Vector Quantized Latent Concepts: A Scalable Alternative to Clustering-Based Concept Discovery*

WARNING: The appendix contains some examples, which may be disturbing to the reader.

Anonymous ACL submission

Abstract

Large language models (LLMs) encode rich semantic information in their hidden states, yet it remains difficult to understand what information these internal representations capture. Latent concepts extracted from hidden states offer a promising direction for interpreting LLMs, but existing clustering-based methods face a trade-off: hierarchical clustering produces coherent concepts but is limited to small datasets due to its quadratic memory cost, while K-Means scales efficiently but may yield less semantically coherent concepts. We propose Vector Quantized Latent Concept (VQLC), a discrete concept learning framework that learns a codebook of latent concepts on frozen hidden states. Across 12 dataset-model settings, VQLC stays close to K-Means in computational cost, scales better than hierarchical clustering, and remains competitive in faithfulness, with the clearest gains on decoder-only models. LLMs-based evaluation, qualitative analysis, and a Sparse Autoencoder (SAE) comparison demonstrate that the learned concepts are interpretable and task-relevant.

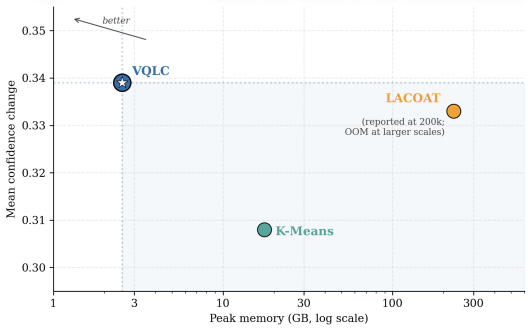


Figure 1: Faithfulness and scalability trade-off: Each point compares a concept-discovery method using the average confidence change across 12 dataset-model settings and the peak memory measured during the scalability evaluation. VQLC achieves the most favorable trade-off, obtaining the highest mean confidence change while requiring the lowest peak memory.[1]

1 Introduction

Current LLMs achieve impressive capabilities, but their internal representations remain difficult to interpret, and the relationship between these representations and the model output is often opaque (Rudin, 2019; Huang et al., 2023b; Shi et al., 2024; Dodge et al., 2021; Sheng et al., 2021). Prior work often explains individual predictions through input attribution methods, such as integrated gradients (IG), SmoothGrad, and SHAP (Ribeiro et al., 2016; Sundararajan et al., 2017; Smilkov et al., 2017; Lundberg and Lee, 2017). These methods identify input tokens that are salient to a model's prediction, providing a token-level explanation of prediction saliency. However, token-level explanations provide limited insight into the high-level semantic information encoded in internal representations.

Another line of work aims to interpret hidden representations through latent concepts extracted from contextual representations (Kim et al., 2018; Ghorbani et al., 2019; Dalvi et al., 2022; Jourdan et al., 2023; Zhao et al., 2024; Yu et al., 2024; Lam et al., 2024; Sharma et al., 2025). The main idea is that a word can have different contextualized representations depending on the context, where each representation captures a different meaning. Representations with similar semantic or functional behavior can be viewed as *concepts* (Dalvi et al., 2022). By organizing representations into latent concepts, concept discovery provides a more structured semantic view of representation space than isolated token-level saliency explanations. Most existing methods discover concepts through post-hoc clustering. For example, LACOAT (Yu et al.,

[0]Anonymous code repository: https://anonymous.4open.science/r/VQLC-46FE.

[1]VQLC and K-Means are measured at 500k tokens. Latent concept attribution (LACOAT) is shown at its largest completed scale at 200k tokens.

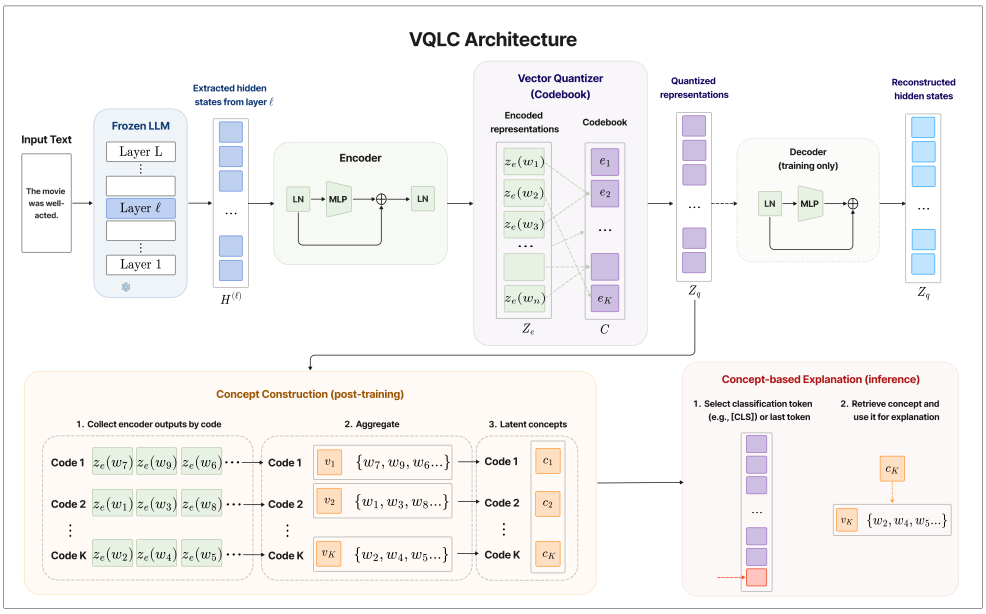


Figure 2: Architecture overview: Contextual token representations extracted from a LLM are encoded and discretized through vector quantization. During training, the decoder reconstructs the token representations. After training, latent concepts are constructed by aggregating encoder outputs with the same code assignments. At inference time, the learned concepts are used to explain the corresponding representations.

2024) applies agglomerative hierarchical clustering to token representations. While hierarchical clustering can discover meaningful concepts, it scales poorly as the number of tokens grows. Alternatively, K-Means is more computationally efficient, but it often provides a weaker trade-off between scalability and semantic concept quality.

Vector quantization provides a natural alternative by mapping continuous hidden states to a finite set of learned discrete codes (van den Oord et al., 2017). Each representation is assigned to its nearest codebook vector in a shared learned codebook. The reconstruction objective encourages these codebook vectors to preserve information from the original representations while grouping tokens that can be represented by similar codebook vectors. This aligns with the goal of concept discovery, where the codebook provides a finite set of codes that can be interpreted as concepts learned from hidden states. The codebook also keeps the cost of assigning a token to a code constant with respect to dataset size, removing the quadratic dependence that limits hierarchical clustering at scale.

Based on this motivation, we propose VQLC, a discrete concept learning framework. It uses a lightweight residual encoder to map token representations into a code space, a vector quantizer to assign them to their nearest vectors in a learnable codebook, and a residual decoder to reconstruct the original hidden states. After training, the learned codebook maps hidden representations to a finite set of latent concepts. Token representations assigned to the same code are treated as sharing a similar semantic facet. Our goal is to discover latent concepts that identify the semantic structure encoded in hidden representations.

We evaluate VQLC against hierarchical clustering and K-Means across models. The results show that VQLC remains close to K-Means in computational cost, scales better than hierarchical clustering in a representative large-scale setting, and remains competitive in concept quality, with the clearest gains on decoder-only models. Figure 1 previews this trade-off between faithfulness and scalability. Overall, this paper contributes a vector quantized framework for scalable latent concept discovery and a multiple evaluations covering scalability, faithfulness, LLMs-based judgments, qualitative analysis, and comparisons with SAEs.

2 Problem Formulation

We consider a language model $\mathbb{M}$ with layers $\ell \in L$. Given an input instance $s = w_1, w_2, \ldots, w_N$, let $h^{(\ell)}_{w_i}$ denote the contextual representation of token w_i at layer ℓ, and $H^{(\ell)}$ denote the set of represen-

tations over the training split $\mathbb{D}_{\text{train}}$. Our goal is to learn a discrete codebook $\mathcal{E} = \{e_1, \ldots, e_K\}$ such that token representations with similar semantic meaning are mapped to the same discrete codes. From the resulting assignments, we derive concept vectors $\mathcal{V} = \{v_1, \ldots, v_K\}$ by averaging the encoded representations assigned to each code, and corresponding latent concepts $\mathcal{C} = \{c_1, \ldots, c_K\}$, where each c_k consists of a concept vector v_k and its associated tokens. Tokens in the same concept are expected to encode similar semantic facets.

3 Methodology

VQLC adapts vector quantization to latent concept discovery on hidden states. As shown in Figure 2, it consists of three main modules: an **encoder** that maps contextual representations extracted from a chosen layer of frozen LLMs into a codebook space; a **vector quantizer** that assigns each encoded representation to its nearest vector in a learnable codebook; and a **decoder** that reconstructs the original representations from the quantized vectors. During training, the codebook learns to support discrete assignment and reconstruction. After training, we freeze the model and run a concept-construction pass over the training split to derive token-to-code assignments and construct concept vectors. At test time, token representations of a test instance are assigned to the learned codebook, and the corresponding latent concepts describe the semantic information encoded in those representations. When the representation chosen for explanation is the one driving model prediction (e.g., the last layer classification token), the assigned concept corresponds to the semantics underlying that prediction.

3.1 Encoder

The encoder is a residual multi-layer perceptron (MLP) that maps contextual representations into the code space. It combines a direct linear projection with a lightweight nonlinear correction branch. The linear path preserves a direct dependence on the original hidden states, while the non-linear branch provides limited reshaping before vector quantization. This design keeps encoded representations close to the original model space while making it suitable for stable discrete assignment.

Let $h_{w_i}^{(\ell)}$ denote the contextual representation of token w_i extracted from layer ℓ. The encoder produces an output $z_e(w_i)$ defined as:

$$u_{w_i}^{(\ell)} = \mathrm{LN}_{\text{in}}\left(h_{w_i}^{(\ell)}\right), \tag{1a}$$

$$r_{w_i}^{(\ell)} = W_2\ \mathrm{GELU}\left(W_1 u_{w_i}^{(\ell)} + b_1\right) + b_2, \tag{1b}$$

$$z_e(w_i) = \mathrm{LN}_{\text{out}}\left(W_{\text{proj}} h_{w_i}^{(\ell)} + b_{\text{proj}} + r_{w_i}^{(\ell)}\right), \tag{1c}$$

where $W_{\text{proj}} \in \mathbb{R}^{d_c \times d}$ denotes the linear transformation into the code space, $W_1 \in \mathbb{R}^{m \times d}$ and $W_2 \in \mathbb{R}^{d_c \times m}$ are the parameters of the nonlinear residual branch, and LN_{in} and LN_{out} denote layer normalization. In our experiments, we set $d_c = d$ and a hidden dimension of $m = 128$. Appendix F provides an ablation study of the residual encoder.

3.2 Vector Quantizer

The vector quantizer maintains a learnable codebook $\mathcal{E} = \{e_1, \ldots, e_K\}$, where each $e_j \in \mathbb{R}^{d_c}$, and assigns each encoder output to its nearest codebook vector. These assignments are then used to organize token representations into discrete groups for latent concept construction.

Codebook Initialization The codebook is initialized using encoder outputs extracted from the training split. We first remove near-duplicate candidates by rounding each coordinate to four decimal places and then deduplicating identical rounded vectors. We select K initial vectors with a greedy farthest-first traversal: the first vector is chosen as the candidate farthest from the candidate mean, and each following vector maximizes its distance to the nearest already selected vector. This initialization spreads the initial codebook vectors across diverse regions of the encoded representation space rather than concentrating them in dense local regions. Additional initialization details are provided in Appendix D. Appendix G compares the farthest-first initialization against random initialization.

Code Assignment During quantization, we compute the squared Euclidean distance between each encoder output $z_e(w_i)$ and every codebook vector:

$$D(z_e(w_i), e_j) = \|z_e(w_i) - e_j\|_2^2. \tag{2}$$

Each encoder output is then assigned to its nearest codebook vector:

$$j^* = \arg\min_j D(z_e(w_i), e_j), \tag{3a}$$

$$z_q(w_i) = e_{j^*}. \tag{3b}$$

During training, we use a straight-through estimator so that the forward pass uses the quantized representation while gradients continue to flow through the encoder. At inference time, token representations are assigned deterministically using the learned codebook.

Codebook Learning We update the codebook using the exponentially moving average (EMA)-based method proposed by Łukasz Kaiser et al. (2018), which provides stable and smooth updates compared to gradient-based approaches. For each codebook vector e_j, EMA maintains both an accumulated usage frequency n_j and a moving average of the encoder outputs assigned to it m_j.

The usage frequency of each vector e_j is calculated as:

$$n_j \leftarrow \lambda n_j + (1-\lambda) \sum_i \mathbb{1}\left[z_q\left(w_i\right) = e_j\right], \quad (4)$$

where $z_q(w_i)$ denotes the codebook vector assigned to token w_i, $\mathbb{1}$ is an indicator function, and λ is a decay parameter.

The corresponding codebook vector e_j is updated towards the average of its assigned encoder outputs:

$$m_j \leftarrow \lambda m_j + (1-\lambda) \sum_i \mathbb{1}\left[z_q\left(w_i\right) = e_j\right] z_e(w_i), \quad (5a)$$

$$e_j \leftarrow \frac{m_j}{n_j}, \quad (5b)$$

where $z_e(w_i)$ is the encoder output for token w_i. The decay parameter λ controls how much the update relies on the previous codebook state, and we set $\lambda = 0.99$ in all experiments.

Although EMA stabilizes codebook learning, some codes may remain persistently unused during training. We maintain an inactivity counter for each code and recover codes that do not receive any assignments for a fixed patience window. When recovery is triggered, we select encoder outputs with the largest assignment errors in the current batch and use them to reinitialize eligible dead codes. Appendix D gives details on the dead code recovery, and reports its ablation study in Appendix I.

Latent Concept Construction After training, we use the trained VQLC components and run a final concept-construction pass over the training split. The EMA codebook is then used to assign each encoded token representation to a discrete code index. For each code k, we derive a concept vector v_k by averaging the encoded representations assigned to that code:

$$v_k = \frac{1}{|A_k|} \sum_{z_e(w_i) \in A_k} z_e(w_i), \quad (6)$$

where A_k denotes the set of encoded token representations assigned to code k. We use v_k from this concept construction pass rather than the EMA codebook vector e_k as a concept vector. The codebook vector e_k is updated during training and reflects an exponentially weighted training history. In contrast, v_k is computed in a final pass with the frozen encoder and directly summarizes the representations assigned to that code for interpretation.

We then define a latent concept explanation c_k as the combination of the concept vector v_k and the tokens assigned to e_k. At test time, token representations from a new input instance are assigned to the learned codebook. We use the latent concept assigned to each input token to explain the information encoded in its representation.

3.3 Decoder

The decoder reconstructs the original contextual representations from the quantized representation. It consists of a direct linear reconstruction path and a nonlinear correction branch. This design preserves a simple reconstruction route from the codebook space while allowing the decoder to model residual nonlinear structure.

Let $z_q(w_i) \in \mathbb{R}^{d_c}$ denote the quantized representation assigned to token w_i. The decoder output $\hat{h}_{w_i}^{(\ell)} \in \mathbb{R}^d$ is defined as:

$$u_q(w_i) = \mathrm{LN}(z_q(w_i)), \quad (7a)$$

$$r_q(w_i) = W_2\ \mathrm{GELU}(W_1 u_q(w_i) + b_1) + b_2, \quad (7b)$$

$$\hat{h}_{w_i}^{(\ell)} = W_{\mathrm{rec}} z_q(w_i) + b_{\mathrm{rec}} + r_q(w_i), \quad (7c)$$

where $W_{\mathrm{rec}} \in \mathbb{R}^{d \times d_c}$ denotes the direct linear reconstruction from the code space to the original hidden-state space, and $W_1 \in \mathbb{R}^{m \times d_c}$ and $W_2 \in \mathbb{R}^{d \times m}$ are the parameters of the nonlinear correction branch. The layer normalization is applied before the nonlinear branch to make the decoder less sensitive to variations in the scale of the quantized representations.

The direct linear reconstruction path captures the dominant structure needed to recover the original

hidden states, while the nonlinear branch complements this reconstruction path by modeling residual nonlinear structure in the reconstruction.

3.4 Training Objective

The overall training objective consists of two components, as shown in Equation 8. The **reconstruction loss** trains the decoder to reconstruct the original contextual representations from the quantized vectors. Because gradients propagate back to both the decoder and encoder, this objective encourages the encoder to produce representations that preserve contextual information while remaining compatible with vector quantization. The **commitment loss** encourages the encoder outputs to stay close to their assigned codebook vectors, which stabilizes discrete assignment during training. The optimizer updates only the encoder and decoder parameters. The codebook is updated through EMA rather than direct gradient steps.

$$\mathcal{L} = \mathcal{L}_{\textbf{rec}} + \beta \mathcal{L}_{\textbf{commit}} \tag{8}$$

$$\mathcal{L}_{\textbf{rec}} = \left\| h_{w_i}^{(\ell)} - \hat{h}_{w_i}^{(\ell)} \right\|_2^2 \tag{9}$$

$$\mathcal{L}_{\textbf{commit}} = \| z_e(w_i) - \mathrm{sg}[z_q(w_i)] \|_2^2 , \tag{10}$$

where sg is the stop-gradient operator, which prevents gradient updates on the codebook. β controls the strength of the commitment constraint. We set $\beta = 0.25$ in our experiments. Appendix J provides a sensitivity analysis for the commitment weight.

4 Experiment Setup

Data We use three sequence classification tasks: ERASER Movie Reviews (Pang and Lee, 2004), Jigsaw Toxicity (cjadams et al., 2017), and AG News (Gulli, 2005). Appendix B provides detailed dataset information.

Models We evaluate on two fine-tuned encoder models, BERT-base-cased (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), and two decoder-only LLMs, Llama-2-7B-chat-hf (Touvron et al., 2023) and Qwen2.5-3B (Team, 2024), in a zero-shot prompting setting. For comparison with SAE, we use pretrained SAEs matched to the corresponding model: Qwen-Scope (Deng et al., 2026) for the Qwen3.5-2B base model (Qwen Team, 2026), and Gemma-Scope (McDougall et al., 2025) for Gemma-3-4B-IT (Gemma Team, 2025). Fine-tuning performance and hardware details are reported in Appendices C and A.

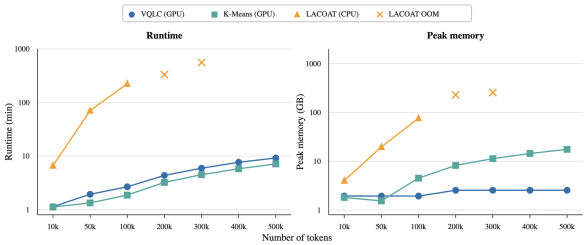


Figure 3: Scalability experiment: runtime and peak memory usage as the number of token representations increases. Dataset: ERASER movie; Model: Qwen.

Representation extraction We use NeuroX toolkit (Dalvi et al., 2023) to extract last-layer token activations, following prior work showing that the last layer tends to contain the most task-aligned representations (Ansuini et al., 2019; Roeder et al., 2021; Yu et al., 2024). The task decision representation is the classification token (e.g. `[CLS]`) for encoder-based models and the final token representation for decoder-only models. Appendix B provides data processing details.

Baselines To test whether VQLC can preserve concept quality while improving scalability, we compare it with two clustering-based baselines:

- **LACOAT** (Yu et al., 2024) uses agglomerative hierarchical clustering to discover latent concepts and trains a classifier to map tokens to concepts. Latent concept vectors are computed by averaging token representations within each cluster.
- **K-Means** applies K-Means clustering directly to token representations and assigns tokens to the nearest concept vector using cosine similarity at inference time. This baseline tests whether VQLC can improve concept quality over scalable centroid-based methods.

Because both VQLC and K-Means are sensitive to initialization, we run experiments with three seeds $\{0, 42, 999\}$. We also include SAEs as a complementary comparison to contrast different forms of explanations. Full VQLC hyperparameter settings are provided in Appendix E.

5 Evaluation

We evaluate VQLC against both K-Means and LACOAT along four axes: scalability, faithfulness, LLMs-based evaluation, and qualitative analysis. We additionally compare VQLC with SAEs to highlight how VQLC differs from mechanistic feature-based explanation methods. Our main question is whether VQLC can serve as a scalable alternative

Dataset	Method	RoBERTa	LLaMA	Qwen
AG News	VQLC	**0.726 ± 0.003**	**0.028 ± 0.000**	0.195 ± 0.004
	K-Means	0.718 ± 0.000	0.010 ± 0.000	**0.203 ± 0.000**
	LACOAT	0.722	0.026	0.196
ERASER	VQLC	0.493 ± 0.002	**0.060 ± 0.000**	**0.090 ± 0.004**
	K-Means	**0.499 ± 0.000**	0.023 ± 0.001	0.063 ± 0.000
	LACOAT	0.484	0.041	0.063
Jigsaw	VQLC	0.466 ± 0.001	**0.123 ± 0.004**	**0.168 ± 0.016**
	K-Means	**0.483 ± 0.000**	0.050 ± 0.003	0.157 ± 0.000
	LACOAT	0.481	0.080	0.167

Table 1: Faithfulness evaluations across datasets and models: Confidence change after removing the assigned concept direction by orthogonal projection. **Higher is better** (↑). The best results are **bolded**. Datasets: AG News, Jigsaw, ERASER movie; Models: RoBERTa, Qwen, LLaMA.

Top: Average Rank (↓)

Dataset	Method	RoBERTa	LLaMA	Qwen
AG News	VQLC	2.198 ± 0.129	2.082 ± 0.016	**1.999 ± 0.063**
	K-Means	2.011 ± 0.078	**1.186 ± 0.053**	2.046 ± 0.061
	LACOAT	**1.773 ± 0.026**	1.914 ± 0.011	2.074 ± 0.061
ERASER	VQLC	**1.821 ± 0.076**	2.027 ± 0.075	**1.915 ± 0.083**
	K-Means	2.139 ± 0.096	2.259 ± 0.140	2.135 ± 0.085
	LACOAT	2.052 ± 0.092	**1.659 ± 0.073**	1.992 ± 0.039
Jigsaw	VQLC	**1.839 ± 0.070**	1.632 ± 0.093	**1.878 ± 0.106**
	K-Means	2.247 ± 0.067	2.796 ± 0.085	2.104 ± 0.096
	LACOAT	2.010 ± 0.058	**1.576 ± 0.047**	2.065 ± 0.053

Bottom: Mean Kendall's W (↑)

Dataset	RoBERTa	LLaMA	Qwen
AG News	0.744 ± 0.036	0.300 ± 0.036	0.741 ± 0.022
ERASER	0.665 ± 0.025	0.617 ± 0.032	0.752 ± 0.047
Jigsaw	0.699 ± 0.042	0.715 ± 0.024	0.778 ± 0.023

Table 2: LLMs-based evaluation: **Top:** Average rank values for each method across datasets and models. **Bottom:** Mean Kendall's W scores over the three-method rankings, measuring inter-LLMs agreement.

to clustering-based concept discovery without compromising concept quality. Figure 1 summarizes this central trade-off. Across 12 dataset-model settings, VQLC maintains lower peak memory usage than both clustering baselines as the number of tokens grows, while also achieving the highest mean faithfulness confidence change. We next examine each evaluation axis in detail.

5.1 Scalability Evaluation

We evaluate scalability on the ERASER movie dataset using the Qwen model. We scale the number of training token representations from 10k to 500k, where each representation is 2048-dimensional. Due to the difference in execution regimes of baselines and VQLC, we report peak GPU memory for VQLC and K-Means, and peak CPU resident memory for LACOAT.

Figure 3 shows that VQLC and K-Means both scale substantially better than LACOAT. VQLC increases from about 1 minute at 10k tokens to about 9 minutes at 500k tokens, while K-Means increases from about 1 minute to about 7 minutes over the same range. LACOAT already requires about 3.5 hours at 100k tokens. The memory results show the clearest difference: VQLC remains nearly constant, increasing from 1.94GB at 10k tokens to about 2.54GB at 500k tokens. K-Means grows from 1.81GB to 17.52GB. LACOAT reaches 228.99GB at 200k tokens before failing at larger scales. This behavior is consistent with the methods' computational structure. VQLC uses mini-batches with a fixed-size codebook, while hierarchical clustering requires pairwise token similarity computations, resulting in quadratic growth with the number of tokens. K-Means avoids this quadratic dependence, but its clustering procedure becomes increasingly memory-intensive as the number of tokens grows.

5.2 Faithfulness Evaluation

We hypothesize that if a latent concept encodes task-relevant information in the task-decision representation, then removing its direction from that representation should lead to a larger change in model output. We test this by ablating the concept direction through orthogonal projection and comparing the perturbed output with the original output. Appendix K provides details on the projection procedure. We report confidence change, which measures the change in prediction confidence after projection, and additionally report predicted label changes in Appendix K.

Table 1 reports the confidence change for RoBERTa, Qwen, and LLaMA across the three datasets. VQLC yields the largest confidence change in 6 out of 9 cases. VQLC achieves its strongest gains in decoder-only models, outperforming the baselines in 5 out of 6 Qwen and LLaMA settings. These results suggest that VQLC is particularly effective at identifying task-relevant concept directions in decoder-only models while remaining competitive with clustering-based baselines in encoder-based settings. The prediction change results and the BERT comparison are reported in Appendix K (Table 12 and Table 11).

5.3 LLMs-Based Evaluation

Following the growing use of LLM-as-a-judge evaluation in LLM research (Zheng et al., 2023; Li et al., 2024; Shi et al., 2025), we use multiple LLMs

Dataset	Method	Confidence Change	# Concepts	Active Rate
AG News	VQLC	**0.444**	399	**0.890**
	SAE-concept top1	0.338	1834	0.375
ERASER	VQLC	**0.296**	398	**0.739**
	SAE-concept top1	0.058	601	0.216
Jigsaw	VQLC	**0.304**	398	**0.779**
	SAE-concept top1	0.093	918	0.199

Table 3: Comparison between VQLC and top-1 SAE features on the Qwen model. We report faithfulness results, the size of the learned concept inventory, and the concept activation rate on the test set.

to judge how well each method's discovered concepts align with the model's output. For each test instance, we provide the sentence, the ground-truth label, the predicted label, and the concept contents generated by VQLC, LACOAT, and K-Means to the LLM judges. Prompt templates are provided in Appendix L. The LLM judges assign ranks from 1 to 3, where lower ranks indicate better alignment. To mitigate potential position bias, we randomly shuffle the order in which the three candidate concept explanations are presented. Ties are allowed.

We use three LLMs: Claude Haiku, Gemini Flash, and DeepSeek (Liu et al., 2024), and evaluate 50 samples for each of the 12 dataset-model settings, for a total of 600 test instances and 1,800 individual LLM rankings before agreement filtering. For each sample, final ranks are determined by majority vote across the three evaluators. Samples without an agreement from at least two judges are excluded. We report the average rank over the resolved samples and measure inter-LLM agreement using Kendall's W, computed per sample over the three method rankings and averaged within each setting. Appendix L gives the aggregation details.

Table 2 reports the average rank (top) and the agreement scores (bottom). VQLC obtains the lowest average rank in 8 of the 12 dataset-model settings. LACOAT is best in 3 settings, and K-Means is best in only 1 setting. These results suggest that VQLC generally yields more task-aligned latent concepts. Agreement scores are generally moderate to strong, although agreement weakens in AG News with the LLaMA model setting. The BERT results are reported in Appendix L.

5.4 Qualitative Evaluation

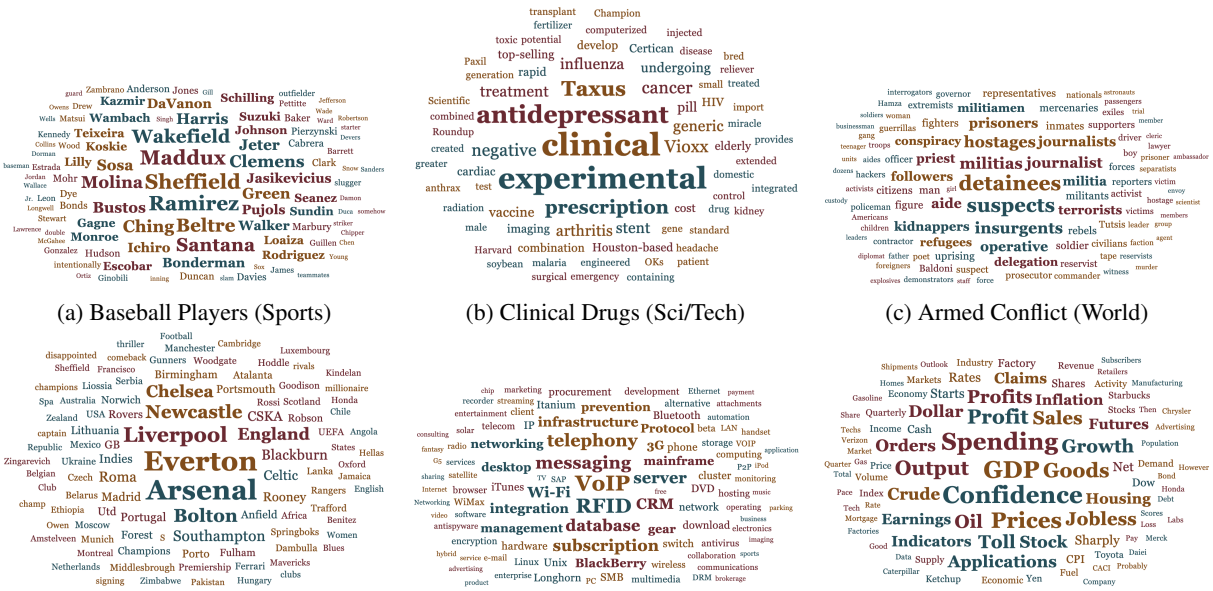
We analyze the latent concepts discovered by VQLC. Each concept is visualized using a word-cloud constructed from the top-100 most frequent tokens. Figure 4 presents examples of latent concepts learned by Qwen on the AG News dataset, with concepts drawn from each of the *Sports*, *Sci/Tech*, *World*, and *Business*. Within the *Sports* category, Figure 4a captures baseball players, with player names such as "Sheffield", "Ramirez", and "Jeter". Figure 4d focuses on football, emphasizing team names such as "Everton", "Arsenal", and "Chelsea". For the *Sci/Tech* category, Figure 4b highlights clinical drugs and cancer treatment, with concept words such as "clinical", "cancer", and "vaccine". Figure 4e captures networking and telecom infrastructure, including terms such as "VoIP", "Wi-Fi", and "networking". Within the *World* category, Figure 4c captures a conflict-related concept focused on insurgents, and militias, with words such as "detainees", "militias", and "insurgents". For the *Business* category, Figure 4f focuses on macroeconomics and market indicators, with words such as "GDP", "Oil" and "Earnings".

We also include qualitative examples to check whether the discovered concepts capture task-related semantic information in Appendix N. For example, Figure 5 shows a correct *Business* prediction, where VQLC retrieves a business concept focused on oil prices, GDP-related economic data, and treasury market reactions, while K-Means and LACOAT return broader economic clusters. Figure 6 shows an incorrect *Business* prediction for a *Science/Tech* instance. All methods reveal that the representation follows a business semantic direction, but VQLC more directly captures corporate competition and consumer-hardware themes.

5.5 Comparison with Sparse Autoencoders

SAEs expose sparse feature directions distributed across the model representation space, while VQLC organizes representations into a set of task-level latent concepts. Because the two methods represent and organize semantic information differently, we treat SAEs as a complementary comparison rather than as another baseline. We compare them across three datasets using the Qwen model. Token representations from the last layer are passed through a pretrained SAE. SAE-concept assigns tokens to feature indices, and forms a concept vector by averaging the hidden representations assigned to the same feature. In SAE-concept top1, each token is assigned to its maximally activated feature. Tokens sharing the same top-1 feature are treated as instances of a concept, and the corresponding averaged hidden representations are used as the ablated vector.

Table 3 characterizes how each method's units



(a) Baseball Players (Sports) (b) Clinical Drugs (Sci/Tech) (c) Armed Conflict (World)

(d) Football (Sports) (e) Network Infrastructure (Sci/Tech) (f) Market Indicators (Business)

Figure 4: Examples of latent concepts identified in the Qwen model for the AG News classification task

behave on the task-decision representation. SAE produces more features, but rarely activates them in the test set. VQLC generates fewer concepts, activates on a larger fraction of test examples, and produces greater confidence changes. Appendix M provides additional results on Qwen and Gemma.

6 Related Work

Early interpretability methods focused on attributing input features to predictions, including IG (Sundararajan et al., 2017), SmoothGrad (Smilkov et al., 2017), SHAP (Lundberg and Lee, 2017), and LIME (Ribeiro et al., 2016). Concept-based methods move beyond individual input features. TCAV (Kim et al., 2018) and CEBab (Abraham et al., 2022) use human-defined concepts, while later works like ACE (Ghorbani et al., 2019) automatically discover concepts via clustering. More recent work extracts latent concepts directly from hidden representations to support higher-level semantic explanations (Rajani et al., 2020; Dalvi et al., 2022; Jourdan et al., 2023; Zhao et al., 2024; Yu et al., 2024; Lam et al., 2024; Sharma et al., 2025).

SAEs, motivated by the superposition hypothesis, have become a central tool in mechanistic interpretability for decomposing representations into sparse feature directions (Cunningham et al., 2023; Härle et al., 2024; Templeton et al., 2024; Lan et al., 2024). Unlike latent concept methods, SAE operate at the feature level and aim to disentangle the representation into monosemantic features rather than task-related concepts. Therefore, we treat them as a complementary comparison rather than as a direct baseline.

Vector quantized-variational autoencoder (VQ-VAE) (van den Oord et al., 2017) learns discrete latent representations via a finite codebook, enabling a natural discretization of the representation space. Prior work has mainly used this idea for representation learning and compression (Łukasz Kaiser et al., 2018; Guo et al., 2020; Yu et al., 2021; Bhardwaj et al., 2022; Huang et al., 2023a). VQLC instead uses vector quantization as a scalable mechanism for discovering latent concepts in LLM hidden states.

7 Conclusion

This work introduced VQLC, a vector quantization-based framework for latent concept discovery. Across 12 dataset-model settings, VQLC demonstrates a favorable balance between concept quality and scalability: it remains competitive with hierarchical and K-Means clustering on faithfulness, with the clearest gains on decoder-only LLMs, while requiring near-constant memory as the number of tokens grows. Compared with SAE, VQLC offers a concept-level explanation that encodes more task-relevant information in the task-decision representation. Overall, these results position VQLC as a practical and scalable approach for latent concept discovery in LLMs.

Limitations

The VQLC framework involves design choices that affect the learned concepts, including the codebook size, commitment weight, and dead-code recovery settings. These choices influence both reconstruction quality and concept granularity, and therefore require careful tuning. Moreover, the current study focuses on classification tasks. Since generative models rely on token-by-token generation and reasoning, extending latent concept-based explanation to analyze intermediate generation behavior is an essential direction for future work.

References

Eldar D Abraham, Karel D'Oosterlinck, Amir Feder, Yair Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. 2022. Cebab: Estimating the causal effects of real-world concepts on nlp model behavior. *Advances in Neural Information Processing Systems*, 35:17582–17596.

Alessio Ansuini, Alessandro Laio, Jakob H Macke, and Davide Zoccolan. 2019. Intrinsic dimension of data representations in deep neural networks. *Advances in Neural Information Processing Systems*, 32.

Rishabh Bhardwaj, Amrita Saha, Steven C.H. Hoi, and Soujanya Poria. 2022. Vector-quantized input-contextualized soft prompts for natural language understanding. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6776–6791, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

cjadams, Jeffrey Sorensen, Julia Elliott, Lucas Dixon, Mark McDonald, nithum, and Will Cukierski. 2017. Toxic comment classification challenge.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Fahim Dalvi, Nadir Durrani, and Hassan Sajjad. 2023. Neurox library for neuron analysis of deep nlp models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 75–83, Toronto, Canada. Association for Computational Linguistics.

Fahim Dalvi, Abdul Rafae Khan, Firoj Alam, Nadir Durrani, Jia Xu, and Hassan Sajjad. 2022. Discovering latent concepts learned in BERT. In *International Conference on Learning Representations*.

Boyi Deng, Xu Wang, Yaoning Wang, Yu Wan, Yubo Ma, Baosong Yang, Haoran Wei, Jialong Tang, Huan Lin, Ruize Gao, Tianhao Li, Qian Cao, Xuancheng Ren, Xiaodong Deng, An Yang, Fei Huang, Dayiheng Liu, and Jingren Zhou. 2026. Qwen-Scope: Turning sparse features into development tools for large language models. *Preprint*, arXiv:2605.11887.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL-HLT '19, pages 4171–4186, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gemma Team. 2025. Gemma 3.

Amirata Ghorbani, James Wexler, James Zou, and Been Kim. 2019. Towards Automatic Concept-based Explanations. *arXiv preprint*. ArXiv:1902.03129 [cs, stat].

Antonio Gulli. 2005. Ag's corpus of news articles.

Daya Guo, Duyu Tang, Nan Duan, Jian Yin, Daxin Jiang, and Ming Zhou. 2020. Evidence-aware inferential text generation with vector quantised variational autoencoder. *Preprint*, arXiv:2006.08101.

Ruben Härle, Felix Friedrich, Manuel Brack, Björn Deiseroth, Patrick Schramowski, and Kristian Kersting. 2024. Scar: Sparse conditioned autoencoders for concept detection and steering in llms. *arXiv preprint arXiv:2411.07122*.

Mengqi Huang, Zhendong Mao, Zhuowei Chen, and Yongdong Zhang. 2023a. Towards accurate image coding: Improved autoregressive image generation with dynamic vector quantization. *Preprint*, arXiv:2305.11718.

Yue Huang, Zihao Huang, Hengxing Cui, Qitian Meng, and Cynthia Rudin. 2023b. Augmenting interpretable models with large language models. *arXiv preprint arXiv:2307.05310*.

Fanny Jourdan, Agustin Picard, Thomas Fel, Laurent Risser, Jean Michel Loubes, and Nicholas Asher. 2023. Cockatiel: Continuous concept ranked attribution with interpretable elements for explaining neural net classifiers on nlp tasks. *arXiv preprint arXiv:2305.06754*.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Ramprasaath Sayres. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the*

35th International Conference on Machine Learning, volume 80 of *Proceedings of Machine Learning Research*, pages 2668–2677.

Michelle S. Lam, Janice Teoh, James A. Landay, Jeffrey Heer, and Michael S. Bernstein. 2024. Concept induction: Analyzing unstructured text with high-level concepts using lloom. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. 2024. Quantifying feature space universality across large language models via sparse autoencoders. *arXiv preprint arXiv:2410.06981*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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:1907.11692*.

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

Callum McDougall, Arthur Conmy, János Kramar, Tom Lieberum, Senthooran Rajamanoharan, and Neel Nanda. 2025. Gemma scope 2 - technical paper.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, page 271–es, USA. Association for Computational Linguistics.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Nazneen Fatema Rajani, Ben Krause, Wengpeng Yin, Tong Niu, Richard Socher, and Caiming Xiong. 2020. Explaining and improving model behavior with k nearest neighbor representations. *arXiv preprint arXiv:2010.09030*.

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.

Geoffrey Roeder, Luke Metz, and Durk Kingma. 2021. On linear identifiability of learned representations. In *International Conference on Machine Learning*, pages 9030–9039. PMLR.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Arushi Sharma, Vedant Pungliya, Christopher J Quinn, and Ali Jannesari. 2025. Analyzing latent concepts in code language models. *arXiv preprint arXiv:2510.00476*.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4275–4293, Online. Association for Computational Linguistics.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. *Preprint*, arXiv:2310.16789.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Integrated Gradients: Axiomatic attribution for deep networks. In the Proceedings of ICML.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, and 3 others. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti

Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. *CoRR*, abs/1711.00937.

Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. 2021. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*.

Xuemin Yu, Fahim Dalvi, Nadir Durrani, Marzia Nouri, and Hassan Sajjad. 2024. Latent concept-based explanation of NLP models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12435–12459, Miami, Florida, USA. Association for Computational Linguistics.

Ruochen Zhao, Tan Wang, Yongjie Wang, and Shafiq Joty. 2024. Explaining language model predictions with high-impact concepts. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 995–1012.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Łukasz Kaiser, Aurko Roy, Ashish Vaswani, Niki Parmar, Samy Bengio, Jakob Uszkoreit, and Noam Shazeer. 2018. Fast decoding in sequence models using discrete latent variables. *Preprint*, arXiv:1803.03382.

A Experiment Setup: Hardware

Our experiments were conducted on high-performance computing (HPC) cluster equipped with NVIDIA H100 and L40 GPUs, and standard CPU resources. We performed all training and inference runs for VQLC and K-Means on H100 GPUs.

B Dataset

Data statistics We conduct experiments on three sequence classification tasks: ERASER Movie Reviews (Pang and Lee, 2004) for sentiment classification task, Jigsaw Toxicity (cjadams et al., 2017) for toxicity classification task, and AG News (Gulli, 2005) for multi-class news topic classification.

Benchmark	Train	Dev	Tags
ERASER Movie	13878	856	2
JIGSAW Toxicity	9000	800	2
AG News	16000	1200	4

Table 4: Data statistics for the benchmarks used in the evaluation.

Processing To keep the comparison with Yu et al. (2024) aligned, we apply the same token filtering criteria, excluding those with frequencies lower than five, and randomly select 20 contextual occurrences of each token. We retain all occurrences for the representative classification tokens. For concept presentation and evaluations, we apply a lightweight post-processing step to concept token lists. We remove empty tokens, punctuation tokens, and common stopwords, while retaining digits and special tokens.

C Finetuning Performance of 12-layered pre-trained Models

We finetuned two 12-layered pre-trained models: BERT-base-cased (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) with their standard data split.

Benchmark	BERT	RoBERTa
ERASER Movie	93.74	95.98
JIGSAW Toxicity	91.30	91.66
AG News	94.88	95.18

Table 5: The fine-tuned performance of models across all benchmarks. Model: BERT, RoBERTa

Dataset	Model	Encoder	Confidence Change	% of Prediction Change
ERASER	RoBERTa	Residual MLP	**0.4926 ± 0.0015**	**0.4560 ± 0.0089**
		Linear	0.4869 ± 0.0013	0.4171 ± 0.0085
	Qwen	Residual MLP	**0.0897 ± 0.0030**	**0.0829 ± 0.0011**
		Linear	0.0782 ± 0.0045	0.0716 ± 0.0020
AG News	RoBERTa	Residual MLP	**0.7260 ± 0.0028**	**0.8750 ± 0.0029**
		Linear	0.7168 ± 0.0011	0.8089 ± 0.0167
	Qwen	Residual MLP	**0.1950 ± 0.0045**	**0.1960 ± 0.0045**
		Linear	0.1906 ± 0.0033	0.1857 ± 0.0015

Table 6: Ablation analysis of the nonlinear encoder branch layer. Each cell reports confidence change and the percentage of predictions changed after orthogonal projection.

Dataset	Model	Initialization	Confidence Change	% of Prediction Change
ERASER	RoBERTa	Farthest-First	**0.4926 ± 0.0015**	**0.4560 ± 0.0089**
		Random	0.4783 ± 0.0011	0.2932 ± 0.0229
	Qwen	Farthest-First	**0.0897 ± 0.0037**	**0.0829 ± 0.0014**
		Random	0.0777 ± 0.0046	0.0728 ± 0.0036
AG News	RoBERTa	Farthest-First	**0.7260 ± 0.0028**	**0.8750 ± 0.0029**
		Random	0.7132 ± 0.0006	0.8008 ± 0.0029
	Qwen	Farthest-First	0.1950 ± 0.0045	0.1960 ± 0.0045
		Random	**0.2264 ± 0.0008**	**0.2264 ± 0.0025**

Table 7: Comparison between Farthest-First and Random initialization. Each cell reports confidence change and the percentage of predictions changed after orthogonal projection.

D Methodology: Additional Details

Farthest-first initialization Let $\mathcal{Z} = \{z_1, \ldots, z_M\}$, $z_i \in \mathbb{R}^{d_c}$ denote the encoded token representations collected for initialization, and let $\tilde{\mathcal{Z}} = \{\tilde{z}_1, \ldots, \tilde{z}_{M'}\}$ be the deduplicated candidates after rounded-value deduplication. Let $\bar{z}$ denote the mean of the deduplicated candidate pool. We choose the first vector as:

$$s_1 = \arg\max_{\tilde{z} \in \tilde{\mathcal{Z}}} \|\tilde{z} - \bar{z}\|_2^2. \tag{11}$$

This initialization favors a point that is well separated from the center of the candidates and provides a strong starting anchor in the encoded space. After selecting $t-1$ vectors, the next vector is chosen as the point whose distance to its nearest previously selected vector is maximal:

$$s_t = \arg\max_{\tilde{z} \in \tilde{\mathcal{Z}} \setminus \mathcal{S}_{t-1}} \min_{s \in \mathcal{S}_{t-1}} \|\tilde{z} - s\|_2^2, \tag{12}$$

where $\mathcal{S}_{t-1} = \{s_1, \ldots, s_{t-1}\}$. We repeat this process until K vectors are selected, and use them to initialize the codebook:

$$e_k \leftarrow s_k, \qquad k = 1, \ldots, K. \tag{13}$$

Deadcode Recovery We maintain an inactivity counter for each code. At training step t, let

$$u_j^{(t)} = \sum_i \mathbb{1}\left[z_q(w_i) = e_j\right] \tag{14}$$

denote the number of assignments received by code e_j in the current batch. We update the inactivity counter $d_j^{(t)}$ as

$$d_j^{(t)} = \begin{cases} d_j^{(t-1)} + 1, & \text{if } u_j^{(t)} \le \delta, \\ 0, & \text{otherwise,} \end{cases} \tag{15}$$

where δ is a small usage threshold. A code is considered eligible for recovery once its inactivity counter exceeds a patience threshold T. For each valid token, we define the assignment error as the squared Euclidean distance to its assigned code:

$$a_i^{(t)} = \|z_e(w_i) - z_q(w_i)\|_2^2. \tag{16}$$

Recovered codes are reinitialized using encoder outputs with the largest assignment errors in the current batch.

E VQLC hyperparameters

For the encoder, the nonlinear branch uses a hidden dimension of 128. During vector quantization, the

codebook size is set to $K = 400$, the commitment weight to 0.25, and the EMA decay to 0.99. We use dead-code recovery with a zero-assignment threshold, a patience of 100 training steps, and at most two code recoveries per step. Appendix H analyzes the codebook size choice.

F Ablation Study: Encoder

We have an ablation study to evaluate the usefulness of the nonlinear encoder branch layer. We ablate the nonlinear encoder branch by replacing the default residual-MLP encoder (**ResidualMLP**) with a linear projection encoder (**Linear**). We compare the faithfulness performance. Table 6 shows that the nonlinear encoder branch layer provides benefits across both encoder-based models and decoder-only models.

G Comparison of Codebook Initialization Methods

We study the effect of codebook initialization by comparing the default farthest-first initialization against a random initialization strategy. In our method, the codebook is initialized from encoded training representations using the farthest-first mechanism in Section 3.2. In random initialization, the codebook is initialized from a uniform distribution. Table 7 indicates farthest first initialization is favored in most settings, which provides the more stable overall choice.

H Sensitivity Analysis: Codebook Size

We have a sensitivity analysis of the codebook size on the AG News dataset using the RoBERTa model. Table 8 shows that the codebook size of 400 achieves the best performance. In addition, 400 corresponds to the number of clusters used in the evaluation experiments of the LACOAT baseline (Yu et al., 2024).

Setting	Confidence Change	% of Prediction Change
200	0.720 ± 0.002	0.818 ± 0.010
400 (default)	$\mathbf{0.729 \pm 0.003}$	$\mathbf{0.879 \pm 0.007}$
800	0.712 ± 0.001	0.789 ± 0.014

Table 8: Codebook size sensitivity on the AG News/RoBERTa setting under faithfulness evaluation. Each cell reports confidence change and the percentage of predictions changed after orthogonal projection. **Higher is better** (↑). The best results are **bolded**.

I Ablation Study: Dead-code Recovery

We have an ablation study to evaluate the usefulness of dead-code recovery. We compare the faithfulness performance with and without the dead-code recovery mechanism. Table 9 shows that dead-code recovery mechanism provides better performance.

Setting	Confidence Change	% of Prediction Change
Enabled (default)	$\mathbf{0.729 \pm 0.003}$	$\mathbf{0.879 \pm 0.007}$
Disabled	0.714 ± 0.011	0.834 ± 0.016

Table 9: Dead-code recovery ablation on the AG News/RoBERTa setting under faithfulness evaluation. **Higher is better** (↑). The best results are **bolded**.

J Sensitivity Analysis: Commitment Weight

Weight	Confidence Change	% of Prediction Change
0.10	0.715 ± 0.003	0.806 ± 0.012
0.25 (default)	$\mathbf{0.729 \pm 0.003}$	$\mathbf{0.879 \pm 0.007}$
0.50	0.717 ± 0.001	0.812 ± 0.008

Table 10: Commitment weight sensitivity on the AG News-RoBERTa setting under faithfulness evaluation. **Higher is better** (↑). The best results are **bolded**.

The commonly recommended commitment weight in prior literature is 0.25. In addition, we conduct a sensitivity analysis of different commitment weights on the AG News dataset using the RoBERTa model. We evaluate how varying the weight affects the faithfulness performance. Table 10 demonstrates that a commitment weight of 0.25 achieves the best overall performance. It has the highest confidence change and prediction percentage change. These results indicate that 0.25 is the most stable choice.

K Faithfulness Evaluation

Orthogonal Projection for Concept Ablation To measure whether an assigned latent concept encodes a direction that the underlying model relies on for prediction, we remove the corresponding latent concept direction from a sentence representation via orthogonal projection. Let v_j denote the concept vector, and let $h_{w_i}^{(\ell)}$ denote the sentence representation at layer ℓ.

The projection of $h_{w_i}^{(\ell)}$ onto v_j is defined as:

$$\mathrm{proj}_{v_j}\left(h_{w_i}^{(\ell)}\right) = \frac{h_{w_i}^{(\ell)} \cdot v_j}{\|v_j\|^2}\, v_j \qquad (17)$$

% of Prediction Change (↑)					
Dataset	**Method**	**RoBERTa**	**BERT**	**LLaMA**	**Qwen**
AG News	VQLC	**0.875 ± 0.003**	0.879 ± 0.003	**0.354 ± 0.002**	0.196 ± 0.004
	K-Means	0.827 ± 0.000	0.336 ± 0.000	0.350 ± 0.017	**0.204 ± 0.000**
	LACOAT	0.829	**0.881**	0.350	0.190
ERASER	VQLC	0.456 ± 0.011	0.480 ± 0.006	**0.369 ± 0.003**	**0.083 ± 0.001**
	K-Means	**0.482 ± 0.000**	**0.519 ± 0.000**	0.265 ± 0.004	0.061 ± 0.000
	LACOAT	0.350	0.510	0.315	0.059
Jigsaw	VQLC	0.436 ± 0.011	0.498 ± 0.023	**0.136 ± 0.006**	**0.158 ± 0.016**
	K-Means	0.445 ± 0.000	**0.543 ± 0.000**	0.036 ± 0.009	0.150 ± 0.002
	LACOAT	**0.477**	0.506	0.058	0.157

Table 11: Additional faithfulness results measured by percentage of predictions changed after projection-based concept ablation. **Higher is better** (↑). The best results are **bolded**.

Dataset	**Method**	**BERT**
AG News	VQLC	**0.749 ± 0.004**
	K-Means	0.504 ± 0.000
	LACOAT	0.748
ERASER	VQLC	0.495 ± 0.005
	K-Means	**0.508 ± 0.000**
	LACOAT	0.502
Jigsaw	VQLC	0.479 ± 0.001
	K-Means	**0.482 ± 0.000**
	LACOAT	0.480

Table 12: Additional faithfulness results for BERT, measured by confidence change after projection-based concept ablation. **Higher is better** (↑). The best results are **bolded**.

This projection isolates the component of the representation that aligns with the concept direction.

We then remove this concept direction by subtracting the projection from the original representation:

$$h^{(\ell)}_{w_i,\perp} = h^{(\ell)}_{w_i} - \mathrm{proj}_{v_j}\left(h^{(\ell)}_{w_i}\right) \tag{18}$$

The resulting representation $h^{(\ell)}_{w_i,\perp}$ preserves information of the original representation except for the contribution along the concept direction.

Additional Faithfulness Results Table 12 and Table 11 report additional BERT confidence change and prediction change of all dataset-model settings. For the BERT model, VQLC and LACOAT remain close on all three datasets. For prediction change metrics, the overall pattern is consistent with confidence change results: decoder-only settings remain the advantages for VQLC. VQLC has comparable performance for encoder-based settings.

L LLMs-Based Evaluation

Average Ranking Formula Let S denote the set of evaluation samples, A the set of method approaches, and E the set of the LLM judges. For each sample $i \in S$, each LLM judge $e \in E$ assigns a ranking value $r^{(e)}_{i,a} \in \{1, 2, 3\}$ for each method $a \in A$, where a lower rank indicates a better explanation.

For each sample i and method a, we have an aggregated rank value $\hat{r}_{i,a}$ by majority vote across the evaluators. A valid majority voting rank is defined only when at least two of the three evaluators assign the same rank value to that method. Let $\hat{S} \subseteq S$ denote the set of resolved samples. The average rank of method a is then computed as:

$$\mathrm{AvgRank}(a) = \frac{1}{|\hat{S}|} \sum_{i \in \hat{S}} \hat{r}_{i,a}, \tag{19}$$

where $|\hat{S}|$ denotes the number of resolved samples. A lower average rank value indicates that the method is preferred more often by the LLM evaluators.

Additional LLMs-based Evaluation Results Table 13 shows additional LLMs-based evaluation results for BERT across all datasets. VQLC achieves the best results in all settings. In additional, agreement scores are generally moderate to strong.

Prompt Template We use the following prompt template for each LLM judge. For each sample, the order of the three candidate explanations is randomly shuffled to reduce position bias. The prompt always includes the input text, the model prediction, and three candidate concept-based explanations. Concepts are presented in one of two formats.

Dataset	Method	Avg. Rank (↓)	Kendall's W (↑)
Jigsaw	VQLC	**1.721 ± 0.167**	0.797 ± 0.011
	K-Means	2.483 ± 0.102	
	LACOAT	1.751 ± 0.122	
ERASER	VQLC	**1.506 ± 0.118**	0.804 ± 0.025
	K-Means	1.624 ± 0.037	
	LACOAT	2.870 ± 0.038	
AG News	VQLC	**1.850 ± 0.083**	0.726 ± 0.079
	K-Means	2.008 ± 0.027	
	LACOAT	2.032 ± 0.037	

Table 13: Additional LLMs-based evaluation results for BERT. Average rank measures evaluator preference, and Kendall's W measures inter-LLMs agreement over the three-method rankings

.

If more than half of a concept consists of special tokens such as `[CLS]` , we randomly sample five such tokens and provide their original sentences. Otherwise, we provide up to ten frequent tokens from the concept to fit the API context limit. Thus, concept content is presented in one of these two forms: (i) five sample sentences for special token dominated concepts, or (ii) a list of up to ten tokens.

```
You are an expert judge of local concept-based
explanations.
Your task is to rank candidate explanations
for why the model made its prediction for a
single input.
Input Text: [INPUT_TEXT]
Model Prediction: [MODEL_PREDICTION]
Please evaluate the candidate explanations
using the following principles:
1. The best explanation should identify the
most important semantic reason for why the
model made its prediction.
2. Prefer explanations whose concept content
matches the specific topic, event, entity, or
semantic pattern in the input.
3. Do not reward generic topical overlap if
another explanation is more specific and
directly relevant to the prediction.
4. When concepts are weak, noisy, or only
loosely related to the prediction, rank them
lower.
Candidate Explanations:
Explanation 1:
Name: [METHOD_NAME_1]
Concept Content: [CONCEPT_CONTENT_1]
Explanation 2:
Name: [METHOD_NAME_2]
Concept Content: [CONCEPT_CONTENT_2]
Explanation 3:
Name: [METHOD_NAME_3]
Concept Content: [CONCEPT_CONTENT_3]
Return a JSON object with one field "ranking",
mapping each explanation name to a rank from 1
to 3, where 1 is best. Ties are allowed.
Also include a short field "reason" explaining
the ranking.
```

Dataset	Method	Confidence Change	% Pred. Change	# Concepts	Active Rate
AG News	VQLC	**0.4444**	**0.6045**	399	**0.8897**
	SAE-concept top1	0.3379	0.3878	1,834	0.3751
	SAE-concept top5	0.1668	0.2943	176,251	0.0868
Jigsaw	VQLC	**0.3035**	**0.5676**	398	**0.7789**
	SAE-concept top1	0.0930	0.1071	918	0.1993
	SAE-concept top5	0.1830	0.1154	62,149	0.1007
ERASER	VQLC	**0.2960**	**0.4417**	398	**0.7387**
	SAE-concept top1	0.0579	0.0206	601	0.2163
	SAE-concept top5	0.1131	0.0874	42,572	0.1096

Table 14: Full comparison between VQLC, SAE-concept top1, and SAE-concept top5 on the Qwen model. SAE-concept first assigns tokens to SAE feature indices and then uses the average hidden state of the assigned tokens as the concept direction. We report faithfulness results, concept inventory size, and active rate.

M Comparison with SAE

Table 14 and Table15 show the full comparison results between VQLC and SAE on the Qwen model. For SAE-concept, we use the average vector of the hidden representations assigned to the same feature. In the top1 setting, each token is assigned to its most highly activated feature. In the top5 setting, each token is assigned to its unordered top5 SAE feature set. This makes the SAE-concept comparable to VQLC, where each explanation is represented by the average vector over its assigned token representations.

SAE-feature instead directly ablates the active SAE features by setting the selected activations to zero. The edited latent is then decoded back into hidden presentations and forward pass to measure the resulting performance effect. In general, SAEs produces more features, but their top1 and top5 features are activated in fewer test examples. VQLC use a smaller codebook and is activated on most inputs. VQLC concepts encode more task-relevant information used by the model for prediction in the Qwen setting.

Dataset	Method	Confidence Change	% Pred. Change
AG News	SAE-feature top1	0.0300	0.0060
	SAE-feature top5	**0.2066**	**0.2089**
Jigsaw	SAE-feature top1	0.1027	0.0153
	SAE-feature top5	**0.3274**	**0.2844**
ERASER	SAE-feature top1	0.0433	**0.0328**
	SAE-feature top5	**0.0623**	0.0291

Table 15: Additional faithfulness results for SAE-feature top1 and SAE-feature top5 on the Qwen model. SAE-feature directly uses SAE feature directions, without averaging assigned token representations.

We also compare VQLC with SAE in `Gemma-3-4b-IT` at the last layer (see Table 16 and

Table 17). These results broadly support the same pattern as the Qwen comparison. VQLC gives a much smaller concept inventory and a higher active rate. For SAE-concept, ERASER movie is the only exception where SAE-concept top5 gives a slightly higher performance change in faithfulness, but it has large concept inventory and a much lower active rate. For the SAE-feature, it is stronger in AG News. This shows that individual SAE feature can perturb output in some cases, but they do not provide the same consistently active concept as VQLC.

Dataset	Method	Confidence Change	% Pred. Change	# Concepts	Active Rate
AG News	VQLC	**0.1803**	**0.1758**	397	**0.8060**
	SAE-concept top1	0.0209	0.0192	2,788	0.4706
	SAE-concept top5	0.0478	0.0427	420,462	0.0541
Jigsaw	VQLC	**0.1296**	**0.1224**	398	**0.8040**
	SAE-concept top1	0.0438	0.0383	2,346	0.3176
	SAE-concept top5	0.0739	0.0744	280,026	0.0421
ERASER	VQLC	0.0292	0.0269	399	**0.7519**
	SAE-concept top1	0.0002	0.0000	1,583	0.3797
	SAE-concept top5	**0.0312**	**0.0292**	170,980	0.0487

Table 16: Full comparison between VQLC, SAE-concept top1, and SAE-concept top5 on the Gemma model.

Dataset	Method	Confidence Change	% Pred. Change
AG News	SAE-feature top1	0.1974	0.1817
	SAE-feature top5	**0.2290**	**0.1933**
Jigsaw	SAE-feature top1	**0.0392**	**0.0357**
	SAE-feature top5	0.0244	0.0167
ERASER	SAE-feature top1	0.0123	0.0129
	SAE-feature top5	**0.0129**	**0.0140**

Table 17: Additional faithfulness results for SAE-feature top1 and SAE-feature top5 on the Gemma model.

N Examples of Qualitative Evaluation

Comparing Latent Concept Methods Figure 5 shows a correct prediction example, where the sample sentence discusses economic contraction, rising oil prices, and a widening grade gap, and the model correctly predicts the *Business News* label. The concept identified by VQLC is the most closely aligned with the sample sentence. The content retrieved directly focuses on oil prices, GDP-related economic data, and treasury market reactions. K-Means also captures economically related content, but the concept is broader, mixing oil-price movements with general stock-market reactions. LACOAT discovers generic macroeconomic indicators such as jobless claims, factory output, and durable-goods orders, which are less directly related to the oil prices and economic slowdown emphasized in the sample sentence. Overall, VQLC provides a more precise latent concept to explain the model predictions for this sample.

Figure 6 in the Appendix provides an example where the model incorrectly predicts *Business News* label for a *Science/Technology News*. The sample sentence concerns PC customer satisfaction and hardware support. All three methods reveal that the model relies on a business-oriented semantic direction for this prediction. K-Means finds the concept is centered more on specific contracts, acquisitions, and product launches. LACOAT provides the concept related to the broader enterprise IT and branding. The concept identified by VQLC focuses on business-oriented technology themes such as corporate competition and consumer hardware, which more closely match the semantic content of the input. It helps to understand why the model incorrectly predicts *Business News* instead of *Science/Technology News* for this sample explicitly.

Figure 7 shows a correct prediction example for the Jigsaw dataset using the Qwen model. The concepts of both LACOAT and K-Means mainly contains toxic sentences. In contrast, most sentences in the concept of VQLC are non-toxic. VQLC captures the discourse behaviors and non-toxic semantic, making it more suitable for interpreting model prediction.

Figure 8 displays a case where a non-toxic sentence is misclassified as a toxic label. Although all three methods provide mixed concepts that contain both benign and aggressive examples, the underlying representation region itself is semantically ambiguous. K-Means associate emotional expression with strongly toxic lexical samples. LACOAT groups together a broad range of conflict discussion. In contrast, VQLC produces a concept focused on the semantics of argumentative interaction.

Discovered Latent Concepts Figure 9 presents examples of latent concepts learned by the RoBERTa model on the AG News dataset, with two concepts drawn from each of the *Sports*, *Sci/Tech*, and *World*. Within the *Sports* category, Figure 9a and Figure 9d show two distinct concepts. Figure 9a captures Olympic sports, with terms related to swimming and track events such as “Phelps”, “freestyle”, “200-meter”, and “Thorpe”. Figure 9d focuses on baseball game recaps, emphasizing team names, inning-level events, and scoring actions, in-

Sentence:

Now It #39;s Official: Economy Shrunk The US economy slowed more sharply in the second quarter than first thought as oil prices rose and the trade gap swelled, the government said on Friday in a report that confirmed momentum faltered in the spring.

Ground Truth: 2 (Business News) **Prediction:** 2 (Business News)

Salient Token: Answer:729

LACOAT

1. US weekly jobless claims rise The US Labor Department said Thursday the number of workers filing first-time applications for unemployment benefits rose by 10,000 last week.
2. Fed officials ease crude concerns A recovery in the world's largest economy is likely to continue, despite oil prices surging to record levels, US Federal Reserve officials say.
3. Durable-goods orders up; new-home sales down WASHINGTON - Factory orders for costly manufactured goods in July recorded the biggest gain in four months. Meanwhile, new-home sales slid, according to a pair of reports that offered a mixed picture of economic activity.
4. Greenspan Says Effects of Japan Currency Sales Hard to Gauge The effect of Japan #39;s currency sales this year and last on the value of the yen are difficult to judge, #39; #39; Federal Reserve Chairman Alan Greenspan said.
5. August Mid-Atlantic Factory Output Slows NEW YORK (Reuters) - Output at U.S. Mid-Atlantic factories slowed in August as new orders dropped sharply, suggesting recent weakness in the economy is likely to persist into coming months.

K-Means

1. Nikkei Down, Oil Worries Hit Exporters (Reuters) Reuters - Tokyo's Nikkei fell 1.66 percent by\midday on Monday, extending losses into a third day as another\surge in oil prices deepened worries about the global economic\impact and knocked down exporters such as Toyota Motor Corp.
2. Wal-Mart warning dents Wall Street rally NEW YORK Stocks rose Monday as investors cheered a drop in oil prices but a disappointing sales forecast by Wal-Mart Stores stoked concerns about a slowing economy.
3. Durable Goods, Housing Data Trip Stocks NEW YORK (Reuters) - U.S. stocks were slightly lower on Wednesday after government reports for July on durable goods orders and new home sales cast some doubt on the strength of the economy.
4. Treasuries Trim Gains After GDP Data NEW YORK (Reuters) - Treasury debt prices trimmed early gains on Friday after U.S. economic data proved to be not quite as weak as some bulls had been betting.
5. Stocks Up Despite Oil Nearing \$50 a Barrel NEW YORK - Stocks edged higher in early trading Friday as investors tried to uncouple share prices from skyrocketing oil costs, even as crude reached new highs and flirted with \$50 per barrel. A barrel of light crude was quoted at \$48.97, up 27 cents, on the New York Mercantile Exchange after setting a new intraday record earlier in the session at \$49.40...

VQLC

1. Oil Hits \$44 as U.S. Crude Stocks Slide NEW YORK (Reuters) - Oil prices soared to \$44 a barrel on Wednesday, ending an eight-day streak of losses as a U.S. government report showed crude oil stocks falling to the lowest level in five months.
2. Manufacturing Activity Up for 15th Month Manufacturing activity rose in August for the 15th consecutive month, but at a slower rate than reported in July, according to a monthly report released Wednesday by the Institute for Supply Management, a research group.
3. Treasuries Trim Gains After GDP Data (Reuters) Reuters - Treasury debt prices trimmed early\gains on Friday after U.S. economic data proved to be not quite\as weak as some bulls had been betting.
4. Dollar Bounces; Eye on Data, Greenspan The dollar bounced off recent four-week lows against the euro and yen on Monday in thin August trade with investors focusing on US data and a speech by the Federal Reserve chief later in the week.
5. Oil Dips Below \$46 as Iraq Flow Rises Oil prices dipped below \$46 on Tuesday in a third day of falls as a more optimistic Iraq export picture helped unwind some of the supply worries that have lifted the market to historic levels.

Figure 5: Correct prediction example from AG Qwen.

cluding words such as “marlines”, “Cubes”, “innings”, and “homered”. For the *Sci/Tech* category, Figure 9b highlights biomedical research, with concept words such as “mice”, “brain”, “gene”, and “cancer”. Figure 9e captures computer hardware and networking, including terms such as “Cisco”, “Dell”, “laptop”, and “networking”. Within the *World* category, Figure 9c focuses on geopolitics and nuclear diplomacy, with words such as “Pyongyang”, “nuclear”, “Tehran”, and “weapons”. Figure 9f captures the Shiite uprising in Iraq through specific conflict-related words such as “Najaf”, “al-Sadr”, “cleric”, and “militants”.

Sentence:

Intuit Posts Wider Loss After Charge (Reuters) Reuters - Intuit Inc. (INTU.O), maker of\the No. 1 U.S. tax presentation software TurboTax, on Wednesday\posted a wider quarterly loss after taking a goodwill\impairment charge during its seasonally weaker fourth quarter.

Ground Truth: 3 (Science/Technology News) **Prediction:** 2 (Business News)

Salient Token: Answer:207

LACOAT

1. Before-the-Bell: Accredo Health Drops NEW YORK (Reuters) - Shares of Accredo Health Inc. <A HREF="http://www.investor.reuters.com/FullQuote.aspx?ticker=ACDO.O target=/stocks/quickinfo/fullquote">ACDO.O</A> sank in trade before the opening bell after the provider of pharmacy and treatment services for patients with chronic illnesses slashed its fiscal 2005 profit forecast.
2. Heinz Profit Meets Forecasts; Shares Rise CHICAGO (Reuters) - Ketchup maker H.J. Heinz Co. <A HREF="http://www.investor.reuters.com/FullQuote.aspx?ticker=HNZ.N target=/stocks/quickinfo/fullquote">HNZ.N</A> on Tuesday said quarterly profit rose 7.8 percent, excluding a year-earlier tax benefit, as demand for new products such as Ore-Ida Extra Crispy fries boosted sales.
3. Stocks to Watch Wednesday (Reuters) Reuters - Stocks to watch on Wednesday: FOREST\LABORATORIES INC.
4. Last Call for Investors to Bid on Google (Reuters) Reuters - Time is running out for prospective\investors to submit their offers to buy shares of Google Inc,\the Web's No. 1 search company.
5. Google has strong first day of public trading After a bumpy ride toward becoming a publicly traded company, Google Inc. finally saw its stock start trading on the Nasdaq exchange at around noon Eastern Daylight Time Thursday and with a strong opening at \$100.01, up from its \$85 initial offering price. The stock, which trades under the GOOG ticker symbol, closed at \$100.34, up 18 percent.

K-Means

1. After Months of Hoopla, Google Debut Fits the Norm In the stock #39;s first day of trading, investors bought, sold and flipped shares at a furious pace, with the price ending just above \$100 - 18 percent above where it started.
2. Deutsche Bank, Weisel to Pay \$100 Mln to End Probe (Update2) Deutsche Bank AG and Thomas Weisel Partners LLC will pay a total of \$100 million to settle allegations they published misleading stock research to win investment-banking business, becoming
3. Wall Street set for soft start Wall Street is set to open lower on Monday, as a dearth of corporate news and wariness ahead of economic data later in the week keep investors out of the market.
4. Google IPO Moves Ahead Despite Playboy Interview Google Inc. (GOOG) forged ahead with its IPO auction Friday, even as the online search engine leader acknowledged a newly published magazine interview with its founders contained misleading information.
5. Wal-Mart Lowers Sales Forecast for August LITTLE ROCK, Ark. - Wal-Mart Stores Inc., the world's largest retailer, lowered its sales forecast for August, citing slow back-to-school sales and the effects of Hurricane Charley...

VQLC

1. S amp;P raises outlook on Ingersoll-Rand Standard amp; Poor #39;s on Thursday raised its outlook on Ingersoll-Rand Co. to quot;positive quot; from quot;stable, quot; citing the proposed sale of its Dresser-Rand business unit.
2. Oracle extends offer for PeopleSoft again Oracle yesterday extended its controversial bid for enterprise resource software provider PeopleSoft, while both companies await the final judge #39;s ruling in the Department of Justice (DoJ) anti-competition case.
3. US Stocks Gain for a Third Day, Led by Home Depot, Retailers Aug. 17 (Bloomberg) -- US stocks rose for a third straight day after an unexpected drop in consumer prices encouraged investors that the Federal Reserve may be able to slow the pace of interest rate increases. Benchmark indexes pared gains as oil prices ...
4. MCI Offers to Pay Part of States' Claims (Reuters) Reuters - MCI Inc. (MCIP.O) has offered to pay\states from 2 cents to 30 cents on the dollar to settle charges\it owes them as much as #36;2 billion in back taxes, people\familiar with the matter said on Thursday.
5. Krispy Kreme Scales Back as Profit Dives NEW YORK (Reuters) - Krispy Kreme Doughnuts Inc. <A HREF="http://www.investor.reuters.com/FullQuote.aspx?ticker=KKD.N target=/stocks/quickinfo/fullquote">KKD.N</A> on Thursday posted sharply lower quarterly earnings, signaled that the second half of the year would be weak and scaled back store openings after it was battered by the low-carbohydrate diet craze.

Figure 6: Incorrect prediction example from AG Qwen.

Sentence: ::::I am curious how it got rated B

Ground Truth: 0 (Non-toxic Comment) **Prediction:** 0 (Non-toxic Comment)

Salient Token: Answer:348

LACOAT

1. Fuck off please. Isn't my fault he is a leech on society with no job, friends, or scholastic ambition. Is it really vandalism if it's the truth I think not. kthnxbai
2. This indo-fijian dude is a helpful janitor, but sucks terribly at adding useful content to the bag o'trivia. Moreover, after whining because he failed his request for adminship here, he was promoted on the retarded-english wiki and on meta-wiki. Get DerHexer's tools so you can revert faster you negrito.
3. Please post your ignorant comments here so I can edit them and make you look silly!
4. fuck this motherfucker asshole This asshole should not have his own page, nor should any other hijackers. Giving them their own pages on Wikipedia is glorifying their evil. I hope this motherfucker rots in hell. FUCK MOHAMMED ATTA, SCUM OF THE EARTH.
5. Why are you ignoring what I just said though? There was clearly a consensus to not include wild and speculative claims in the infobox during the straw poll and subsequent discussion where all editors were taking part. The later discussion is meaningless because we can't reasonably expect all editors to stay on a DRN for such a long amount of days. The discussion had already gotten stale. Are you disputing any of this? Thanks.

K-Means

1. what facts you fucking idiot. the facts from turkish security forces. scumbag.
2. I based the close on the arguments in the DRV itself. But I reworded the last few words to make it more clear.v/r - P
3. What I THINK THIS ARTICLE NEED MORE AND MORE PICTURES , SEE People's Republic of China ANDSouth Korea ANDJapan ANDSingapore , THESE ARTICLE HAVE A LOT OF PICTURES ,SHOULDN SOMEBODY TO ADD MORE PICTURES?
4. Yer all a bunch of wankers! ppppttttthhhh! nipple...
5. Binksternet is an anti-Semitewho loves telling lies about Jews and who supports Islamic terrorism. Binksternet is an anti-Semitewho loves telling lies about Jews and who supports Islamic terrorism.

VQLC

1. digg trolls vandalize Volume_license_key#FCKGW you can't get full keys added to this article, so you add them to the Volume license key article. nice. Please note that you need to sign your posts on talk pages like this one, as you have previously been told on your talk page. Also please remember to keep headings neutral and on topic. Try to add comments that are useful and relevant; if you notice vandalism, see Wikipedia:Vandalism for details on how to revert or report vandalism. Thanks! (talk) censorship at wikipedia all of you who were saying that wikipedia is being censored are right. i'm apparently going to get banned soon (see my talk page) because some admin doesn't like anon users. welcome to the night of long knives. Talk pages are created for the purpose of discussing or suggesting changes to the article, not your personal blog or soap box. Agreed, this section is superfluous and should be deleted censoring allegations of censorship. nice.
2. " que? oh lighten up. seriously, the Randstad or however the hell you spell it already has ""Hot Issues"" which is already risgay. why not make it smokin' hot. i do not see the harm. "
3. Sorry about the prod I thought it had been removed without good reason, so I reverted the removal, my apologies.
4. without any information about the forest elephant specifically, it does not need it's own article.
5. To Bad you dont have the balls to sign in, I cant wait to see what happens to you when your V-W gets broadsided or rear ended, either way, your dead. I dont drive an Ambulance, I drive a Monte Carlo, unlike traitors like yourself who favor imports. I live in Royal Oak Michigan. Go ahead. Sign in, I would be more than happy to vandalize your sorry ass. Ha... this is so hilarious. I love reading these discussions. Remember one thing guys, Arguing online is like running in the special olympics, even if you win, you're still a retard !! ha...

Figure 7: Correct prediction example from Jigsaw Qwen.

Sentence:

" ==Liberalism sidebar AGAIN== Deuces, the liberalism sidebar has once again expanded its lists even though the code did not change. This is the sidebar from hell. I've tried everything I can think of to fix it...but nothing worked. The fate of the world is at stake Deuces. If you know how, fix that freaking sidebar please. "

Ground Truth: 0 (Non-toxic Comment) **Prediction:** 1 (Toxic Comment)

Salient Token: hell / freaking

LACOAT

1. " New straw poll You are a user who responded to RFC: Use of logos on sports team pages. As someone interested in the discussion a new straw poll has been laid out to see where we currently stand with regards to building a consensus. For the sake of clarity, please indicate your support or opposition (or neutrality) to each section, but leave discussion to the end of each section. talk "
2. " List of celebrity and notable guest appearances in Doctor Who There's a new AfD nomination for an article you've previously discussed. Please stop by to voice your opinions again. '''''' | "
3. I bet the only thing you like better than sucking cock is editing on fagopedia
4. Well on a free-for-all like Wikipedia I'm not sure how else I'm supposed to interact.
5. Who the HELL do you think you are? How dare you revert my edit to Quillacollo Province. I was making a friend smile, and you ruined it. In doing that, you also ruined my life. And my friends life. I'm going to cut your dog in half.

K-Means

1. *::::I disagree. A general genre shouldn't NEED a citation, because so many people agree with it. When have you seen a song by Slipknot that wasn't some type of metal (excluding ""Snuff"")? Perhaps you're taking this a little too far. There aren't going to be many dead-giveaway citations because they expect their readers to know that kind of stuff. Until we do find one, which may take a while, why don't we just leave it at Metal and from there just leave it alone? "
2. You are such a faggot ..dont tell me what to do...ill fuck you up
3. Yes, it must be. I didn't know you were anti-Semitic. I definately didn't know you hated blacks. Strange.
4. MuYou suck asshole ch
5. Category:Salt pans I have nominated category:Salt pans for merger into category:Salt flats. As the editor who created the category, you may wish to comment at the discussion. I would have emailed you to draw this to your attention, but you have not enabled this feature in your preferences. -(talk)

VQLC

1. }} {{unblock|Dude come on its hilarious
2. Semi-protected edit request on 9 March 2014 You can't semi-block unregistered people. We have the right to change things, as well as the freedom of speech, so unblock it now. 165.138.93.203
3. Banging Dudes Hi there. Grockl, please refrain from banging dudes on Wikipedia. Wikipedia has a No banging dudes policy, and you are flagrantly disregarding this rule at the expense of other members. If possible, please hang yourself immediately.
4. IT'S YOU AND PEOPLE LIKE YOU WHO ARE VANDALIZING! How dare you write nonsense about his 'alleged' criminal activities. Based on what facts? Milosevic's inditements? Cos only Milosevic and Kostunica has caled him a criminal. You better get that part of the article out!!!
5. * Antagonism Keep it up, ******. Keep on following me around Wikipedia and thwarting my edits. Does it make you quake with pleasure to do so, I wonder? Was it because I put your panties in a bind before? In this case I'm talking about your speedy delete of my Phantasy Star IV image, which was uploaded in no violation of WP policy, and for the purpose of improving the quality of the article. I do believe your actions constitute an abuse of power. Your little e-cock stroking edit-block a few days ago was lost on me, as I was away on vacation. I hope the idea of it really got you off, though. But yeah, keep on antagonizing me, see how far it takes you. We'll see how the higher-ups regard your behavior, including your warning about civility including a LACK thereof (i.e. calling me a ""dick""). The censored word, in case you were wondering, has two G's, ends in a T, begins with an F, and is an expletive wrongfully targeting homosexuals but rightfully targeting POS Wikipedia editors like yourself. ''''''T C W "

Figure 8: Incorrect prediction example from Jigsaw Qwen.

(a) Olympic Sports (Sports)

(b) Biomedical Research (Sci/Tech)

(c) Geopolitics (World)

(d) Baseball (Sports)

(e) Technology (Sci/Tech)

(f) Shiite Uprising in Iraq (World)

Figure 9: Examples of latent concepts identified in the RoBERTa model for the AG News classification task