** BERTopic and CETopic assign each word to exactly one concept and therefore satisfies *hard*-normalization. SHAP/LIME produce importance scores that are not normalized. In contrast, VALC learns fractional concept interpretations γ_m and ϕ_{mj} and therefore satisfies *soft*-normalization, which is more flexible and intuitive.

(3) **Additivity.** Baselines perform addition or subtraction of concepts only at a single level (word/document), while our additivity and concept editing (Alg. 2) work for both levels.

(4) **Mutual Information Maximization.** Baselines either use a multi-step pipeline or produce importance scores; they are therefore prone to lose information between FLM embeddings and final clustering/scoring results. In contrast, VALC is theoretically guaranteed to maximally preserve information (Theorem 4.1).

5.3 Results

Accuracy Gain. We perform greedy concept editing (Koh et al., 2020) for BERTopic, CETopic, and our VALC to evaluate the quality of their learned concepts. Higher accuracy gain after pruning indicates better performance.

Table 2 show the results for different methods in three real-world datasets, where 'Finetune (Oracle)' refers to finetuning both the backbone and the classifier of BERT. VALC's concept editing can improve the accuracy upon the unedited model by more than 11% in 20 Newsgroups and M10, almost on par with 'Finetune (Oracle)'. Compared with the baselines, VALC achieves the most accuracy gain in 20 Newsgroups and M10 and the second most accuracy gain in BBC News, demonstrating the effectiveness of VALC's four properties in Definition 3.1. Note that SHAP and LIME both interpret the CLS token's embedding and therefore has identical accuracy gain (details in Appendix D).

Table 3: **VALC Editing Accuracy (%).** We mark the best results with **bold face**, second best with underline.

Dataset	Unedited	Random	Unweighted	Weighted	Finetune (Oracle)
20 Newsgroups	51.26	51.13	54.63	**62.54**	64.38
M10	69.74	69.76	73.56	**80.74**	82.54
BBC News	93.72	93.72	95.52	**96.41**	97.76

Ablation Study. Thanks to its full additivity (Definition 3.1), VALC is capable of different concept editing schemes, including 'Random', 'Unweighted', and 'Weighted'. Specifically, *weighted* pruning uses the concept editing algorithm in Alg. 2 with the optimal hyperparameter ω; *unweighted* pruning runs Alg. 2 with $\omega = 1$; *random* pruning first randomly picks a concept k ($k \in \{1, ..., K\}$), sets $\omega \cdot x_k = 1/K$, and then runs Alg. 2. Table 3 shows accuracy for VALC's different schemes. As expected, random pruning barely improves upon the unedited model. Unweighted pruning improves upon the unedited model by $1.5 \sim 3.5\%$. Weighted pruning improves the accuracy by around 11% upon the unedited model on 20 Newsgroups and M10.

Faithfulness. Table 4 shows the faithfulness of VALC and baselines on the 20 Newsgroups, M10, and BBC News datasets. These results show that our VALC significantly outperforms the baseline models, achieving the highest faithfulness accuracy scores in the 20 Newsgroups (89.8%), M10 (99.5%), and BBC News (100.0%) datasets.

Note that the dataset size of 20 Newsgroups, M10, and BBC News is 16,309, 8,355, and 2,225, respectively. BBC News contains significantly less data, making it easier to achieve a high faithfulness score. This explains why both CETopic and our VALC obtain a faithfulness score of 100.0%.

Baseline methods such as BERTopic and CETopic represent language concepts as discrete bags of words, which lack flexibility and accuracy. In contrast, VALC infers continuous concepts for datasets, documents, and words with theoretical guarantees. Consequently, it provides optimal and faithful conceptual explanations of high quality.

See Appendix G for more quantitative results.

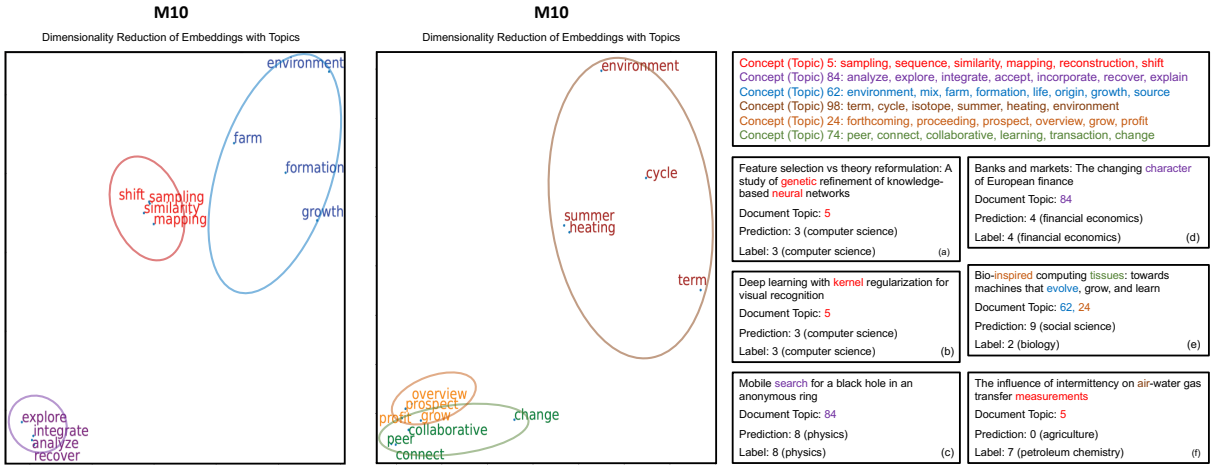

Figure 4: Visualization of VALC's three-level conceptual interpretation. **Left and Middle:** Dataset-level interpretation with 6 concepts' $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ with nearest word embeddings (3 concepts per plot for clarity). **Right:** Top words in each concept and 6 example documents with the associated document-level and word-level interpretations.

Table 4: Additional results for the faithfulness (in terms of accuracy percentage (%)) of VALC and baselines on the 20 Newsgroups, M10, and BBC News datasets. We mark the best results with **bold face**.

Methods	20 NG	M10	BBC	Average
SHAP/LIME	5.8	13.9	22.9	14.2
BERTopic	17.2	87.6	64.6	56.5
CETopic	79.2	96.4	**100.0**	91.9
VALC	**89.8**	**99.5**	**100.0**	**96.4**

5.4 Conceptual Interpretation (More for Different Tasks in Appendix F)

Dataset-Level Interpretations. As a case study, we train VALC on M10, sample 6 concepts (topics) from the dataset, and plot the word embeddings of the top words (closest to the center $\boldsymbol{\mu}_k$) in these concepts using PCA in Fig. 4(left and middle). We can observe Concept 5 is mostly about data analysis, including words such as 'sampling' and 'similarity'. Concept 84 is mostly about reasoning, with words 'explore', 'accept', 'explain', etc. Concept 62 is mostly about nature, with words 'environment', 'formation', 'growth', etc. Concept 98 is mostly about farming, with words 'term', 'summer', 'heating', etc. Concept 24 is mostly about economics, with words 'forthcoming', 'prospect', 'grow', etc. Concept 74 is mostly about social contact, containing words such as 'peer', 'connect', and 'collaborative'. Interestingly, Concept 24 (economics) and Concept 74 (social contact) are both related to social science and are therefore closer to each other in Fig. 4(middle), while Concept 98 (farming) is farther away, showing VALC's cability of capturing concept similarity.

Document-Level Interpretations. Fig. 4(right) shows that VALC can provide conceptual interpretations on why correct or incorrect FLM predictions happen for specific documents. For example, document (e) belongs to class 2 (**biology**), but BERT misclassifies it as class 9 (**social science**); our VALC interprets that this is because document (e) involves Concept 24 (economics), which is related to **social science**. On the other hand, document (b) is related to machine learning and BERT correctly classifies it as class 3 (**computer science**); VALC interprets that this is because document (b) involves Concept 5 (data analysis).

Word-Level Interpretations. Fig. 4(right) also shows that VALC can interpret which words and what concepts of these words lead to specific FLM predictions. For example, document (f) belongs to class 7 (**petroleum chemistry**), but BERT misclassifies it as class 0 (**agriculture**); VALC attributes this to the word 'air', which belongs to Concept 98 (farming). For document (b), VALC interprets that BERT correctly classifies it as class 3 (**computer science**) because the document contains the word 'kernel' that belongs to Concept 5 (data analysis).

6 Conclusion

We address the challenge of multi-level interpretations for FLM predictions by defining conceptual interpretation and introducing VALC, the first method to infer such interpretations effectively. Empirical results are promising, and theoretical analysis confirms that VALC reliably produces optimal conceptual interpretations by our definition.

7 Limitations

Our proposed method assumes access to the hidden layers of Transformer-based models, and therefore can be naturally extended to Transformer-based models including RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2021), ALBERT (Lan et al., 2019), Electra (Clark et al., 2020), and decoder-only models, such as GPTs (Radford et al., 2019; Brown, 2020). Although our VALC is initially designed for Transformer-based models, it is also generalizable to other architectures, such as Convolutional Neural Networks (CNNs) (LeCun et al., 2015) and Long Short-Term Memory networks (LSTMs) (Hochreiter and Schmidhuber, 1997), by simply setting identical attention weights. Future work may include extending VALC beyond Transformer variants and natural language applications. However, many other foundation language models provided by proprietary sources may not expose their internal states, limiting the applicability of our method in such cases.

8 Ethical Considerations

VALC, as the first to comprehensively interpret FLMs at the concept level, holds significant promise for advancing societal and technological progress. By elucidating the inner workings of these complex FLMs, we enable greater transparency and trust in AI systems, which is crucial for their widespread adoption. This transparency ensures that AI-driven decisions in critical areas such as healthcare, law, and finance are more explainable and accountable, thus safeguarding against biases and errors. Additionally, our VALC fosters enhanced collaboration between AI and human experts, as interpretable models can provide insights that are more easily understood and acted upon by domain specialists. This symbiotic relationship has the potential to accelerate innovation, improve decision-making processes, and ultimately lead to more ethical and equitable AI applications, thereby benefiting society at large.

9 Acknowledgements

We extend our sincere gratitude to Akshay Nambi and Tanuja Ganu from Microsoft Research for their invaluable insights and guidance, which significantly enhanced the quality of this work. We also express our deep appreciation to Microsoft Research AI & Society Fellowship, NSF Grant IIS-2127918, NSF CAREER Award IIS-2340125, NIH Grant 1R01CA297832, and the Amazon Faculty Research Award for their generous support. This research is also supported by NSF National Artificial Intelligence Research Resource (NAIRR) Pilot and the Frontera supercomputer supported by the National Science Foundation (award NSF-OAC 1818253) at the Texas Advanced Computing Center (TACC) at The University of Texas at Austin. Additionally, we thank the anonymous reviewers and the area chair/senior area chair for their thoughtful feedback and for recognizing the significance and contributions of our research. Finally, we would like to thank the Center for AI Safety (CAIS) for providing the essential computing resources that enabled this work.

References

Kayhan Batmanghelich, Ardavan Saeedi, Karthik Narasimhan, and Sam Gershman. 2016. Nonparametric spherical topic modeling with word embeddings. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2016, page 537. NIH Public Access.

David M Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

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

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Stefan Bunk and Ralf Krestel. 2018. Welda: enhancing topic models by incorporating local word context. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, pages 293–302.

Dallas Card, Chenhao Tan, and Noah A Smith. 2017. A neural framework for generalized topic models. *arXiv preprint arXiv:1705.09296*.

Jonathan Chang and David Blei. 2009. Relational topic models for document networks. In *Artificial intelligence and statistics*, pages 81–88. PMLR.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian lda for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 795–804.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Zhibin Duan, Dongsheng Wang, Bo Chen, Chaojie Wang, Wenchao Chen, Yewen Li, Jie Ren, and Mingyuan Zhou. 2021. Sawtooth factorial topic embeddings guided gamma belief network. In *International Conference on Machine Learning*, pages 2903–2913. PMLR.

Derek Greene and Pádraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd international conference on Machine learning*, pages 377–384.

Maarten Grootendorst. 2020. Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics. *Zenodo, Version v0*, 9.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *ICLR*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Matthew Hoffman, Francis Bach, and David Blei. 2010. Online learning for latent dirichlet allocation. *advances in neural information processing systems*, 23:856–864.

Pengfei Hu, Wenju Liu, Wei Jiang, and Zhanlei Yang. 2012. Latent topic model based on gaussian-lda for audio retrieval. In *Chinese Conference on Pattern Recognition*, pages 556–563. Springer.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. 1998. An introduction to variational methods for graphical models. In *Learning in graphical models*, pages 105–161. Springer.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Kar Wai Lim and Wray Buntine. 2015. Bibliographic analysis with the citation network topic model. In *Asian conference on machine learning*, pages 142–158. PMLR.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Stuart Lloyd. 1982. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, pages 1–70.

Chengzhi Mao, Kevin Xia, James Wang, Hao Wang, Junfeng Yang, Elias Bareinboim, and Carl Vondrick. 2022. Causal transportability for visual recognition. In *CVPR*.

Jon Mcauliffe and David Blei. 2007. Supervised topic models. *Advances in neural information processing systems*, 20.

Leland McInnes and John Healy. 2017. Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 33–42. IEEE.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Topic discovery via latent space clustering of pretrained language model representations. In *Proceedings of the ACM Web Conference 2022*, pages 3143–3152.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An information bottleneck approach for controlling conciseness in rationale extraction. *arXiv preprint arXiv:2005.00652*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035.

Nicole Peinelt, Dong Nguyen, and Maria Liakata. 2020. tbert: Topic models and bert joining forces for semantic similarity detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7047–7055.

Jacob Portes, Alexander R Trott, Sam Havens, Daniel King, Abhinav Venigalla, Moin Nadeem, Nikhil Sardana, Daya Khudia, and Jonathan Frankle. 2023. Mosaicbert: How to train bert with a lunch money budget. In *Workshop on Efficient Systems for Foundation Models@ ICML2023*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Jie Ren, Mingjie Li, Qirui Chen, Huiqi Deng, and Quanshi Zhang. 2023. Defining and quantifying the emergence of sparse concepts in dnns. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20280–20289.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Jessica Schrouff, Sebastien Baur, Shaobo Hou, Diana Mincu, Eric Loreaux, Ralph Blanes, James Wexler, Alan Karthikesalingam, and Been Kim. 2021. Best of both worlds: local and global explanations with human-understandable concepts. *arXiv preprint arXiv:2106.08641*.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. 2020. Restricting the flow: Information bottlenecks for attribution. *arXiv preprint arXiv:2001.00396*.

Tian Shi, Xuchao Zhang, Ping Wang, and Chandan K Reddy. 2021. Corpus-level and concept-based explanations for interpretable document classification. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(3):1–17.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021. Octis: comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Chong Wang, David Blei, and David Heckerman. 2012. Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298*.

Dongsheng Wang, Dandan Guo, He Zhao, Huangjie Zheng, Korawat Tanwisuth, Bo Chen, and Mingyuan Zhou. 2022. Representing mixtures of word embeddings with mixtures of topic embeddings. *arXiv preprint arXiv:2203.01570*.

Hao Wang, SHI Xingjian, and Dit-Yan Yeung. 2016. Natural-parameter networks: A class of probabilistic neural networks. In *NIPS*, pages 118–126.

Hao Wang and Dit-Yan Yeung. 2016. Towards bayesian deep learning: A framework and some existing methods. *TDKE*, 28(12):3395–3408.

Hao Wang and Dit-Yan Yeung. 2020. A survey on bayesian deep learning. *CSUR*, 53(5):1–37.

Hengyi Wang, Shiwei Tan, and Hao Wang. 2024. Probabilistic conceptual explainers: Towards trustworthy conceptual explanations for vision foundation models. In *ICML*.

Sean Xie, Soroush Vosoughi, and Saeed Hassanpour. 2023. Proto-lm: A prototypical network-based framework for built-in interpretability in large language models. *arXiv preprint arXiv:2311.01732*.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Zihao Xu, Guangyuan Hao, Hao He, and Hao Wang. 2023. Domain indexing variational bayes: Interpretable domain index for domain adaptation. In *ICLR*.

Jingquan Yan and Hao Wang. 2023. Self-interpretable time series prediction with counterfactual explanations. In *ICML*.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. 2023. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Mert Yuksekgonul, Maggie Wang, and James Zou. 2023. Post-hoc concept bottleneck models.

Zihan Zhang, Meng Fang, Ling Chen, and Mohammad-Reza Namazi-Rad. 2022. Is neural topic modelling better than clustering? an empirical study on clustering with contextual embeddings for topics. *arXiv preprint arXiv:2204.09874*.

He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray Buntine. 2020. Neural topic model via optimal transport. *arXiv preprint arXiv:2008.13537*.

A Details on Learning VALC

Update Rules. Similar to Sec. 3.4.1 of the main paper, we expand the ELBO in Eq. 2 of the main paper, take its derivative w.r.t. $\boldsymbol{\mu}_k$ and set it to $\mathbf{0}$:

$$\frac{\partial L}{\partial \boldsymbol{\mu}_k} = \sum_{m,j} \phi_{mjk} w_{mj} \boldsymbol{\Sigma}_k^{-1} (\mathbf{e}_{mj} - \boldsymbol{\mu}_k) = 0, \tag{10}$$

yielding the update rule for learning $\boldsymbol{\mu}_i$:

$$\boldsymbol{\mu}_k = \frac{\sum_{m,j} \phi_{mjk} w_{mj} \mathbf{e}_{mj}}{\sum_{m,j} \phi_{mjk} w_{mj}}, \tag{11}$$

where $\boldsymbol{\Sigma}_k^{-1}$ is canceled out. Similarly, setting the derivatives w.r.t. $\boldsymbol{\Sigma}$ to $\mathbf{0}$, i.e.,

$$\frac{\partial L}{\partial \boldsymbol{\Sigma}_k} = \frac{1}{2} \sum_{m,j} \phi_{mjk} w_{mj} (-\boldsymbol{\Sigma}_k^{-1} + \boldsymbol{\Sigma}_k^{-1} (\mathbf{e}_{mj} - \boldsymbol{\mu}_k)(\mathbf{e}_{mj} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}), \tag{12}$$

we have

$$\boldsymbol{\Sigma}_k = \frac{\sum_{m,j} \phi_{mjk} w_{mj} (\mathbf{e}_{mj} - \boldsymbol{\mu}_k)(\mathbf{e}_{mj} - \boldsymbol{\mu}_k)^T}{\sum_{m,j} \phi_{mjk} w_{mj}}. \tag{13}$$

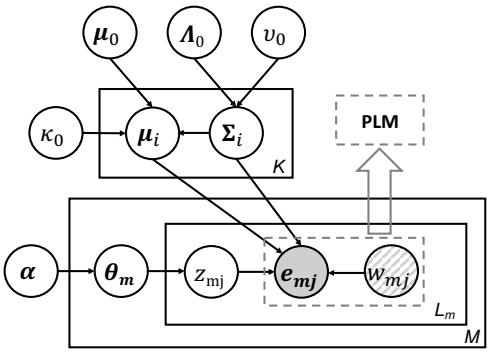

Figure 5: Probabilistic graphical model of smoothed VALC.

Smoothing with Prior Distributions on $\{(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}_{k=1}^K$. To alleviate overfitting and prevent singularity in numerical computation, we impose priors distributions on $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ to smooth the learning process (Fig. 5). Specifically, we use a Normal-Inverse-Wishart prior on $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$:

$$\boldsymbol{\Sigma}_k \sim \mathcal{IW}(\boldsymbol{\Lambda}_0, \nu_0),$$
$$\boldsymbol{\mu}_k | \boldsymbol{\Sigma}_k \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_k / \kappa_0),$$

where $\boldsymbol{\Lambda}_0$, ν_0, $\boldsymbol{\mu}_0$, and κ_0 are hyperparameters for the prior distributions. Taking the expectations of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ over the posterior distibution $\mathcal{NIW}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k | \boldsymbol{\mu}_k^{(n)}, \boldsymbol{\Lambda}_k^{(n)}, \kappa_k^{(n)}, \nu_k^{(n)})$, we have the update rules as:

$$\boldsymbol{\mu}_k \leftarrow \mathbb{E}_{\mathcal{NIW}}[\boldsymbol{\mu}_k] = \frac{\kappa_0 \boldsymbol{\mu}_0 + n_k \tilde{\boldsymbol{\mu}}_k}{\kappa_0 + n_k}, \tag{14}$$

$$\boldsymbol{\Sigma}_k \leftarrow \mathbb{E}_{\mathcal{NIW}}[\boldsymbol{\Sigma}_k] = \frac{\boldsymbol{\Lambda}_0 + \mathbf{S}_k + \frac{\kappa_0 n_k}{\kappa_0 + n_k} (\tilde{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_0)(\tilde{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_0)^T}{\nu_0 + n_k - K - 1}, \tag{15}$$

$$\mathbf{S}_k = \sum_{m,j} \phi_{mjk} w_{mj} (\mathbf{e}_{mj} - \tilde{\boldsymbol{\mu}}_k)(\mathbf{e}_{mj} - \tilde{\boldsymbol{\mu}}_k)^T. \tag{16}$$

where $n_k = \sum_{m,j} \phi_{mjk} w_{mj}$ is the total virtual word counts used to estimate $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$. Eq. 14 and Eq. 15 are the smoothed version of Eq. 7 of the main paper. From the Bayesian perfective, they correspond to the

expectations of $\boldsymbol{\mu}_k$'s and $\boldsymbol{\Sigma}_k$'s posterior distributions. Alg. 1 of the main paper summarizes the learning of VALC.

Online Learning of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$. Note that FLMs are deep neural networks trained using minibatches of data, while Eq. 14 and Eq. 15 need to go through the whole dataset before each update. Inspired by Hoffman et al. (2010); Oord et al. (2017), we using exponential moving average (EMA) to work with minibatchs. Specifically, we update them as:

$$\begin{aligned}
\boldsymbol{\mu}_k &\leftarrow \rho \cdot N \cdot \boldsymbol{\mu}_k + (1-\rho) \cdot B \cdot \tilde{\boldsymbol{\mu}}_k, \\
\boldsymbol{\Sigma}_k &\leftarrow \rho \cdot N \cdot \boldsymbol{\Sigma}_k + (1-\rho) \cdot B \cdot \tilde{\boldsymbol{\Sigma}}_k, \\
N &\leftarrow \rho \cdot N + (1-\rho) \cdot B, \\
\boldsymbol{\mu}_k &\leftarrow \frac{\boldsymbol{\mu}_k}{N}, \quad \boldsymbol{\Sigma}_k \leftarrow \frac{\boldsymbol{\Sigma}_k}{N},
\end{aligned}$$

where B is the minibatch size, N is a running count, and $\rho \in (0, 1)$ is the momentum hyperparameter. $\tilde{\boldsymbol{\mu}}_k$ and $\tilde{\boldsymbol{\Sigma}}_k$ are the updated $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ after applying Eq. 14 and Eq. 15 only on the *current minibatch*.

Effect of Attention Weights. Interestingly, we also observe that FLMs' attention weights on stop words such as 'the' and 'a' tend to be much lower; therefore VALC can naturally ignore these concept-irrelevant stop words when learning and inferring concepts (as discussed in Sec. 3.4.2). This is in contrast to typical topic models (Blei et al., 2003; Blei, 2012) that require preprocessing to remove stop words.

Phrase-Level Interpretations. We can easily infer phrase-level concepts from word-level concepts by treating phrases as sub-documents and adapting Eq. 6 (which provides document-level concepts) in the paper. Specifically, suppose for a given phrase spanning from the r-th word to the s-th word in document m, we can adapt Eq. 6 to provide phrase-level conceptual explanations as $\gamma_{mk}^{(r,s)} = \alpha_k + \sum_{j=r}^{s} \phi_{mjk} w_{mj}$. Here $\gamma_{mk}^{(r,s)}$ is the strength of concept k for the given phrase in document m. In this way, $\gamma_{mk}^{(r,s)}$ can serve as the phrase-level concept explanation of the phrase spanning from r-th word to the s-th word; this is another interesting complementary sub-document-level concept explanation between the word level and the document level.

B Interpretation of the ELBO

VALC's evidence lower bound (ELBO), i.e., Eq. 2 in the paper, is

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\gamma}_m, \{\boldsymbol{\phi}_{m[1:J_m]}\}; \alpha, \{(\boldsymbol{\mu}_{[1:K]}, \boldsymbol{\Sigma}_{[1:K]})\}) &= \mathbb{E}_q[\log p(\boldsymbol{\theta}_m|\alpha)] + \sum_{j=1}^{J_m} \mathbb{E}_q[\log p(\mathbf{z}_{mj}|\boldsymbol{\theta}_m)] \\
&+ \sum_{j=1}^{J_m} \mathbb{E}_q[\log p(\mathbf{e}_{mj}|\mathbf{z}_{mj}, \mu_{\mathbf{z}_{mj}}, \Sigma_{\mathbf{z}_{mj}})] \\
&- \mathbb{E}_q[\log q(\boldsymbol{\theta}_m)] - \sum_{j=1}^{J_m} \mathbb{E}_q[\log q(\mathbf{z}_{mj})]. \qquad (17)
\end{aligned}$$

Derivation of the Evidence Lower Bound. We derive the evidence lower bound by computing the log likelihood of each term. For example, by definition, $p(\mathbf{e}_{mj}|\mathbf{z}_{mj}, \mu_{\mathbf{z}_{mj}}, \Sigma_{\mathbf{z}_{mj}}) = [\mathcal{N}(\mathbf{e}_{mj}; \boldsymbol{\mu}_{mj}, \boldsymbol{\Sigma}_{mj})]^{w_{mj}}$, where $\mathcal{N}(\cdot)$ is the Gaussian distribution. Then we derive the third term $\sum_{j=1}^{J_m} \mathbb{E}_{||}[\log p(\mathbf{e}_{mj}|\mathbf{z}_{mj}, \mu_{\mathbf{z}_{mj}}, \Sigma_{\mathbf{z}_{mj}})]$ in Eq. 17 as follows:

$$\begin{aligned}
\mathbb{E}_q[\log p(\mathbf{e}_{mj}|\mathbf{z}_{mj}, \mu_{\mathbf{z}_{mj}}, \boldsymbol{\Sigma}_{\mathbf{z}_{mj}})] &= \sum_k \phi_{mjk} w_{mj} \log \mathcal{N}(\mathbf{e}_{mj}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
&= \sum_k \phi_{mjk} w_{mj} \{-\frac{1}{2}(\mathbf{e}_{mj} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{e}_{mj} - \boldsymbol{\mu}_k) \\
&- \log[(2\pi)^{d/2} |\boldsymbol{\Sigma}_k|^{1/2}]\}. \qquad (18)
\end{aligned}$$

Table 5: Dataset statistics, including the number of documents (M), vocabulary size (V), the number of corpus categories (L), and the average document length ($\overline{J}$).

Dataset	M	V	L	$\overline{J}$
20 Newsgroups	16,309	1,612	20	48
M10	8,355	1,696	10	5.9
BBC News	2,225	2,949	5	120

Expanding the ELBO to the Loss Function. We can expand the ELBO in Eq. 2 of the main paper as:

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{\gamma}, \boldsymbol{\phi}; \boldsymbol{\alpha}, \{\boldsymbol{\mu}\}_{k=1}^{K}, \{\boldsymbol{\Sigma}\}_{k=1}^{K}) = & \log \boldsymbol{\Gamma}(\sum_{k=1}^{K} \alpha_k) - \sum_{k=1}^{K} \log \boldsymbol{\Gamma}(\alpha_k) + \sum_{k=1}^{K} (\alpha_k - 1)(\Psi(\boldsymbol{\gamma}_k) - \Psi(\sum_{k'=1}^{K} \boldsymbol{\gamma}_{k'})) \\
& + \sum_{j=1}^{J} \sum_{k=1}^{K} \phi_{jk} (\Psi(\boldsymbol{\gamma}_k) - \Psi(\sum_{k'=1}^{K} \boldsymbol{\gamma}_{k'})) \\
& + \sum_{j,k} \phi_{jk} w_j \{ -\frac{1}{2} (\mathbf{e}_j - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{e}_j - \boldsymbol{\mu}_k) - \log[(2\pi)^{d/2} |\boldsymbol{\Sigma}_k|^{1/2}] \} \\
& - \log \boldsymbol{\Gamma}(\sum_{k=1}^{K} \boldsymbol{\gamma}_j) + \sum_{k=1}^{K} \log \boldsymbol{\Gamma}(\boldsymbol{\gamma}_k) - \sum_{k=1}^{K} (\boldsymbol{\gamma}_k - 1)(\Psi(\boldsymbol{\gamma}_k) - \Psi(\sum_{k'=1}^{K} \boldsymbol{\gamma}_{k'})) \\
& - \sum_{j=1}^{J} \sum_{k=1}^{K} \phi_{jk} \log \phi_{jk}. \qquad (19)
\end{aligned}
$$

Definition and Interpretation of the Loss Function. We can interpret the meaning of each term of ELBO as follows:

- **Regularization Term for Document-Level Explanations.** The sum of the first and the fourth terms, namely $\mathbb{E}_q[\log p(\boldsymbol{\theta}_m|\boldsymbol{\alpha})] - \mathbb{E}_q[\log q(\boldsymbol{\theta}_m)]$, is equal to $-KL(q(\boldsymbol{\theta}_m)|p(\boldsymbol{\theta}_m|\boldsymbol{\alpha}))$, which is the negation of KL Divergence between the variational posterior probability $q(\boldsymbol{\theta}_m)$ and the prior probability $p(\boldsymbol{\theta}_m|\alpha)$ of the topic proportion $\boldsymbol{\theta}_m$ for document m. Therefore maximizing the sum of these two terms is equivalent to minimizing the KL Divergence $KL(q(\boldsymbol{\theta}_m)|p(\boldsymbol{\theta}_m|\boldsymbol{\alpha}))$; this serves as a regularization term to make sure the inferred $q(\boldsymbol{\theta}_m)$ is close to its prior distribution $p(\boldsymbol{\theta}_m|\boldsymbol{\alpha})$.

- **Regularization Term for Word-Level Explanations.** Similarly, the sum of the second and the last terms (ignoring the summation over the word index j for simplicity), namely $\mathbb{E}_q[\log p(z_{mj}|\boldsymbol{\theta}_m)] - \mathbb{E}_q[\log q(z_{mj})]$ is equal to $-KL(q(z_{mj})|p(z_{mj}|\boldsymbol{\theta}_m))$, which is the negation of the KL Divergence between the variational posterior probability $q(z_{mj})$ and the prior probability $p(z_{mj}|\boldsymbol{\theta}_m)$ of the word-level topic assignment z_{mj} for word j of document m. Therefore maximizing the sum of these two terms is equivalent to minimizing the KL Divergence $KL(q(z_{mj})|p(z_{mj}|\boldsymbol{\theta}_m))$; this serves as a regularization term to make sure the inferred $q(z_{mj})$ is close to its 'prior' distribution $p(z_{mj}|\boldsymbol{\theta}_m)$.

- **Likelihood Term to Indicate How Much FLM Information is Explained.** The third term $\sum_{j=1}^{J_m} \mathbb{E}_q[\log p(\mathbf{e}_{mj}|z_{mj}, \boldsymbol{\mu}_{z_{mj}}, \boldsymbol{\Sigma}_{z_{mj}})]$ is to maximize the log likelihood $p(\mathbf{e}_{mj}|z_{mj}, \boldsymbol{\mu}_{z_{mj}}, \boldsymbol{\Sigma}_{z_{mj}})$ of every contextual embedding $\mathbf{e}_{mj}$ (for word j of document m) conditioned on the inferred z_{mj} and the parameters $(\boldsymbol{\mu}_{z_{mj}}, \boldsymbol{\Sigma}_{z_{mj}})$.

In this way, we expand the ELBO to a concrete loss function. Each line of Eq. 19 corresponds to the expansion of each of the five terms in the ELBO mentioned above (i.e., Eq. 2 in the paper).

C Experimental Settings and Implementation Details

We will release all code, models, and data. Below we provide more details on the experimental settings and practical implementation.

Data Preprocessing and More Datasets. We follow Terragni et al. (2021) and Zhang et al. (2022) to pre-process these datasets. The statistics of the datasets are summarized in Table 5. We use the standard 8:1:1 train/validation/test set split. We also use the GLUE benchmark (Wang et al., 2018) to

perform *additional* conceptual interpretation in this section and Sec. F. This benchmark includes multiple sub-tasks of predictions, with the paired sentences as inputs. In this paper, we use 4 datasets from GLUE (MRPC, RTE, STS-B, and QQP) to show contextual interpretations. Specifically, we apply VALC to multiple complex natural language understanding (NLU) tasks in the GLUE benchmark. For example, in Appendix F, we show the three-level conceptual explanations of *four different tasks* in the GLUE benchmark using VALC, i.e.,

- **Microsoft Research Paraphrase Corpus (MRPC)**, where the task is paraphrase identification and semantic textual similarity,
- **Recognizing Textual Entailment (RTE)**, where the task is to determine whether one sentence (the premise) entails another sentence (the hypothesis),
- **Semantic Textual Similarity Benchmark (STS-B)**, where the task is to measure the degree of semantic similarity between pairs of sentences (from 0 to 5), and
- **Quora Question Pairs (QQP)**, where the task is to classify whether one question is the duplicate of the other.

Implementation. We implemented and trained the model using PyTorch (Paszke et al., 2019) on an A5000 GPU with 24GB of memory. The training duration was kept under a few hours for all datasets. We utilized the Adam optimizer (Kingma and Ba, 2014) with initial learning rates varying between $10^{-5} \sim 10^{-3}$, tailored to the specific requirements of each dataset.

Visualization Postprocessing. For better showcase the dataset-level concepts as in Fig. 4 of the main paper, we may employ simple linear transformations on the embedding of words after the aforementioned PCA step, in order to scatter all the informative words on the same figures. However, for some datasets such as STS-B, this is not necessary; therefore we do not use it for these datasets.

Topic (Concept) Identification. Inspired by Blei et al. (2003), we identify meaningful topics by listing the top-5 topics for each word, computing the inverse document frequency (IDF), and filtering out topics with the lowest IDF scores. Note that although GLUE benchmark are datasets that consists of documents with small size, making it particularly challenging for traditional topic models (such as LDA) to learn topics; interestingly our VALC can still perform well in learning the topics. We contribute this to the following observations: (1) Compared to traditional LDA using *discrete* word representations, VALC uses *continuous* word embeddings. In such a continuous space, topics learned for one word can also help neighboring words; this alleviates the sparsity issue caused by short documents and therefore learns better topics. (2) VALC's attention-based continuous word counts further improves sample efficiency. In VALC, important words have larger attention weights and therefore larger continuous word counts. In this case, *one* important word in a sentence possesses statistical (sample) power equivalent to *multiple* words; this leads to better sample efficiency in VALC.

Computational Complexity. Our VALC introduces minimal overhead in terms of model training cost. Specifically, VALC's computational complexity is $O(TKd^2)$, where T is the number of epochs (a small number, such as 3, is sufficient for convergence), K is the number of concepts, and d is the dimension of the embeddings (in hidden layers). This means that VALC's computational cost *scales linearly* with the number of concepts K (similar to existing methods).

More NLP Tasks. VALC can be naturally applied to other NLP tasks, such as named entity recognition (NER), reading comprehension, or question answering. Specifically, these tasks involve transformer predictions from multiple positions within the context, rather than relying solely on the 'CLS' token. For example, NER predicts each token in the document as the beginning ('B') of an entity, the inside ('I') of entities, etc. To accommodate this and use VALC to explain each token j in the context, we can substitute the attention from the 'CLS' token with (1) the attention from the 'CLS' token to all tokens of the previous layer with (2) the attention from token j to all tokens of the previous layer in transformers (e.g., using the attention weights from the predicted label 'B' to all tokens of the previous layer as $\mathbf{a}_m$ in VALC). This adaptation allows VALC to maintain its explanatory power across various NLP applications, demonstrating its versatility and effectiveness in a wide range of tasks.

D More Details on Concept Editing

We perform concept pruning to the CLS embeddings for VALC (details in Alg. 2). Since BERTopic and CETopic can infer concepts (topics) only at the document level, their only choice is to prune a concept by completely removing input tokens assigned to the concept (as mentioned in Sec. 5.1 and 5.2). To compare our learned concepts with the baseline models, we first follow their configurations (Grootendorst, 2020; Zhang et al., 2022) to fix BERT model parameters when learning the topics/concepts, train a classifier on top of the fixed contextual embeddings, and then perform concept pruning (Koh et al., 2020) for different evaluated models on the same classifier. Note that concept editing is deterministic; therefore, we conduct our experiments with a single run.

Specifically, we assume each BERT model contains a backbone and a classifier. To perform concept editing:

(1) We first train a classifier on top of the *fixed* BERT embeddings generated by the *fixed* backbone to get the original accuracy in the 'Unedited' column (in Table 2 and Table 3 of the main paper).

(2) We then apply the same embedding cluster methods to these BERT embeddings to infer the concepts/topics for each dataset.

(3) Finally, with the inferred concepts/topics from the baselines (SHAP/LIME, BERTopic and CETopic in Table 2 of the main paper) and our VALC variants (Unweighted and Weighted in Table 3 of the main paper), we perform concept editing and feed the concept-edited embeddings into the trained classifier from Step (1) to compute the editing accuracy for different methods.

Since here one *does not fully finetune the BERT model* (i.e., keeping the backbone fixed), the editing accuracy is expected to be lower than the 'Finetune' column (in Table 2 and Table 3 of the main paper), which serves as the oracle. Table 2 of the main paper shows that our VALC learns better concepts than the baselines, and Table 3 of the main paper shows that the weighted variant of VALC performs better.

Algorithm 3: Algorithm for VALC Document-Level Concept Editing

Input: FLM $f(\cdot)$, classifier $g(\cdot)$, classification loss L, dataset $\{\mathcal{D}_m\}_{m=1}^{M}$, labels $\mathbf{y}$, constant factor ω.

for $m = 1 : M$ **do**

$\mathbf{c}_m = f(\mathcal{D}_m)$

$\mathbf{x}^* = QP(\mathbf{c}_m, \{\boldsymbol{\mu}_k\}_{k=1}^{K})$

$k^* = \arg\min_{k=1}^{K} L(g(\mathbf{c}_m - \omega \cdot x_k^* \boldsymbol{\mu}_k), y_m)$

$\mathbf{c}_m \leftarrow \mathbf{c}_m - \omega \cdot x_{k^*}^* \boldsymbol{\mu}_{k^*}$

Note that SHAP and LIME both interpret the CLS token's embedding, and hence their concept vectors have the same dimension as the FLM embedding vector (768 in our case). When we conduct concept editing on the k'th dimension/concept, we simply subtract the CLS embedding's dimension k with the average value in the batch on dimension k (which means that we know little about the concept/dimension k on this document), and keep values of the other dimensions unchanged. Note that the pruning process is exactly the same for SHAP and LIME. Therefore SHAP and LIME have identical test accuracy and accuracy gain.

Document-Level Concept Editing. We describe the document-level concept editing algorithm of VALC in Alg. 3. c_m denotes the 'CLS' embedding of document m (see Fig. 2 of the main paper).

E Connections Between the Defined Properties and Empirical Results

VALC is able to show which words or embeddings contributed to the document-level concept k. Specifically, our variational parameter (a vector) $\boldsymbol{\phi}_{mj} \in \mathbb{R}^K$ describes how much word j contributes to document m. For example, the k-th entry of $\boldsymbol{\phi}_{mj}$, denoted as ϕ_{mjk} in the paper, describes how much word j contributes to document m in terms of concept k. Therefore, one could use $\arg\max_j \phi_{mjk}$ to find the word that contributes most to document m's concept k. Below, we will explain these four properties using Fig. 4 as a running example.

(1) **Multi-Level Structure** ensures that VALC learns the dataset-, document-, and word-level concepts jointly. In Fig. 4:

- **Dataset-level** concepts are highlighted by the top words of each concept (the top right box of Fig. 4) and the distribution of their embeddings in the FLM (left and middle figures of Fig. 4); for example, *Concept 5 (data analysis)* is marked in red.
- **Document-level** concepts are demonstrated by each document's topic; for instance, in the box for document (a) in Fig. 4 (right), VALC identifies *Topic (Concept) 5* as key to the FLM's prediction of the label 3 (computer science).
- **Word-level** concepts are identified by words in documents. For example, in the box for document (a) in Fig. 4, VALC highlights the words 'genetic' and 'neural' because they are highly related to *Concept 5 (data analysis)*. Terms like 'genetic algorithms' and 'neural networks' are related to data analysis, aligning with the document-level concept.

(2) **Normalization** ensures that concept learning is regulated and smoothed, with inferred concepts appearing reasonable. Specifically, in the document-level explanation $\boldsymbol{\theta}_m$ and word-level explanation $\boldsymbol{\phi}_{mj}$, all concepts are assigned a value within the range of $0 \sim 1$, and all entries sum up to 1, i.e., $\sum_{k=1}^{K} \phi_{mijk} = 1$ and $\sum_{k=1}^{K} \theta_{mk} = 1$. This introduces 'competition' among different concepts; a larger strength for one concept means smaller strength for other concepts. Therefore, together with the help of the Dirichlet prior, it implicitly encourages sparser concept-level explanations θ_m, which are more aligned with humans' cognitive processes and more human-understandable (humans tend to make decisions with a *small* set of concepts).

(3) **Additivity** enables FLMs to incorporate relevant concepts and exclude irrelevant ones, thereby enhancing prediction accuracy (as shown in Table 2 and Table 3). For example, in document (a) of Fig. 4, VALC identifies *Concept 5* as a highly related concept, distinguishing it from less related concepts. In practice, this may help practitioners identify key concepts in model prediction and more effectively intervene to improve model prediction accuracy (e.g., an expert may find that a concept is relevant and manually down-weight the concept to enhance the model's prediction).

(4) **Mutual Information Maximization** ensures a strong correlation between (1) VALC's generated concept explanations and (2) the explained model's representation and predictions. In other words, it ensures that VALC is explaining the target FLM, rather than generating concept explanations irrelevant to the target FLM. For instance, in document (a) of Fig. 4, the inferred document-level *Concept 5* (data analysis) effectively explains the FLM prediction, i.e., label 3 (computer science), by highlighting the intrinsic link between the data analysis concept and the class label computer science. This connection is evidenced by the words in dataset-level *Concept 5* (top right box). The mutual information between the inferred *Concept 5* (data analysis) and label 3 (computer science) contributes to generating high-quality explanations.

F More Conceptual Interpretation Results in Different Downstream Tasks

Dataset-Level Interpretations. As in the main paper, we leverage VALC as an interpreter on MRPC, RTE, STS-B and QQP, respectively, sample $3, 3, 4, 4$ concepts (topics) for each dataset respectively, and plot the word embeddings of the top words (closest to the center $\boldsymbol{\mu}_i$) in these concepts using PCA. Fig. 6(left) shows the concepts from MRPC. We can observe Concept 20 is mostly about policing, including words such as 'suspect', 'police', and 'house'. Concept 24 is mostly about politics, including words such as 'capital', 'Congress', and 'Senate'. Concept 27 contains mostly names such as 'Margaret' and 'Mary'. Similarly, Fig. 6(right) shows the concepts from RTE. We can observe Concept 67 is related to West Asia and includes words such as 'Quran' and 'Pasha'. Concept 13 is related to Europe and includes European countries/names such as 'Prussia' and 'Salzburg'. Concept 91 is mostly about healthcare and includes words such as 'physiology' and 'insulin'. Fig. 7 shows the concepts from STS-B. We can observe Concept 63 is mostly about household and daily life, including words such as 'trash', 'flowers', 'airs', and 'garden'. Concept 60 is mostly about tools, including words such as 'stations', 'rope', 'parachute', and 'hose'. Concept 84 is mostly about national security, including words such as 'guerilla', 'NSA', 'espionage', and 'raided'. Concept 55 contains mostly countries and cities such as 'Kiev', 'Moscow',

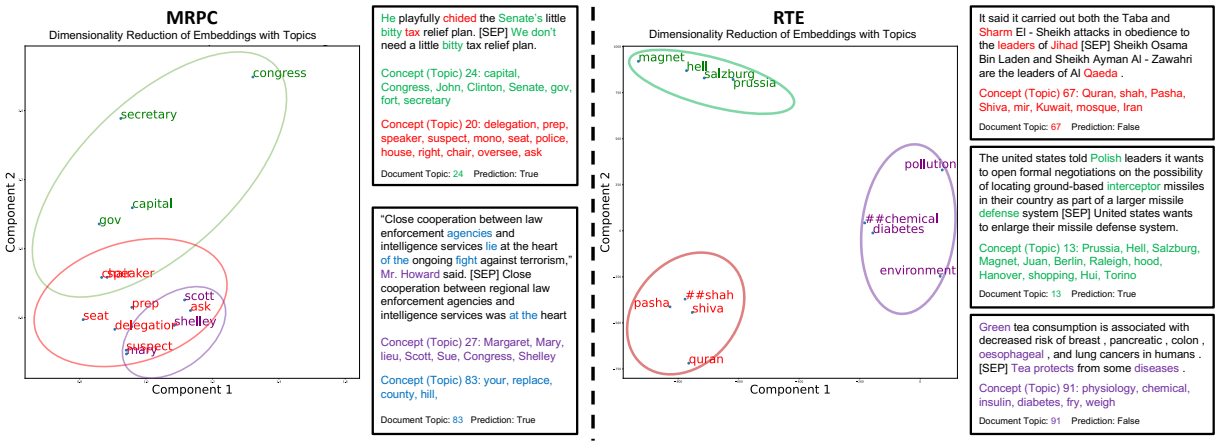

Figure 6: Visualization of VALC's learned topics of contextual word embeddings. **Left:** MRPC's dataset-level interpretation with two example documents. Concept 83 is relatively far from the other three concepts in the embedding space; therefore we omit it on the left panel for better readability. **Right:** RTE's dataset-level interpretation with three example documents.

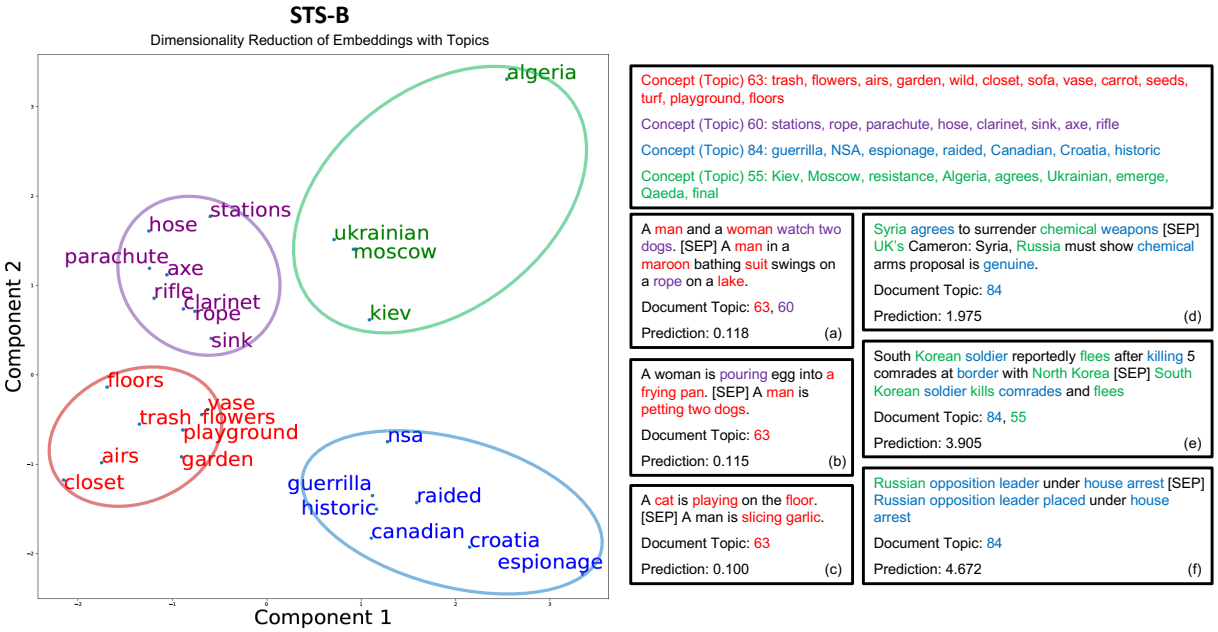

Figure 7: Visualization of VALC's learned topics of contextual word embeddings. We show STS-B's dataset-level interpretation with six example documents. The prediction of VALC is between the range of $[0, 5]$.

'Algeria', and 'Ukrainian'. Similarly, Fig. 8 shows the concepts from QQP. We can observe that Concept 12 is mostly about negative attitude, including words such as 'boring', 'criticism', and 'blame'. Concept 73 is mostly about Psychology, including words such as 'adrenaline', 'haunting', and 'paranoia'. Concept 34 is mostly about prevention and conservatives, including words such as 'destroys', 'unacceptable', and 'prohibits'. Concept 64 is mostly about strategies, including words such as 'rumours', 'boycott', and 'deportation'.

Document-Level Interpretations. For document-level conceptual interpretations, we sample two example documents from MRPC (Fig. 6(left)), three from RTE (Fig. 6(right)), six from STS-B (Fig. 7) and eight from QQP (Fig. 8), respectively, where each document contains a pair of sentences. The MRPC task is to predict whether one sentence paraphrases the other. For example, in the first document of MRPC, we can see that our VALC correctly interprets the model prediction 'True' with Concept 24 (politics). The RTE task is to predict whether one sentence entail the other. For example, in the second document of RTE, VALC correctly interprets the model prediction 'True' with Concept 13 (countries). The STS-B task is to predict the semantic similarity between two sentences with the score range of $[0, 5]$. For example,

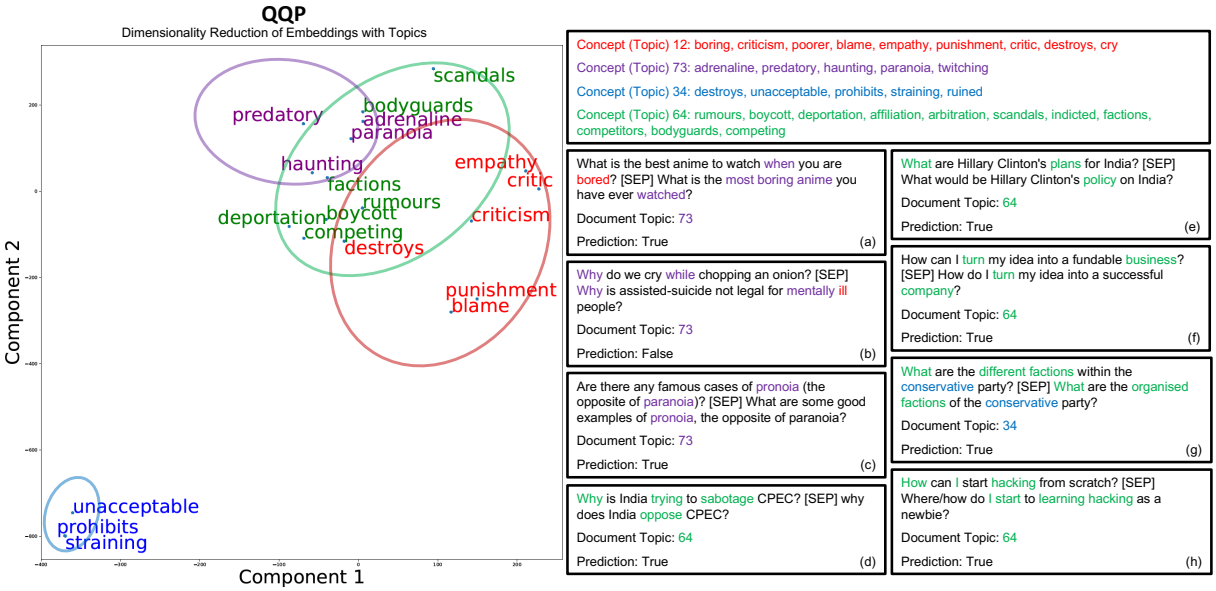

Figure 8: Visualization of VALC's learned topics of contextual word embeddings. We show QQP's dataset-level interpretation with eight example documents.

Table 6: Example concepts on RTE dataset learned by VALC.

Concepts	Top Words							
bio-chem	cigarette	biological	ozone	cardiovascular	chemist	liver	chemical	toxin
citizenship	indies	bolivian	fiji	surrey	jamaican	dutch	latino	caribbean
names	mozart	spielberg	einstein	bush	kurt	liszt	hilton	lynn
conspiracy	secretly	corrupt	disperse	infected	ill	hidden	illegally	sniper
administration	reagan	interior	ambassador	prosecutor	diplomat	legislative	spokesman	embassy
crime	fraud	laundering	sheriff	prosecutor	corruption	fool	robber	greed

in Document (a) of Fig. 7, we can see that VALC correctly interpret the model's predicted similarity score '0.118' (which is relatively low,) with Concept 63 (household and daily life) and Concept 60 (tools). Similarly, in Document (f) of Fig. 7, we can see that VALC correctly interpret the model's predicted similarity score '4.672' (which is relatively high) with Concept 84 (national security). The QQP task is to predict whether the two questions are paraphrase of each other. For example, in Document (b) of Fig. 8, we can see that VALC correctly interprets the model's predicted label 'False' with Concept 73 (Psychology). Similarly, in Document (e) of Fig. 8, we can see that VALC correctly interprets the model's predicted label 'True' with Concept 64 (strategies).

Word-Level Interpretations. For word-level conceptual interpretations, we can observe that VALC interpret the FLM's prediction on MRPC's first document (Fig. 6(left)) using words such as 'senate' and 'bitty' that are related to politics. Note that the word 'bitty' is commonly used (with 'little') by politicians to refer to the small size of tax relief/cut plans. Similarly, for RTE's first document (Fig. 6(right)), VALC correctly identifies Concept 67 (West Asia) and interprets the model prediction 'False' by distinguishing between keywords such as 'Jihad' and 'Al Qaeda'. likewise, we can observe that VALC interprets FLM's prediction on Document (c) of Fig. 7 using words such as 'cat', 'floor', and 'garlic' that are related to household and daily life. Also, VALC interprets FLM's prediction on Document (e) of Fig. 7 using words such as 'soldier' and 'border' that are related to national security. Similarly, for QQP's Document (d) (Fig. 8), VALC correctly interprets the model prediction 'True' by identifying keywords such as 'sabotage' and 'oppose' with similar meanings in the topic of strategies. For QQP's Document (g), (Fig. 8), VALC interprets the words in the both sentences with the same semantics, such as 'conservative' that is related to prevention and conservatives (note that in politics, 'conservative' refers to parties that tend to prevent/block new policies or legislation), and thereby predicting the correct label 'True'.

Example Concepts. Following Blei et al. (2003), we show the learned concepts on the RTE dataset in Table 6, which is complementary to aforementioned explanations. We select several different topics

Table 7: Comparison of Unedited and Unedited+$\boldsymbol{\theta}$ on 20 Newsgroups, M10, and BBC News. We mark the best results with **bold face**.

	Unedited	Unedited+$\boldsymbol{\theta}$
20 Newsgroups	51.26	**51.74**
M10	69.74	**70.76**
BBC News	93.72	**94.90**

from Fig. 6. As in Sec. 5.4 of the main paper, we obtain top words from each concept via first calculating the average of the each word's corresponding contextual embeddings over the dataset, and then getting the nearest words to each topic center ($\boldsymbol{\mu}_k$) in the embedding space. As we can see in Table 6, VALC can capture various concepts with profound and accurate semantics. Therefore, although FLM embeddings are contextual and continuous, our VALC can still find conceptual patterns of words on the dataset-level.

G More Quantitative Results.

Document Classification with VALC Concepts. We conducted additional experiments to perform document classification using the 'CLS' token's embedding and $\boldsymbol{\theta}$ (inferred from VALC) as features. Table 7 shows the results on three datasets. The results show that our VALC can learn meaningful concept vector $\boldsymbol{\theta}$, which can improve model predictions of document labels.

H Theory on the Mutual Information Maximization Property

We provide the following proof of Theorem 4.1 of the main paper.

For convenience, let $\Omega = (\boldsymbol{\mu}_{k=1}^K, \Sigma_{k=1}^K)$, and $\beta = (\boldsymbol{\theta}_m, \mathbf{z}_m)$.

We then introduce a helper joint distribution of the variables $\mathbf{e}_m$ and β, $s(\mathbf{e}_m, \beta) = p(\mathbf{e}_m)q(\beta|\mathbf{e}_m)$.

According to the definition of ELBO of Section 3.4.1, in Eq. 9, we have

$$LHS = \mathcal{L}(\gamma_m, \phi_m; \alpha, \Omega) = \mathbb{E}_{p(\mathbf{e}_m)}[\mathbb{E}_{q(\beta)}[\log p(\mathbf{e}_m|\Omega, \beta)]] + \mathbb{E}_{q(\beta)}[\log q(\beta|\Omega)]. \tag{20}$$

Since $\mathbb{E}_{q(\beta)}[\log q(\beta|\Omega)] \leq 0$, we only need to prove that

$$\mathbb{E}_{p(\mathbf{e}_m)}[\mathbb{E}_{q(\beta)}[\log p(\mathbf{e}_m|\Omega, \beta)]] \leq I_s(\mathbf{e}_m; \beta) - H(\mathbf{e}_m) = RHS. \tag{21}$$

Then we have that

$$\begin{aligned}
\mathbb{E}_{p(\mathbf{e}_m)}[\mathbb{E}_q[\log p(\mathbf{e}_m|\beta, \Omega)]] &\leq \mathbb{E}_{p(\mathbf{e}_m)}[\mathbb{E}_q[\log p(\mathbf{e}_m|\beta)]] \\
&= \mathbb{E}_{p(\mathbf{e}_m)}[\mathbb{E}_q[\log \frac{q(\mathbf{e}_m|\beta)}{p(\mathbf{e}_m)} \frac{p(\mathbf{e}_m)p(\mathbf{e}_m|\beta)}{q(\mathbf{e}_m|\beta)}]] \\
&= \mathbb{E}_{p(\mathbf{e}_m)}[\mathbb{E}_q[\log \frac{q(\mathbf{e}_m|\beta)}{p(\mathbf{e}_m)}]] + \mathbb{E}_{p(\mathbf{e}_m)}[\mathbb{E}_q[\log p(\mathbf{e}_m)]] + \mathbb{E}_{p(\mathbf{e}_m)}[\mathbb{E}_q[\log \frac{p(\mathbf{e}_m|\beta)}{q(\mathbf{e}_m|\beta)}]] \\
&= I_s(\mathbf{e}_m; \beta) - H(\mathbf{e}_m) - \mathbb{E}_q[KL(q(\mathbf{e}_m|\beta)|p(\mathbf{e}_m|\beta))] \\
&\leq I_s(\mathbf{e}_m; \beta) - H(\mathbf{e}_m) - 0 = RHS,
\end{aligned} \tag{22}$$

which concludes the proof of Theorem 4.1.

I Theoretical Analysis on Continuous Word Counts

Before going to the claims and proofs, first we specify some basic problem settings and assumptions. Suppose there are $K + 1$ topic groups, each of which is regarded to be sampled from a parameterized multivariate Gaussian distribution. In specific, the $K + 1$ 'th distribution of topic has a much larger covariance, and in the same time, closed to the center of embedding space. The prementioned properties can be measured by a series of inequalities:

The approximate marginal log-likelihood of word embeddings, i.e., the third term of the ELBO as mentioned in Eqn. 2 of the main paper, is:

$$\begin{aligned}\mathcal{L}^{(train)} &= \sum\nolimits_{j=1}^{J_m} \mathbb{E}_q[\log p(\mathbf{e}_{mj}|z_{mj}, \boldsymbol{\mu}_{z_{mj}}, \boldsymbol{\Sigma}_{z_{mj}})] \\ &= \sum_{m,j,k} \phi_{mjk} w_{mj} \{-\frac{1}{2}(\mathbf{e}_{mj} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{e}_{mj} - \boldsymbol{\mu}_k) - \log[(2\pi)^{d/2}|\boldsymbol{\Sigma}_k|^{1/2}]\}. \end{aligned} \tag{23}$$

The above equation is the training objective, yet for fair comparison of different training schemes, we calculate the approximated likelihood with word count 1 for all words.

$$\begin{aligned}\mathcal{L}^{(eval)} &= \sum\nolimits_{j=1}^{J_m} \mathbb{E}_q[\log p'(\mathbf{e}_{mj}|z_{mj}, \boldsymbol{\mu}_{z_{mj}}, \boldsymbol{\Sigma}_{z_{mj}})] \\ &= \sum_{m,j,k} \phi_{mjk} \{-\frac{1}{2}(\mathbf{e}_{mj} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{e}_{mj} - \boldsymbol{\mu}_k) - \log[(2\pi)^{d/2}|\boldsymbol{\Sigma}_k|^{1/2}]\}. \end{aligned} \tag{24}$$

I.1 Gaussian Mixture Models

Suppose we have a ground truth GMM model with parameters $\boldsymbol{\pi}^* \in \mathbb{R}^K$ and $\{\boldsymbol{\mu}_k^*, \boldsymbol{\Sigma}_k^*\}_{k=1}^K$, with K different Gaussian distributions. In the dataset, let N and N_s denote the numbers of non-stop-words and stop-words, respectively. Then the marginal log likelihood of a learned GMM model on a given data sample $\mathbf{e}$ can be written as

$$p(\mathbf{e}|\{\boldsymbol{\mu}, \boldsymbol{\Sigma}\}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \boldsymbol{\pi}_k \mathcal{N}(\mathbf{e}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{25}$$

Assuming a dataset of $N + N_s$ words $\{\mathbf{e}_i\}_{i=1}^{N+N_s}$ and taking the associated weights w_i for each word into account, the log-likelihood of the dataset can be written as

$$\sum_{i=1}^{N+N_s} p(\mathbf{e}_i|\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K, \boldsymbol{\pi}) = \sum_{i=1}^{N} \log \sum_{k=1}^{K} w_i \boldsymbol{\pi}_k \mathcal{N}(\mathbf{e}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i=N+1}^{N+N_s} \log \sum_{k=1}^{K} w_i \boldsymbol{\pi}_k \mathcal{N}(\mathbf{e}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{26}$$

Leveraging Jensen's inequality, we obtain a lower bound of the above quantity (denoting as $\boldsymbol{\Theta}$ the collection of parameters $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$ and $\boldsymbol{\pi}$):

$$\mathcal{L}_{\text{GMM}}(\boldsymbol{\Theta}, \{w_i\}) = \sum_{i=1}^{N} w_i \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{e}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + \sum_{i=N+1}^{N+N_s} w_i \log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{e}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) + C, \tag{27}$$

where C is a constant.

In the following theoretical analysis, we consider the following three different configurations of the weights w_i.

Definition I.1 (**Weight Configurations**). We define three different weight configurations as follows:

- Identical Weights: $w_i = \frac{1}{N+N_s}, i \in \{1, 2, \ldots, N + N_s\}$
- Ground-Truth Weights : $w_i = \begin{cases} \frac{1}{N}, & i \in \{1, 2, \ldots, N\} \\ 0, & i \in \{N+1, N+2, \ldots, N+N_s\} \end{cases}$
- Attention-Based Weights: $w_i = \begin{cases} \lambda_1 \in [\frac{1}{N+N_s}, \frac{1}{N}], & i \in \{1, 2, \ldots, N\} \\ \lambda_2 \in [0, \frac{1}{N+N_s}], & i \in \{N+1, N+2, \ldots, N+N_s\} \end{cases}$

Definition I.2 (**Advanced Weight Configurations**). We define three different weight configurations as follows:

- Identical Weights: $w_i = \frac{1}{N+N_s}, i \in \{1, 2, \ldots, N + N_s\}$

- Ground-Truth Weights : $w_i = \begin{cases} \frac{1}{N}, & i \in \{1, 2, \ldots, N\} \\ 0, & i \in \{N+1, N+2, \ldots, N+N_s\} \end{cases}$

- Attention-Based Weights: $w_i \in \begin{cases} [\frac{1}{N+N_s}, \frac{1}{N}], & i \in \{1, 2, \ldots, N\} \\ [0, \frac{1}{N+N_s}], & i \in \{N+1, N+2, \ldots, N+N_s\} \end{cases}$

Definition I.3 (Optimal Parameters). With Definition I.1, the corresponding optimal parameters are then defined as follows:

$$\mathbf{\Theta}_I = \arg\max_{\mathbf{\Theta}} \mathcal{L}(\mathbf{\Theta}; \mathbf{w} \to \text{Identical}), \tag{28}$$

$$\mathbf{\Theta}_G = \arg\max_{\mathbf{\Theta}} \mathcal{L}(\mathbf{\Theta}; \mathbf{w} \to \text{GT}), \tag{29}$$

$$\mathbf{\Theta}_A = \arg\max_{\mathbf{\Theta}} \mathcal{L}(\mathbf{\Theta}; \mathbf{w} \to \text{Attention}), \tag{30}$$

where $\mathbf{w} \to$ Identical, $\mathbf{w} \to$ GT, and $\mathbf{w} \to$ Attention indicates that 'Identical Weights', 'Ground-Truth Weights', and 'Attention-Based Weights' are used, respectively.

Lemma I.1. *Suppose we have two series of functions $\{f_{1,i}(x)\}$ and $\{f_{2,i}(x)\}$, with two non-negative weighting parameters λ_1, λ_2 satisfying $N\lambda_1 + N_s\lambda_2 = 1$. We define the final objective function $f(\cdot)$ as:*

$$f(x; \lambda_1, \lambda_2) = \lambda_1 \sum_{i=1}^{N} f_{1,i}(x) + \lambda_2 \sum_{i=N+1}^{N_s} f_{2,i}(x). \tag{31}$$

We assume two pairs of parameters (λ_1, λ_2) and (λ_1', λ_2'), where

$$\lambda_1 \geq \lambda_1', \tag{32}$$

$$\lambda_2 \leq \lambda_2'. \tag{33}$$

Defining the optimal values of the objective function for different weighting parameters as

$$\hat{x} = \arg\max_x f(x; \lambda_1, \lambda_2), \tag{34}$$

$$\hat{x}' = \arg\max_x f(x; \lambda_1', \lambda_2'), \tag{35}$$

we then have that

$$f(\hat{x}; \frac{1}{N}, 0) \geq f(\hat{x}'; \frac{1}{N}, 0). \tag{36}$$

Proof. We prove this theorem by contradiction. Suppose that we have

$$f(\hat{x}; \frac{1}{N}, 0) < f(\hat{x}'; \frac{1}{N}, 0). \tag{37}$$

According to Eq. 32, i.e., $\lambda_1 \geq \lambda_1'$, and the equation $N\lambda_1 + N_s\lambda_2 = 1$, we have

$$\lambda_1\lambda_2' = \lambda_1 \frac{1 - N\lambda_1'}{N_s} \geq \lambda_1' \frac{1 - N\lambda_1}{N_s} = \lambda_1'\lambda_2. \tag{38}$$

According to Eq. 35, we have the following equality:

$$f(\hat{x}; \lambda_1', \lambda_2') \leq f(\hat{x}'; \lambda_1', \lambda_2'). \tag{39}$$

Combined with the aforementioned assumption in Eq. 37, we have that

$$\lambda_2' f(\hat{x};\lambda_1,\lambda_2) = \lambda_1\lambda_2' \sum_{i=1}^{N} f_{1,i}(\hat{x}) + \lambda_2\lambda_2' \sum_{i=N+1}^{N_s} f_{2,i}(\hat{x}) \tag{40}$$

$$=(\lambda_1'\lambda_2 \sum_{i=1}^{N} f_{1,i}(\hat{x}) + \lambda_2'\lambda_2 \sum_{i=N+1}^{N_s} f_{2,i}(\hat{x})) + (N(\lambda_1\lambda_2' - \lambda_1'\lambda_2) \cdot \frac{1}{N} \sum_{i=1}^{N} f_{1,i}(\hat{x})) \tag{41}$$

$$=\lambda_2 f(\hat{x};\lambda_1',\lambda_2') + N(\lambda_1\lambda_2' - \lambda_1'\lambda_2) f(\hat{x};\frac{1}{N},0) \tag{42}$$

$$<\lambda_2 f(\hat{x}';\lambda_1',\lambda_2') + N(\lambda_1\lambda_2' - \lambda_1'\lambda_2) f(\hat{x}';\frac{1}{N},0) \tag{43}$$

$$=(\lambda_1'\lambda_2 \sum_{i=1}^{N} f_{1,i}(\hat{x}') + \lambda_2'\lambda_2 \sum_{i=N+1}^{N_s} f_{2,i}(\hat{x}')) + (N(\lambda_1\lambda_2' - \lambda_1'\lambda_2) \cdot \frac{1}{N} \sum_{i=1}^{N} f_{1,i}(\hat{x}')) \tag{44}$$

$$=\lambda_1\lambda_2' \sum_{i=1}^{N} f_{1,i}(\hat{x}') + \lambda_2\lambda_2' \sum_{i=N+1}^{N_s} f_{2,i}(\hat{x}') \tag{45}$$

$$=\lambda_2' f(\hat{x}';\lambda_1,\lambda_2), \tag{46}$$

which contradicts the definition of $\hat{x}$ in Eq. 34 (i.e., $\hat{x}$ maximizes $f(x;\lambda_1,\lambda_2)$), completing the proof. □

Lemma I.2. *Suppose we have two series of functions $\{f_{1,i}(x)\}$ and $\{f_{2,i}(x)\}$, with two series of non-negative weighting parameters $\boldsymbol{\lambda}_1 = [\lambda_{1,i}]_{i=1}^{N}, \boldsymbol{\lambda}_2 = [\lambda_{2,i}]_{i=N+1}^{N_s}$ satisfying $\sum_{i=1}^{N} \lambda_{1,i} + \sum_{i=N+1}^{N_s} \lambda_{2,i} = 1$. We define the final objective function $f(\cdot)$ as:*

$$f(x;\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2) = \sum_{i=1}^{N} \lambda_{1,i} f_{1,i}(x) + \sum_{i=N+1}^{N_s} \lambda_{2,i} f_{2,i}(x). \tag{47}$$

We assume two pairs of parameters $(\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2)$ and $(\boldsymbol{\lambda}_1',\boldsymbol{\lambda}_2')$, where

$$\lambda_{1,i} \geq \lambda_{1,i}', \;\; i \in \{1,2,...,N\}, \tag{48}$$

$$\lambda_{2,i} \leq \lambda_{2,i}', \;\; i \in \{N+1,N+2,...,N_s\}. \tag{49}$$

Defining the optimal values of the objective function for different weighting parameters as

$$\hat{x} = \arg\max_x f(x;\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2), \tag{50}$$

$$\hat{x}' = \arg\max_x f(x;\boldsymbol{\lambda}_1',\boldsymbol{\lambda}_2'), \tag{51}$$

$$x^* = \arg\max f(x,\frac{\mathbf{1}}{N},\mathbf{0}). \tag{52}$$

*Under the following **Assumptions** (with $\mathbf{1}$ and $\mathbf{0}$ denoting vectors with all entries equal to 1 and 0, respectively):*

1. $f(\hat{x},\mathbf{0},\boldsymbol{\lambda}_2) \leq f(\hat{x}',\mathbf{0},\boldsymbol{\lambda}_2)$.
2. $f(x;\boldsymbol{\lambda},\mathbf{0}) \geq f(x';\boldsymbol{\lambda},\mathbf{0})$, *iff* $\|x-x^*\| \leq \|x'-x^*\|, \;\; \boldsymbol{\lambda} \geq 0, \;\; \|\boldsymbol{\lambda}\|_1 = 1$.

we have that

$$f(\hat{x};\frac{\mathbf{1}}{N},\mathbf{0}) \geq f(\hat{x}';\frac{\mathbf{1}}{N},\mathbf{0}). \tag{53}$$

Proof. We start with proving the following equality by contradiction:

$$\|\hat{x}-x^*\| \leq \|\hat{x}'-x^*\|. \tag{54}$$

Specifically, if

$$\|\hat{x} - x^*\| > \|\hat{x}' - x^*\|, \tag{55}$$

leveraging the Assumption 1 and 2 above, we have that

$$f(\hat{x}; \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = f(\hat{x}; \boldsymbol{\lambda}_1, \mathbf{0}) + f(\hat{x}; \mathbf{0}, \boldsymbol{\lambda}_2) < f(\hat{x}'; \boldsymbol{\lambda}_1, \mathbf{0}) + f(\hat{x}'; \mathbf{0}, \boldsymbol{\lambda}_2) = f(\hat{x}'; \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2), \tag{56}$$

which contradicts Eq. 50. Therefore, Eq. 54 holds.

Combining Eq. 54 and Assumption 2 above, we have that

$$f(\hat{x}; \frac{\mathbf{1}}{N}, \mathbf{0}) \geq f(\hat{x}'; \frac{\mathbf{1}}{N}, \mathbf{0}), \tag{57}$$

concluding the proof. □

Based on the definitions and lemmas above, we have the following theorems:

Theorem I.3 (Advantage of $\boldsymbol{\Theta}_A$ in the Simplified Case). *With Definition I.1 and Definition I.3, comparing $\boldsymbol{\Theta}_I$, $\boldsymbol{\Theta}_G$, and $\boldsymbol{\Theta}_A$ by evaluating them on the marginal log-likelihood of non-stop-words, i.e., $\mathcal{L}(\cdot, w \to GT)$, we have that*

$$\mathcal{L}_{GMM}(\boldsymbol{\Theta}_I; \mathbf{w} \to GT) \leq \mathcal{L}_{GMM}(\boldsymbol{\Theta}_A; \mathbf{w} \to GT) \leq \mathcal{L}_{GMM}(\boldsymbol{\Theta}_G; \mathbf{w} \to GT). \tag{58}$$

Proof. First, by definition one can easily find that $\boldsymbol{\Theta}_G$ achieves the largest $\mathcal{L}(\cdot; \mathbf{w} \to \text{GT})$ among the three:

$$\max[\mathcal{L}_{GMM}(\boldsymbol{\Theta}_I; \mathbf{w} \to \text{GT}), \mathcal{L}_{GMM}(\boldsymbol{\Theta}_A; \mathbf{w} \to \text{GT})] \leq \max_{\boldsymbol{\Theta}} \mathcal{L}_{GMM}(\boldsymbol{\Theta}; \mathbf{w} \to \text{GT}) = \mathcal{L}_{GMM}(\boldsymbol{\Theta}_G; \mathbf{w} \to \text{GT}). \tag{59}$$

Next, we set $\{w_i\}_{i=1}^{N}$ to λ_1 and $\{w_i\}_{i=N+1}^{N+N_s}$ to λ_2, respectively; we rewrite $\log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{e}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ as $f_{1,i}(x)$ for $i \in \{1, 2, \ldots, N\}$ and $f_{2,i}(x)$ for $i \in \{N+1, N+1, \ldots, N+N_s\}$, where x corresponds to $\boldsymbol{\Theta} \triangleq (\boldsymbol{\pi}, \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K})$. By Lemma I.1, we have that

$$\mathcal{L}_{\text{GMM}}(\boldsymbol{\Theta}_A; \mathbf{w} \to \text{GT}) \leq \mathcal{L}_{\text{GMM}}(\boldsymbol{\Theta}_G; \mathbf{w} \to \text{GT}). \tag{60}$$

Combining Eq. 59 and Eq. 60 concludes the proof. □

Theorem I.3 shows that under mild assumptions, the attention-based weights can help produce better estimates of $\boldsymbol{\Theta}$ in the presence of noisy stop-words and therefore learns higher-quality topics from the corpus, improving interpretability of FLMs.

Theorem I.4 (Advantage of $\boldsymbol{\Theta}_A$ in the General Case). *With Definition I.2 and Definition I.3, comparing $\boldsymbol{\Theta}_I$, $\boldsymbol{\Theta}_G$, and $\boldsymbol{\Theta}_A$ by evaluating them on the marginal log-likelihood of non-stop-words, i.e., $\mathcal{L}_{GMM}(\cdot, w \to GT)$, we have that*

$$\mathcal{L}_{GMM}(\boldsymbol{\Theta}_I; \mathbf{w} \to GT) \leq \mathcal{L}_{GMM}(\boldsymbol{\Theta}_A; \mathbf{w} \to GT) \leq \mathcal{L}_{GMM}(\boldsymbol{\Theta}_G; \mathbf{w} \to GT). \tag{61}$$

Proof. First, by definition one can easily find that $\boldsymbol{\Theta}_G$ achieves the largest $\mathcal{L}(\cdot; \mathbf{w} \to \text{GT})$ among the three:

$$\max[\mathcal{L}_{\text{GMM}}(\boldsymbol{\Theta}_I; \mathbf{w} \to \text{GT}), \mathcal{L}_{\text{GMM}}(\boldsymbol{\Theta}_A; \mathbf{w} \to \text{GT})] \leq \max_{\boldsymbol{\Theta}} \mathcal{L}_{\text{GMM}}(\boldsymbol{\Theta}; \mathbf{w} \to \text{GT}) = \mathcal{L}_{\text{GMM}}(\boldsymbol{\Theta}_G; \mathbf{w} \to \text{GT}). \tag{62}$$

Next, we invoke Lemma I.2 by (1) setting $\{w_i\}_{i=1}^{N}$ to $\boldsymbol{\lambda}_1$ and $\{w_i\}_{i=N+1}^{N+N_s}$ to $\boldsymbol{\lambda}_2$, respectively, and (2) rewriting $\log \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{e}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ as $f_{1,i}(x)$ for $i \in \{1, 2, \ldots, N\}$ and $f_{2,i}(x)$ for $i \in \{N+1, N+1, \ldots, N+N_s\}$, where x corresponds to $\boldsymbol{\Theta} \triangleq (\boldsymbol{\pi}, \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^{K})$. By Lemma I.2, we then have that

$$\mathcal{L}_{\text{GMM}}(\boldsymbol{\Theta}_A; \mathbf{w} \to \text{GT}) \leq \mathcal{L}_{\text{GMM}}(\boldsymbol{\Theta}_G; \mathbf{w} \to \text{GT}). \tag{63}$$

Note that because $f_{1,i}(\cdot)$ and $f_{2,i}(\cdot)$ are Gaussian, therefore Assumption 1 and 2 in Lemma I.2 hold naturally under mild regularity conditions.

Combining Eq. 62 and Eq. 63 concludes the proof. □

I.2 VALC as Interpreters

As mentioned in Eq. B , the ELBO of the marginal likelihood (denoting as $\mathbf{\Theta}$ the collection of parameters $\boldsymbol{\phi}, \boldsymbol{\gamma}$ and $\{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$) is as follows:

$$\begin{aligned}\mathcal{L}_{\text{VALC}}(\mathbf{\Theta}; \{w_i\}) &= \sum_{j=1}^{L'} \mathbb{E}_q[\log p(\mathbf{e}_{mj}|z_{mj}, \boldsymbol{\mu}_{z_{mj}}, \boldsymbol{\Sigma}_{z_{mj}})] \\ &= \sum_{m,j} w_{mj} \sum_k \phi_{mjk}\{-\frac{1}{2}(\mathbf{e}_{mj} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\mathbf{e}_{mj} - \boldsymbol{\mu}_k) - \log[(2\pi)^{H/2}|\boldsymbol{\Sigma}_k|^{1/2}]\}. \end{aligned} \tag{64}$$

Based on the definitions and lemmas above, we have the following theorems:

Theorem I.5 (**Advantage of $\mathbf{\Theta}_A$ in the Simplified Case**). *With Definition I.1 and Definition I.3, comparing $\mathbf{\Theta}_I$, $\mathbf{\Theta}_G$, and $\mathbf{\Theta}_A$ by evaluating them on the marginal log-likelihood of non-stop-words, i.e., $\mathcal{L}(\cdot, w \to GT)$, we have that*

$$\mathcal{L}_{VALC}(\mathbf{\Theta}_I; \mathbf{w} \to GT) \leq \mathcal{L}_{VALC}(\mathbf{\Theta}_A; \mathbf{w} \to GT) \leq \mathcal{L}_{VALC}(\mathbf{\Theta}_G; \mathbf{w} \to GT). \tag{65}$$

Proof. First, by definition one can easily find that $\mathbf{\Theta}_G$ achieves the largest $\mathcal{L}(\cdot; \mathbf{w} \to \text{GT})$ among the three:

$$\max[\mathcal{L}_{\text{VALC}}(\mathbf{\Theta}_I; \mathbf{w} \to \text{GT}), \mathcal{L}_{\text{VALC}}(\mathbf{\Theta}_A; \mathbf{w} \to \text{GT})] \leq \max_{\mathbf{\Theta}} \mathcal{L}_{\text{VALC}}(\mathbf{\Theta}; \mathbf{w} \to \text{GT}) = \mathcal{L}_{\text{VALC}}(\mathbf{\Theta}_G; \mathbf{w} \to \text{GT}). \tag{66}$$

Next, we set $\cup_m\{w_{mj}\}_{j=1}^{N_m}$ to λ_1 and $\cup_m\{w_{mj}\}_{j=N_m+1}^{N_m+N_{m,s}}$ to λ_2, respectively; we rewrite $\sum_i \phi_{mji}\{-\frac{1}{2}(\mathbf{e}_{mj} - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{e}_{mj} - \boldsymbol{\mu}_i) - \log[(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}]\}$ as $f_{1,j}(x)$ for $j \in \cup_m\{1, 2, \ldots, N_m\}$ and $f_{2,j}(x)$ for $j \in \cup_m\{N_m + 1, N_m + 1, \ldots, N_m + N_{m,s}\}$, where x corresponds to $\mathbf{\Theta} \triangleq (\boldsymbol{\phi}, \boldsymbol{\gamma}, \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K)$. By Lemma I.1, we have that

$$\mathcal{L}_{\text{VALC}}(\mathbf{\Theta}_A; \mathbf{w} \to \text{GT}) \leq \mathcal{L}_{\text{VALC}}(\mathbf{\Theta}_G; \mathbf{w} \to \text{GT}). \tag{67}$$

Combining Eq. 66 and Eq. 67 concludes the proof. □

Theorem I.5 shows that under mild assumptions, the attention-based weights can help produce better estimates of $\mathbf{\Theta}$ in the presence of noisy stop-words and therefore learns higher-quality topics from the corpus, improving and interpretability of FLMs.

Theorem I.6 (**Advantage of $\mathbf{\Theta}_A$ in the General Case**). *With Definition I.2 and Definition I.3, comparing $\mathbf{\Theta}_I$, $\mathbf{\Theta}_G$, and $\mathbf{\Theta}_A$ by evaluating them on the marginal log-likelihood of non-stop-words, i.e., $\mathcal{L}_{VALC}(\cdot, w \to GT)$, we have that*

$$\mathcal{L}_{VALC}(\mathbf{\Theta}_I; \mathbf{w} \to GT) \leq \mathcal{L}_{VALC}(\mathbf{\Theta}_A; \mathbf{w} \to GT) \leq \mathcal{L}_{VALC}(\mathbf{\Theta}_G; \mathbf{w} \to GT). \tag{68}$$

Proof. First, by definition one can easily find that $\mathbf{\Theta}_G$ achieves the largest $\mathcal{L}(\cdot; \mathbf{w} \to \text{GT})$ among the three:

$$\max[\mathcal{L}_{\text{VALC}}(\mathbf{\Theta}_I; \mathbf{w} \to \text{GT}), \mathcal{L}_{\text{VALC}}(\mathbf{\Theta}_A; \mathbf{w} \to \text{GT})] \leq \max_{\mathbf{\Theta}} \mathcal{L}_{\text{VALC}}(\mathbf{\Theta}; \mathbf{w} \to \text{GT}) = \mathcal{L}_{\text{VALC}}(\mathbf{\Theta}_G; \mathbf{w} \to \text{GT}). \tag{69}$$

Next, we invoke Lemma I.2 by (1) setting $\cup_m\{w_{mj}\}_{j=1}^{N_m}$ to $\boldsymbol{\lambda}_1$ and $\cup_m\{w_{mj}\}_{j=N_m+1}^{N_m+N_{m,s}}$ to $\boldsymbol{\lambda}_2$, respectively, and (2) rewriting $\sum_i \phi_{mji}\{-\frac{1}{2}(\mathbf{e}_{mj} - \boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{e}_{mj} - \boldsymbol{\mu}_i) - \log[(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}]\}$ as $f_{1,j}(x)$ for $j \in \cup_m\{1, 2, \ldots, N_m\}$ and $f_{2,j}(x)$ for $j \in \cup_m\{N_m + 1, N_m + 1, \ldots, N_m + N_{m,s}\}$, where x corresponds to $\mathbf{\Theta} \triangleq (\boldsymbol{\phi}, \boldsymbol{\gamma}, \{\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K)$. By Lemma I.2, we then have that

$$\mathcal{L}_{\text{VALC}}(\mathbf{\Theta}_A; \mathbf{w} \to \text{GT}) \leq \mathcal{L}_{\text{VALC}}(\mathbf{\Theta}_G; \mathbf{w} \to \text{GT}). \tag{70}$$

Note that because $f_{1,j}(\cdot)$ and $f_{2,j}(\cdot)$ are very close to Gaussian, therefore Assumption 1 and 2 in Lemma I.2 hold naturally under mild regularity conditions.

Combining Eq. 69 and Eq. 70 concludes the proof. □