Crafting Large Language Models for Enhanced Interpretability

Chung-En Sun [1] **Tuomas Oikarinen** [1] **Tsui-Wei Weng** [1]

Abstract

We introduce the Concept Bottleneck Large Language Model (CB-LLM), a pioneering approach to creating inherently interpretable Large Language Models (LLMs). Unlike traditional black-box LLMs that rely on post-hoc interpretation methods with limited neuron function insights, CB-LLM sets a new standard with its built-in interpretability, scalability, and ability to provide clear, accurate explanations. This innovation not only advances transparency in language models but also enhances their effectiveness. Our unique Automatic Concept Correction (ACC) strategy successfully narrows the performance gap with conventional black-box LLMs, positioning CB-LLM as a model that combines the high accuracy of traditional LLMs with the added benefit of clear interpretability — a feature markedly absent in existing LLMs.

1. Introduction

Large Language Models (LLMs), such as BERT (Devlin et al., 2019) and GPT3 (Brown et al., 2020), have become instrumental in advancing Natural Language Processing (NLP) tasks. However, the inherent opacity of these models poses significant challenges in ensuring their reliability, particularly when outcomes are based on unclear or flawed reasoning. This lack of transparency complicates the effort to debug and improve these models.

Recent efforts in the field have primarily focused on post-hoc interpretations of neurons within LLMs (Bills et al., 2023; Dalvi et al., 2019; Antverg & Belinkov, 2022). Given a learned LLM, these studies aim to elucidate the inner workings of black-box language models by finding post-hoc explanations for neurons (Bills et al., 2023; Lee et al., 2023; Dalvi et al., 2019; Antverg & Belinkov, 2022). Nevertheless, the explanations derived from these methods often do not accurately align with the activation behaviors of the neurons. Moreover, they often fall short in offering clear directions for model editing or debugging, thereby limiting their practical application in correcting outputs.

Motivated by these limitations, we propose the Concept Bottleneck Large Language Model (CB-LLM) – the first concept bottleneck model (CBM) for NLP tasks. Our method can transform any pretrained language model into a CBM with an inherently interpretable concept bottleneck layer and a prediction layer. Our contributions are as follows:

- We present the first CBM framework for LLMs that scales to large text classification benchmarks. Our CB-LLM encapsulates the best of both worlds: it matches the high accuracy of traditional black-box models across multiple datasets while also offering clear interpretability, a feature absent in existing LLMs.
- Our proposed pipeline to build CB-LLM is fully automatic and efficient: it eliminates the need for human-annotated concept labels, and the computational cost is almost the same as the standard fine-tuning. Furthermore, our proposed Automatic Concept Correction (ACC) strategy efficiently boosts the performance of our CB-LLM in terms of both accuracy and faithfulness evaluation.
- Our CB-LLM matches the accuracy of the standard black-box models and achieves a $1.39\times$ higher average rating compared to the random baseline on the faithfulness evaluation. This suggests that our CB-LLM provides high-quality interpretability without sacrificing performance.

2. Background and related works

Post-hoc neuron analysis for NLP. Post-hoc analysis is the most popular method for comprehending the inner workings of black-box language models. Traditionally, this analytical approach comprises several methodological categories, each offering distinctive insights. Visualizing-based methods (Li et al., 2016) involve the graphical representation of neuron activations and manually identify the underlying concepts. Corpus-based methods (Kádár et al., 2017;

[1] UC San Diego. Correspondence to: Chung-En Sun <cesun@ucsd.edu>, Tsui-Wei Weng <lweng@ucsd.edu>.

Preprint. Present at ICML 2024 *Mechanistic Interpretability (MI)* Workshop, Vienna, Austria. Our code will be available at: https://github.com/Trustworthy-ML-Lab/Concept-Bottleneck-LLM.

Antverg & Belinkov, 2022) involve aggregating statistical information derived from data activations to uncover the roles of neurons. Probing-based methods (Dalvi et al., 2019) entail training classifiers over activations to pinpoint neurons associated with predefined linguistic concepts. Causation-based approaches (Lakretz et al., 2019) identify neurons through controlling perturbations and observing prediction change.

Recently, with the advent of Large Language Models (LLMs) such as GPT, (Bills et al., 2023) proposes utilizing GPT4 to generate explanations for GPT2 neurons and simulating the real neuron activations. The subsequent comparison of simulated and actual activations facilitates an evaluation of the quality of explanations. Additionally, (Kroeger et al., 2023) delves into the capability of LLMs to explain other predictive models. Given a dataset and model to explain, they perform in-context learning (ICL) to prompt LLMs to give explanations and highlight that LLMs can generate faithful explanations and consistently outperform previous post-hoc methods.

While the notion of utilizing LLMs for post-hoc explanations appears promising, the challenge lies in the fact that the intricate nature of a neuron from a black-box language model may not be effectively articulated through natural language, potentially resulting in oversimplification and overlooking complex behaviors. Moreover, the considerable computational resources required for this approach restrict its applicability to explaining only a small fraction of neurons in a language model. In contrast, our proposed CB-LLM offers intrinsic interpretability without the need to obtain post-hoc interpretations.

CBM in image classification. Recently, Concept Bottleneck Models (CBMs) (Koh et al., 2020) have been revisited in the context of image classification tasks. CBMs incorporate a concept bottleneck layer (CBL), where individual neurons are designed to learn specific concepts that are interpretable by humans. CBL is then followed by the final fully connected layer responsible for making predictions. Training a CBM typically involves utilizing human-annotated concept labels, enabling the CBL to make multilabel predictions for these concepts when presented with an image. However, a significant limitation arises from the computational expense of constructing an entire CBM from scratch and the dependency on human-annotated concept labels. Addressing this challenge, (Yüksekgönül et al., 2023) introduced a computationally economical algorithm that transforms any image classifier into a CBM. This transformation is achieved by leveraging Concept Activation Vectors (CAV) (Kim et al., 2018) or the multi-modal CLIP (Contrastive Language-Image Pretraining) model (Radford et al., 2021). It's important to note that their approach requires either concept labels to obtain CAV or restricting the backbone to the CLIP image encoder if concept labels are unavailable, which does not fully resolve the limitation. Recognizing this constraint, (Oikarinen et al., 2023) proposed a Label-free CBM, which learns a CBM without relying on concept labels by leveraging the interpretability tool CLIP-Dissect (Oikarinen & Weng, 2023).

Despite the extensive exploration of CBMs in the field of image classification tasks, to the best of our knowledge, there is still no CBM that scales to large NLP benchmarks. Consequently, our work focuses on learning an efficient, automated, and high-performance CBM specifically for LLMs.

Sentence embedding models with contrastive learning. Contrastive learning has emerged as a predominant technique in training sentence embeddings, replacing the traditional approach of augmenting word2vec (Mikolov et al., 2013) with n-gram embeddings. A noteworthy method, SimCSE (Gao et al., 2021), has demonstrated success in semantic textual similarity (STS) tasks. They employ supervised contrastive learning to train the sentence embedding model with Natural Language Inference (NLI) datasets. This involves using entailment pairs as positive instances and contradiction pairs as hard negatives.

In our work, we leverage sentence embedding models trained with contrastive learning for Automatic Concept Scoring (ACS). This method yields high-quality concept scores without any human effort, which is a key step in building CB-LLM.

3. CB-LLMs: Building Interpretable Large Language Models

Existing large language models (LLMs), despite their impressive performance, often lack interpretability. This section introduces a methodology that addresses this critical gap by employing a novel strategy. Our method transforms black-box pretrained models into interpretable entities, specifically converting them into Concept Bottleneck Large Language Models (CB-LLMs). This transformation significantly boosts interpretability without sacrificing performance. While our approach is adaptable to both fine-tuning pretrained models and training LLMs from scratch, we predominantly focus on building from pretrained models, as fine-tuning is a more common practice in NLP due to computational costs.

Our proposed method consists of four steps and is illustrated in Figure 1:

1. **Concept Generation:** given a text classification task, generate a concept set for each class by prompting modern language models.
2. **Automatic concept scoring (ACS):** leverage sentence

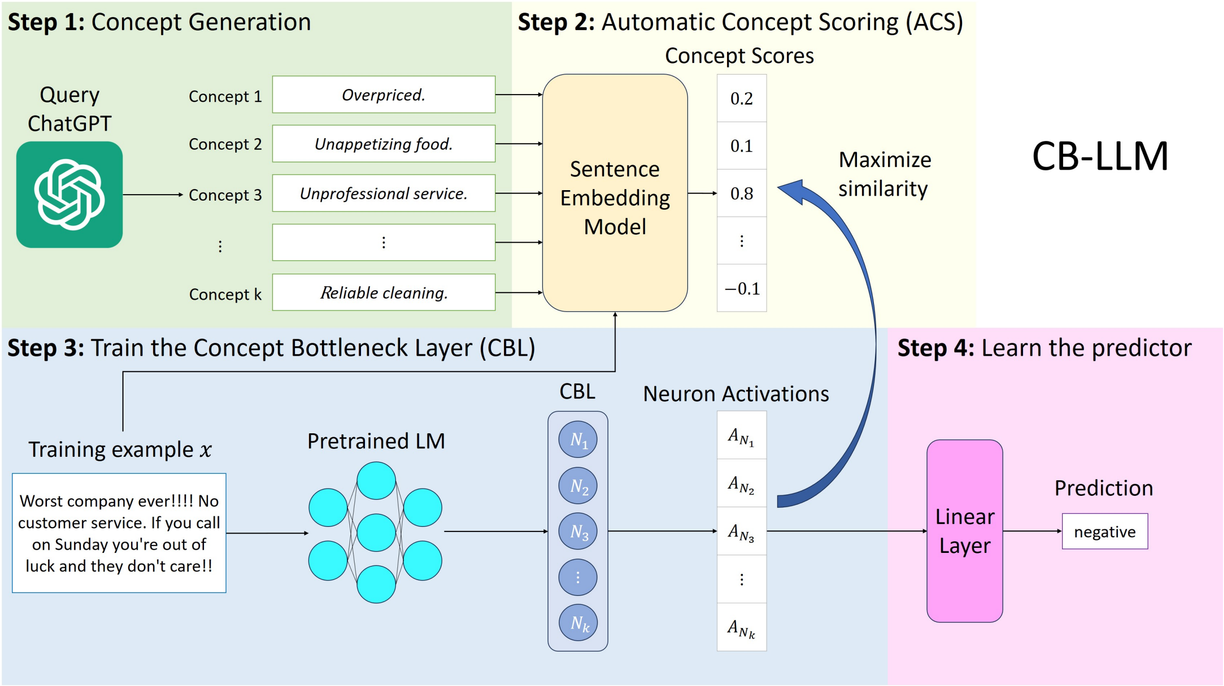


Figure 1. The overview of our CB-LLM.

embedding models to measure the similarity between each concept in the concept set and each text sample in the dataset.

3. **Train the Concept Bottleneck Layer:** learn the concept mapping from uninterpretable features to human-interpretable concepts by maximizing the similarity between the neuron activations and the concept scores.

4. **Learn the predictor:** train the final linear layer to make predictions for the downstream tasks.

The details of steps 1 and 2 can be found in Section 3.1 and 3.2 respectively. The details of steps 3 and 4 can be found in Section 3.3.

3.1. Concept generation

The first step is to generate a set of concepts related to the downstream task. To automate this process, we leverage ChatGPT (Ouyang et al., 2022) as a replacement for the domain experts. For any text classification dataset $\mathcal{D}$ with n classes/labels, we prompt ChatGPT to generate the concept subset $\mathcal{S}_i$ for each class i. Then, the concept set $\mathcal{C}$ is the union of $\mathcal{S}_i$, $\mathcal{C} = \bigcup_{i=0}^{n-1} \mathcal{S}_i$. The following is the template we use to prompt ChatGPT to get $\mathcal{S}_i$:

- "Here are some examples of key features that are often present in a {*class*}. Each feature is shown between the tag <example></example>.
  - <example>{*example 1*}</example>
  - <example>{*example 2*}</example>
  - <example>{*example 3*}</example>
  - <example>{*example 4*}</example>

  List {*concept size per class* $|\mathcal{S}_i|$} other different important features that are often present in a {*class*}. Need to follow the template above, i.e.<example>features</example>."

We use four human-designed concepts as examples for in-context learning. This prompting style requires only n queries to ChatGPT to obtain the full concept set and can be done efficiently through the web interface provided by OpenAI. More prompting details can be found in Appendix A.6.

3.2. Automatic Concept Scoring (ACS)

After generating the concept set $\mathcal{C}$, the next step is to obtain the concept labels for a given text sample x in dataset $\mathcal{D}$. Typically, this stage requires involving domain experts and can be time-consuming. To overcome this challenge, we propose an automatic scoring strategy by utilizing sentence embedding models, which can measure the similarity between each concept and any text sample x. We name this strategy as Automatic Concept Scoring (ACS) and describe the details below.

For any sentence embedding model $\mathcal{E}$ that encodes a text sample into a fixed-size embedding, we calculate the con-

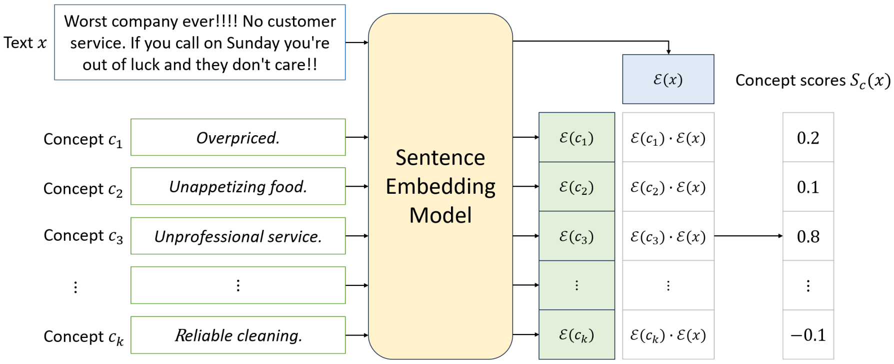


Figure 2. The process of Automatic Concept Scoring (ACS) through sentence embedding models.

cept scores $S_c(x) \in \mathbb{R}^k$ for text sample x by calculating the following:

$$S_c(x) = [\mathcal{E}(c_1)\cdot\mathcal{E}(x), \mathcal{E}(c_2)\cdot\mathcal{E}(x), ..., \mathcal{E}(c_k)\cdot\mathcal{E}(x)]^\top, \quad (1)$$

where $\mathcal{E}(x) \in \mathbb{R}^d$ denotes the text embedding generated by $\mathcal{E}$, c_i is the i-th concept in the concept set $\mathcal{C}$, and k is the size of the concept set. Each component of the vector $S_c(x)$ represents the degree of association between the text x and the concept c_i. This vector will be used as the learning target for CBL in the next section. The process of getting $S_c(x)$ is shown in Figure 2.

It's worth noting that, for a dataset with m text examples $\mathcal{D} = \{x_1, ..., x_m\}$ and a concept set with k concepts $\mathcal{C} = \{c_1, ..., c_k\}$, our ACS strategy requires only $m + k$ inferences to label the entire dataset. This stands in stark contrast to the more expensive alternative of utilizing zero-shot classification models trained with NLI datasets, which would require mk inferences to label each pair of $(x_i, c_j), i \in \{1, ..., m\}, j \in \{1, ..., k\}$.

We use the off-the-shelf sentence embedding models **all-mpnet-base-v2** from Huggingface (Wolf et al., 2019) for ACS. **all-mpnet-base-v2** is fine-tuned from pretrained **MPNet** model (Song et al., 2020) with self-supervised contrastive learning objective using 1 billion sentence pairs. It serves as a computationally efficient option for ACS.

3.3. Learning CB-LLM

After ACS, we have the concept scores $S_c(x)$ for every text example x in dataset $\mathcal{D}$. Our CB-LLM is trained based on these concept scores and the class labels of $\mathcal{D}$. The training process unfolds in two sequential steps: first, a Concept Bottleneck Layer (CBL) is trained to learn the concepts, and subsequently, a linear predictor is trained to make the final predictions.

Training the concept bottleneck layer (CBL): In this step, the goal is to force the neurons in CBL to activate in correlation with the pattern of concept scores. We first send the text sample x into a pretrained LM f_{LM} and use CLS pooling to get a fix size embedding $f_{\text{LM}}(x) \in \mathbb{R}^d$. Then, the CBL f_{CBL} projects the embeddings into a k dimensional interpretable embedding $f_{\text{CBL}}(f_{\text{LM}}(x)) \in \mathbb{R}^k$. Note that f_{CBL} can be a non-linear function and this will not hurt the interpretability, as our focus is solely on the activation behaviors of the neurons in the last layer of CBL. To force the last k neurons in the f_{CBL} learn the k concepts, we maximize the similarity between $f_{\text{CBL}}(f_{\text{LM}}(x))$ and $S_c(x)$ for every x:

$$\max_{\theta_1,\theta_2} \frac{1}{|\mathcal{D}|} \sum_{x\in\mathcal{D}} Sim\big(f_{\text{CBL}}(f_{\text{LM}}(x;\theta_1);\theta_2), S_c(x)\big), \quad (2)$$

where $Sim : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}$ can be any similarity function, θ_1 and θ_2 are the parameters of the pretrained LM and the CBL respectively.

Learning the predictor: After training the CBL, the k neurons from the last layer of CBL learn the corresponding k concepts. Let A_N be the neuron activations from the last layer neurons of CBL $A_N(x) = f_{\text{CBL}}(f_{\text{LM}}(x))$, we set all the negative activations of $A_N(x)$ to zero through a ReLu function $A_N^+(x) = \text{ReLu}(A_N(x))$. We remove the negative activations as the negation of a concept introduces ambiguity (e.g., it is unclear whether the negative activations imply the absence of a concept or the negation of the semantic meaning of a concept). After obtaining A_N^+, we train a final linear layer with sparsity constraint to make predictions:

$$\min_{W,b} \frac{1}{|\mathcal{D}|} \sum_{x,y\in\mathcal{D}} \mathcal{L}_{\text{CE}}(WA_N^+(x) + b, y) + \lambda R(W), \quad (3)$$

where $W \in \mathbb{R}^{n\times k}$ is the weight matrix and $b \in \mathbb{R}^n$ is the bias vector of the final linear layer, y is the label of x, and $R(W) = \alpha||W||_1 + (1-\alpha)\frac{1}{2}||W||_2^2$ is the elastic-net regularization, which is the combination of ℓ_1 and ℓ_2 penalty. Generally, a sparse final layer makes the CBM more interpretable. We will discuss the effect of sparsity in Section 5.

4. Automatic Concept Correction

While ACS offers an efficient way to provide pseudo labels (concept scores), its correctness is dependent on the performance of the sentence embedding model. This introduces a limitation wherein the concept scores may not align with human reasoning, consequently impacting the learning of the CBL and introducing a trade-off in performance. Notably, this challenge is prevalent in recent CBM works that do not rely on human-assigned concept labels.

To address this challenge, we proposed Automatic Concept Correction (ACC), a technique leveraging the knowledge from ChatGPT to improve the quality of concept scores generated by ACS. As shown in our experiment (Table 1), ACC can effectively boost the performance of CBM to a comparable level with black-box models.

Here, we describe the details of ACC. Recall that in Section 3.1, we generate the concept set $\mathcal{C} = \bigcup_{i=0}^{n-1} \mathcal{S}_i$ for dataset $\mathcal{D}$ with n classes, where $\mathcal{S}_i$ is the concept subset for class i. We define the mapping $\mathcal{M} : c \to \{0, ..., n-1\}$ which maps a concept $c \in \mathcal{C}$ to a class:

$$\mathcal{M}(c) = \begin{cases} 0 \text{ if } c \in \mathcal{S}_0 \\ 1 \text{ if } c \in \mathcal{S}_1 \\ \vdots \\ n-1 \text{ if } c \in \mathcal{S}_{n-1} \end{cases} \tag{4}$$

For any text sample x in $\mathcal{D}$, let y be the class label of x and $S_c(x)$ be the concept scores generated by sentence embedding model $\mathcal{E}$ as in Eq. (1). The key idea is to replace $S_c(x)$ with new concept scores $S_c^{\text{ACC}}(x)$, which are corrected by the ACC procedure. The new concept scores $S_c^{\text{ACC}}(x)$ are defined as follows:

$$S_c^{\text{ACC}}(x)_i = \begin{cases} \mathcal{E}(c_i)\mathcal{E}(x), \text{ if } \mathcal{E}(c_i)\mathcal{E}(x) > 0, \mathcal{M}(c_i) = y \\ 0, \text{ otherwise} \end{cases} \tag{5}$$

where $S_c^{\text{ACC}}(x)_i$ is the i-th component of vector $S_c^{\text{ACC}}(x)$. ACC filters out the negative concept scores and forces every component of $S_c^{\text{ACC}}(x)$ to be zero when the corresponding concept c_i and text sample x belong to different classes. This is achievable because we prompt ChatGPT to generate the concept set for each class separately, thereby providing information about the association of concepts with their respective classes.

We utilize ACC to correct inaccurate concept scores before training the CBL, leading to a significant improvement in the accuracy of CB-LLM, which matches and, in certain cases, even surpasses those of finetuned black-box models. Further details on the accuracy of CB-LLM will be discussed in Section 5.1. Unlike prior studies focusing on leveraging test-time intervention to correct the predictions of CBM, ACC occurs before the training of CBM and does not necessitate information about the testing set or any human knowledge. Additionally, our ACC strategy does not require any extra queries to ChatGPT and can be executed with almost zero time cost.

5. Experiment results

In this section, we evaluate our CB-LLM in terms of three crucial aspects: *Accuracy*, *Efficency*, and *Faithfulness*. These aspects are pivotal as our goal is to ensure that CB-LLM achieves high accuracy with minimal additional cost while providing reasonable and human-understandable explanations.

Setup. We conduct experiments on the standard text-classification benchmarks:

- **SST2 (Socher et al., 2013):** comprise 6920 training samples, 872 validation samples, and 1821 test samples of movie reviews with positive and negative classes.
- **Yelp Polarity (YelpP) (Zhang et al., 2015):** comprise 560,000 training samples and 38,000 test samples of Yelp reviews with positive and negative classes.
- **AGnews (Zhang et al., 2015):** comprise 120,000 training samples and 7,600 test samples of news articles with 4 classes.
- **DBpedia (Lehmann et al., 2015):** comprise 560,000 training samples and 70,000 test samples from DBpedia 2014 with 14 classes.

We generate 208 concepts for SST2, 248 concepts for YelpP, 216 concepts for AGnews, and 476 concepts for DBpedia. We use **`RoBERTa-base`** (Liu et al., 2019) pretrained model with 768 output dimensions as the backbone for learning CB-LLM, and compared our CB-LLM with the finetuned **`RoBERTa-base`** (standard black-box model).

5.1. Accuracy of CB-LLM

The test accuracy is shown in Table 1. In general, our CB-LLMs demonstrate high accuracy across various datasets, including large ones such as YelpP and DBpedia. The CB-LLM implementation without ACC already achieves high accuracy: only a 1~5% gap compared to the standard black-box model. This gap can be further eliminated: it can be seen that our ACC strategy, described in Section 4, improves the accuracy significantly to the level of the standard black-box model. This indicates that ACC can effectively correct inaccurate concept scores and enhance learning on the given task. As for the effect of the sparse final layer, we do not observe a large performance drop after incorporating the sparsity constraint. In fact, CB-LLM with a sparse final layer, when combined with ACC, sometimes exhibits better accuracy than the counterpart with only ACC. This observation suggests that our ACC strategy works well with the

Table 1. Test accuracy of CB-LLM. Our CB-LLMs achieve nearly identical performance as the standard black-box model after undergoing ACC. Numbers highlighted in blue indicate accuracy surpassing that of the standard black-box model.

Method	Dataset			
	SST2	YelpP	AGnews	DBpedia
Ours:				
CB-LLM	0.9138	0.9358	0.8989	0.9828
CB-LLM w/ sparse final layer	0.9094	0.9327	0.8972	0.9742
CB-LLM w/ ACC	0.9473	**0.9805**	0.9462	**0.9925**
CB-LLM w/ ACC & sparse final layer	**0.9478**	0.9803	0.9467	**0.9925**
Baseline (standard black-box):				
Roberta-base finetuned	0.9418	0.9803	**0.9478**	0.9922

Table 2. The time cost of ACS and learning CB-LLM. Training CB-LLM requires only a little additional time cost compared to finetuning the black-box language models.

Time cost (hours)	Dataset			
	SST2	YelpP	AGnews	DBpedia
Automatic Concept Scoring (ACS):				
mpnet ACS	0.0024	1.6172	0.2455	1.6578
Finetuning model:				
CB-LLM	0.0984	8.9733	2.0270	9.1800
Standard black-box	0.0289	8.9679	1.3535	9.1996

sparsity constraint on the final layer. Overall, our CB-LLMs sometimes achieve higher accuracy than the standard black-box model (highlighted in blue in Table 1), showcasing the possibility of building an interpretable model without incurring a trade-off in performance loss.

5.2. Efficiency of CB-LLM

The time cost of Automatic Concept Scoring (ACS) and finetuning language model is shown in Table 2. Our ACS strategy takes about 1.6 hours on the largest YelpP and DBpedia dataset when using `all-mpnet-base-v2` as the sentence embedding model. The training time of CB-LLM is approximately equivalent to the time cost of finetuning the standard black-box model. These results indicate that we incur only a small overhead of time cost while significantly improving interpretability through the incorporation of a human-interpretable CBL and a final linear layer.

5.3. Faithfulness of CB-LLM

It is important for an interpretable model to make predictions based on human-understandable and faithful logic. Hence, in this section, we introduce our faithfulness evaluation design and present the results obtained through large-scale human evaluation for our CB-LLM.

We define two metrics to evaluate the faithfulness :

1. **Activation Faithfulness:** This evaluates if the activations of neurons in CBL align with the corresponding concepts they have learned. For a given neuron in CBL, its activation over each text sample in the dataset can be extracted. Subsequently, *Activation Faithfulness* can be evaluated by manually inspecting whether the highly activated samples are related to the concept represented by the given neuron.

2. **Contribution faithfulness:** This evaluates if the activation of neurons in CBL makes reasonable contributions to the final predictions. We first define the contribution of a neuron in CBL. For any text sample x from dataset $\mathcal{D}$ with n classes, the contribution for neuron j to class i is denoted as $W_{ij}{A_N}^{+}(x)_j$, where W is the weight matrix from the final linear layer, ${A_N}^{+}(x)$ is the non-negative activations from CBL, $i \in \{1, ..., n\}$, $j \in \{1, ..., k\}$, n is the number of classes, and k is the number of neurons in CBL. This contribution score directly describes how each neuron in CBL influences the final predictions. For a text sample that is correctly classified, we extract the neurons in CBL with high contributions to the prediction. Subsequently, *Contribution Faithfulness* can be evaluated by manually inspecting whether the concepts represented by the highly contributed neurons are related to the given text sample and are reasonable for making the correct prediction.

Human evaluation design. We perform the human evaluation through Amazon Mechanical Turk (MTurk) to study the faithfulness of our CB-LLM. Based on the above metrics, we design two tasks for human evaluation:

1. **Task 1: Activation Faithfulness.** In this task, workers will be presented with a neuron concept alongside the corresponding top k highly activated text samples. Workers need to provide a rating ranging from 1 (strongly disagree) to 5 (strongly agree) based on the agreement observed between the neuron concept and the top k highly activated samples.

2. **Task 2: Contribution Faithfulness.** In this task, workers will be presented with explanations from two models for a text sample. The explanations are generated by showing the top r neuron concepts with the highest contribution to the prediction. Workers need to compare which model's explanations are better.

We conduct human evaluations for Task 1 and Task 2 to compare our CB-LLMs with the *Random baseline*. The random baseline is generated by the following rules: For Task 1, the highly activated text samples are randomly selected. For Task 2, the explanations are randomly selected from the same concept set. To ensure more reliable results, each question in the tasks mentioned above is evaluated three times by different workers.

Table 3. The human evaluation results for task 1 — Activation Faithfulness. Workers give higher ratings to our CB-LLM w/ ACC, suggesting that the neurons in our CB-LLM w/ ACC are more interpretable than the neurons with random activations.

Task 1 — Activation Faithfulness	Dataset				Average
Method	SST2	YelpP	AGnews	DBpedia	
Human evaluation:					
CB-LLM w/ ACC (Ours)	**4.07**	**4.00**	**4.00**	**4.07**	**4.03**
Random (Baseline)	3.40	3.53	2.86	2.80	3.15
Large-scale human evaluation:					
CB-LLM w/ ACC (Ours)	**3.50**	**4.03**	**4.23**	**3.90**	**3.92**
Random (Baseline)	3.03	3.20	2.97	3.13	3.08

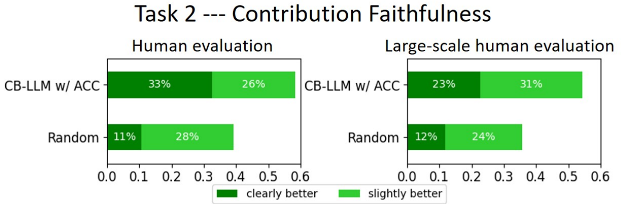


Figure 3. The human evaluation results for task 2 — Contribution Faithfulness. Workers prefer the explanations generated by CB-LLM w/ ACC more than the random explanations.

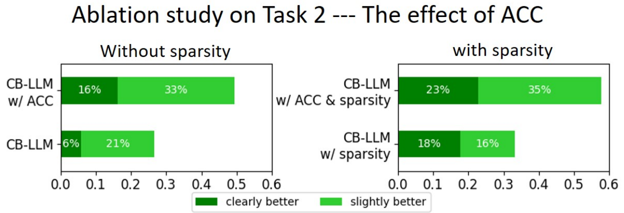


Figure 4. Ablation study on Automatic Concept Correction (ACC). Workers favor the explanations provided by the CB-LLMs with ACC.

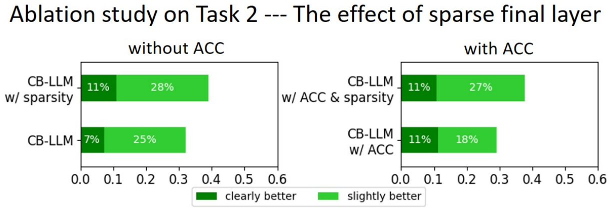


Figure 5. Ablation study on the sparsity. Workers demonstrate only a marginal preference for explanations provided by the CB-LLMs with a sparse final layer.

We also perform an additional large-scale human evaluation to further verify the human evaluation results for CB-LLMs. More details about the survey design and interface can be found in Appendix A.1.

5.3.1. Results of human evaluation

The results of task 1 (Activation Faithfulness) are shown in Table 3. Our CB-LLMs w/ ACC constantly achieve higher ratings than the random baseline. This suggests that the neurons in our CB-LLMs w/ ACC are more interpretable than the neurons with random activations.

The results of task 2 (Contribution Faithfulness) are shown in Figure 3. Workers consistently express a preference for our CB-LLM w/ ACC over the random baseline. This suggests that the explanations generated by our CB-LLM w/ ACC are better than randomly selected explanations.

5.3.2. Ablation study

We conduct an ablation study to evaluate the impact of ACC and sparse final layer on faithfulness through task 2. Figure 4 compares the CB-LLMs with ACC and the ones without ACC. Workers consistently prefer the explanations from CB-LLM with ACC, irrespective of using a sparse final layer or not. This indicates that our ACC strategy not only improves accuracy but also enhances the quality of explanations. Figure 5 compares the CB-LLMs w/ sparse final layer and the ones without. Workers exhibit little preference after the application of a sparse final layer, suggesting that sparsity might offer marginal help for the interpretability of CB-LLM. More details about the ablation study can be found in Appendix A.2.

6. Case study: Concept Unlearning

In this section, we provide use cases to demonstrate how to leverage the interpretability of our CB-LLM in practice.

Concept Unlearning refers to forcing the model to forget a certain concept. In some situations, there might be specific reasons to deactivate the influence of a particular concept on the final prediction. With the interpretable structure of our CB-LLM, we can easily unlearn a concept by manually deactivating a specific neuron in the CBL or removing all the weights connected to this neuron in the final linear layer.

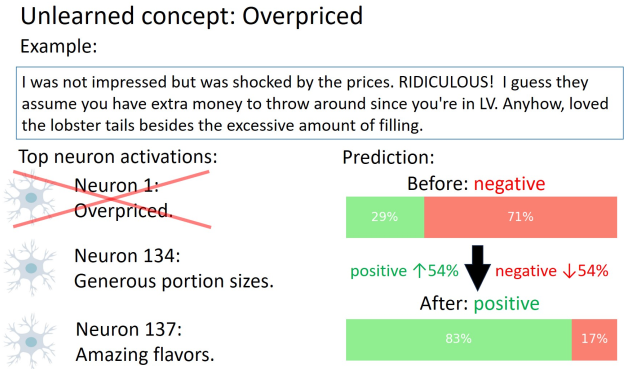
Figure 6 presents an example of unlearning the concept of "overpriced". In practice, we might consider removing the concept of "overpriced" from Yelp reviews due to subjectivity or geographical reasons (as the standard of overpricing varies across individuals and locations). This adjustment can encourage CB-LLM to prioritize the evaluation of product quality. After unlearning the concept of "overpriced," the predictions for 2726 samples in the test set changed from negative to positive. Subsequently, we employed **`bart-large-mnli`**, an NLI model, to assess whether these 2726 samples indeed contain the concept of "overpriced". Our findings reveal that 2162 out of the 2726 samples strongly entail "overpriced", accounting for 79%. This suggests that most of the samples now predicting positive were initially classified as negative due to the presence of the "overpriced" concept.

Table 4. The two neurons from CB-LLM w/ ACC and their corresponding highly activated samples.

Neuron	Highly activated samples
(AGnews) Neuron #16: human rights violations and advocacy.	1. US soldier convicted of torture in Iraq A US military intelligence soldier in Iraq has been sentenced to 8 months in prison for taking part in torturing detainees in Abu Ghraib prison. 2. Pinochet is ordered to stand trial for murder Augusto Pinochet, the former Chilean dictator, was ordered under house arrest yesterday, charged with kidnapping and murder dating back to his 17-year rule. 3. Trial Date Set for Soldier at Abu Ghraib (AP) AP - A military judge ordered a U.S. Army reservist on Friday to stand trial Jan. 7 in Baghdad for allegedly abusing Iraq inmates at the Abu Ghraib prison outside Baghdad. 4. Afghan court convicts US trio of torture KABUL, Afghanistan – Three Americans – led by a former Green Beret who boasted he had Pentagon support – were found guilty yesterday of torturing Afghans in a private jail and were sentenced to prison. 5. Soldier to Plead Guilty in Iraq Abuse Case (AP) AP - An Army reservist charged with abusing Iraqi prisoners plans to plead guilty at a court martial to four counts arising from the Abu Ghraib prison abuse scandal in a plea deal in which eight other counts will be dropped, his lawyer has said.
(DBpedia) Neuron #71: the artist's born date.	1. Joanna Taylor (born 24 July 1978) is an English actress and former model. 2. Jody Miller (born November 29 1941) is an American country music singer. Born as Myrna Joy Miller she was born in Phoenix Arizona and raised in Oklahoma. 3. Priscilla Mitchell (born September 18 1941 in Marietta Georgia) was an American country music singer. 4. Geoffrey Davies (born 15 December 1942 Leeds West Riding of Yorkshire) is a British actor. 5. He was born in Asunción Paraguay on March 27 1950. Son of Carmen Emategui and Rodolfo Barreto.

Table 5. The explanations generated by CB-LLM w/ ACC for two text samples.

Sample	Explanations
(SST2) Sample #330: occasionally funny , always very colorful and enjoyably overblown in the traditional almodóvar style .	1. Charming characters. 2. Clever and unexpected humor. 3. Stunning and exotic locations. 4. Stellar and diverse ensemble cast. 5. Unique and well-developed characters.
(YelpP) Sample #34857: This place has something for everyone. My wife and I started going there out of convenience before attending a movie at the South Pointe. But then we continued going back because we liked the food and the staff is very helpful. This most recent visit I had sushi for the first time and it was very good - and reasonably priced. We have company coming and are going to make it one of our stops on their visit.	1. Welcoming and friendly staff. 2. Clean and inviting ambiance. 3. Amazing flavors. 4. Great warranty and support. 5. Delicious food.


Unlearned concept: Overpriced
Example:
Top neuron activations:
~~Neuron 1: Overpriced.~~
Neuron 134: Generous portion sizes.
Neuron 137: Amazing flavors.
Prediction:
Before: negative
29%
71%
positive ↑54%
negative ↓54%
After: positive
83%
17%

I was not impressed but was shocked by the prices. RIDICULOUS! I guess they assume you have extra money to throw around since you're in LV. Anyhow, loved the lobster tails besides the excessive amount of filling.

Figure 6. An example of concept unlearning. This example is initially classified as negative due to the customer complaining about the high price, despite the lobster tails being great. After unlearning the concept "Overpriced", the concepts "Amazing flavors" and "Generous portion sizes" dominate the prediction, resulting in a positive prediction.

Based on the above case study, we believe our CB-LLM has great potential to facilitate human intervention such as Concept Unlearning for enhancing fairness, as users can easily remove biased, subjective, or unfair elements that could distort the predictions. More examples of Concept Unlearning can be found in Appendix A.3.

7. Visulization of neurons and explanations

We provide some visualizations of neurons in our CB-LLM w/ ACC in Table 4. The neurons are displayed along with their corresponding concepts and the top 5 highly activated samples in the dataset. For a given text sample, we also show the 5 explanations generated by CB-LLM w/ ACC in Table 5. The explanations are generated by selecting the top 5 concepts with the highest contribution to the prediction. More visualizations of neurons and explanations can be found in Appendix A.4 and A.5 respectively.

8. Conclusion

In this work, we introduced CB-LLM, the first CBM that scales to large text classification benchmarks. Our CB-LLM is fully automatic, training-efficient, and achieves accuracy comparable to, and sometimes surpassing, black-box language models while providing faithful interpretability.

Broader impact

CB-LLM represents a notable advancement in the realm of interpretable language models. As demonstrated in Section 6, CB-LLM allows human intervention to identify and remove biased, subjective, or unfair elements that can distort the predictions. We believe that this feature could positively contribute to fairness and transparency in the development of Large Language Models.

Acknowledgements

This work is supported in part by National Science Foundation (NSF) awards CNS-1730158, ACI-1540112, ACI-1541349, OAC-1826967, OAC-2112167, CNS-2100237, CNS-2120019, the University of California Office of the President, and the University of California San Diego's California Institute for Telecommunications and Information Technology/Qualcomm Institute. Thanks to CENIC for the 100Gbps networks. C. Sun, T. Oikarinen and T.-W. Weng are supported by National Science Foundation under Grant No. 2107189 and 2313105. T.-W. Weng also thanks the Hellman Fellowship for providing research support.

References

Antverg, O. and Belinkov, Y. On the pitfalls of analyzing individual neurons in language models. In *ICLR*, 2022.

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., and Saunders, W. Language models can explain neurons in language models, 2023.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In *NeurIPS*, 2020.

Dalvi, F., Durrani, N., Sajjad, H., Belinkov, Y., Bau, A., and Glass, J. R. What is one grain of sand in the desert? analyzing individual neurons in deep NLP models. In *AAAI*, 2019.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

Gao, T., Yao, X., and Chen, D. SimCSE: Simple contrastive learning of sentence embeddings. In *EMNLP*, 2021.

Kádár, Á., Chrupala, G., and Alishahi, A. Representation of linguistic form and function in recurrent neural networks. *Comput. Linguistics*, 2017.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C. J., Wexler, J., Viégas, F. B., and Sayres, R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*, 2018.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In *ICML*, 2020.

Kroeger, N., Ley, D., Krishna, S., Agarwal, C., and Lakkaraju, H. Are large language models post hoc explainers? *CoRR*, 2023.

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., and Baroni, M. The emergence of number and syntax units in LSTM language models. In *NAACL-HLT*, 2019.

Lee, J., Oikarinen, T., Chatha, A., Chang, K.-C., Chen, Y., and Weng, T.-W. The importance of prompt tuning for automated neuron explanations. In *NeurIPS ATTRIB Workshop*, 2023.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. Dbpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 2015.

Li, J., Chen, X., Hovy, E. H., and Jurafsky, D. Visualizing and understanding neural models in NLP. In *NAACL*, 2016.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, 2019.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013.

Oikarinen, T. P. and Weng, T. Clip-dissect: Automatic description of neuron representations in deep vision networks. In *ICLR*, 2023.

Oikarinen, T. P., Das, S., Nguyen, L. M., and Weng, T. Label-free concept bottleneck models. In *ICLR*, 2023.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., and Lowe, R. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*, 2013.

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T. Mpnet: Masked and permuted pre-training for language understanding. In *NeurIPS*, 2020.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and Brew, J. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, 2019.

Yüksekgönül, M., Wang, M., and Zou, J. Post-hoc concept bottleneck models. In *ICLR*, 2023.

Zhang, X., Zhao, J. J., and LeCun, Y. Character-level convolutional networks for text classification. In *NeurIPS*, 2015.

A. Appendix

A.1. MTurk survey design and interface

We perform the human evaluation through Amazon Mechanical Turk (MTurk). The details of two tasks we design are as follows:

1. **Task 1 — Activation Faithfulness:** In this task, workers will be presented with a neuron concept alongside the corresponding top 5 highly activated text samples. Workers need to provide a rating ranging from 1 (strongly disagree) to 5 (strongly agree) based on the agreement observed between the neuron concept and the top 5 highly activated samples.

2. **Task 2 — Contribution Faithfulness.** In this task, workers will be presented with explanations from two models for a text sample. The explanations are generated by showing the top 5 neuron concepts with the highest contribution to the prediction. Workers need to compare which model's explanations are better and select an option from "model 1 is clearly better", "model 1 is slightly better", "equally good", "model 2 is slightly better", and "model 2 is clearly better".

We did human evaluations on MTurk for Task 1 and Task 2 as mentioned in Section 5.3. The details are as follows:

- **Human evaluation:** We evaluate the following 5 models:
  - CB-LLM (Vanilla)
  - CB-LLM w/ ACC
  - CB-LLM w/ sparse final layer
  - CB-LLM w/ ACC & sparse final layer
  - *Random baseline*: For Task 1, the highly activated text samples are randomly selected. For Task 2, the explanations are randomly selected from the same concept set.

  For task 1, we evaluate each model's 5 most highly activated neuron concepts across each dataset. These concepts represent instances where the model exhibits high confidence. For task 2, we evaluate 5 random samples for every dataset.

We also did a large-scale human evaluation to verify the performance of our CB-LLM w/ ACC. The details are as follows:

- **Large-scale human evaluation:** We evaluate the following 2 models:
  - CB-LLM w/ ACC
  - *Random baseline*: For Task 1, the highly activated text samples are randomly selected. For Task 2, the explanations are randomly selected from the same concept set.

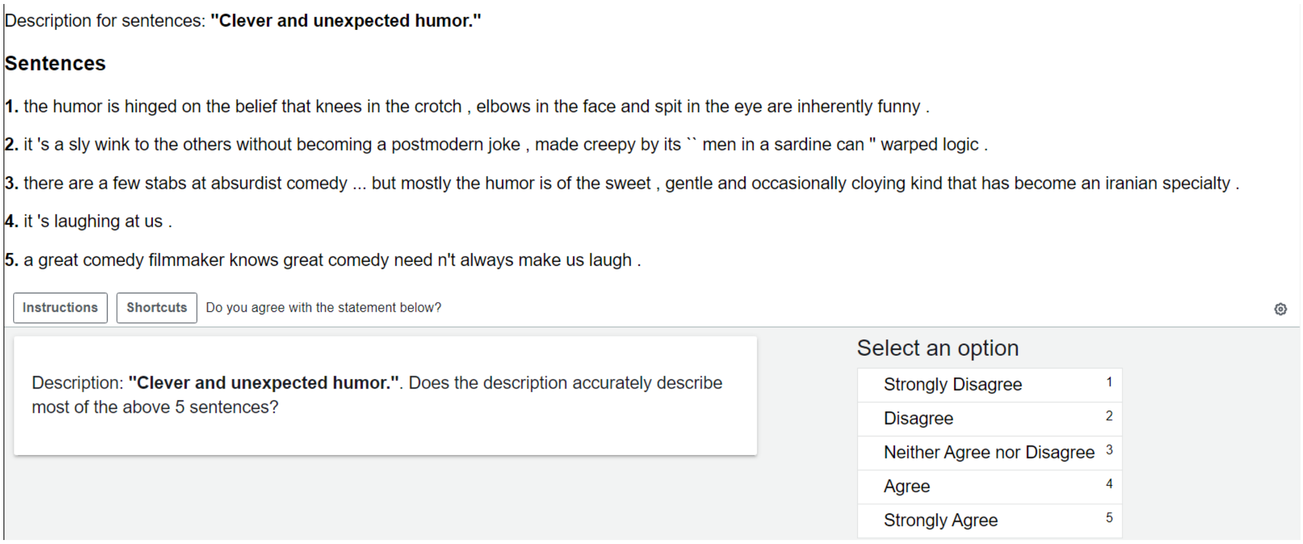
  For Task 1, we evaluate the 10 most highly activated neuron concepts for each model across each dataset. For task 2, we evaluate 40 random samples (20 per class) from SST2, 40 random samples (20 per class) from YelpP, 40 random samples (10 per class) from AGnews, and 70 random samples (5 per class) from DBpedia.

To ensure more reliable results, each question in the tasks mentioned above is evaluated three times by different workers.

The survey interface for task 1 and task 2 is shown in Figure 7 and Figure 8 respectively. In task 2, workers are also asked to provide ratings for each model, similar to task 1. These ratings are utilized to filter out inconsistent results. The following logic is employed for filtering:

- If workers indicate that model 1 is slightly or clearly better than model 2, the rating of model 1 must be no lower than the rating of model 2, and vice versa.

- If workers select "equally good," the two models must have the same rating.



Description for sentences: **"Clever and unexpected humor."**

Sentences

1. the humor is hinged on the belief that knees in the crotch , elbows in the face and spit in the eye are inherently funny .

2. it 's a sly wink to the others without becoming a postmodern joke , made creepy by its `` men in a sardine can " warped logic .

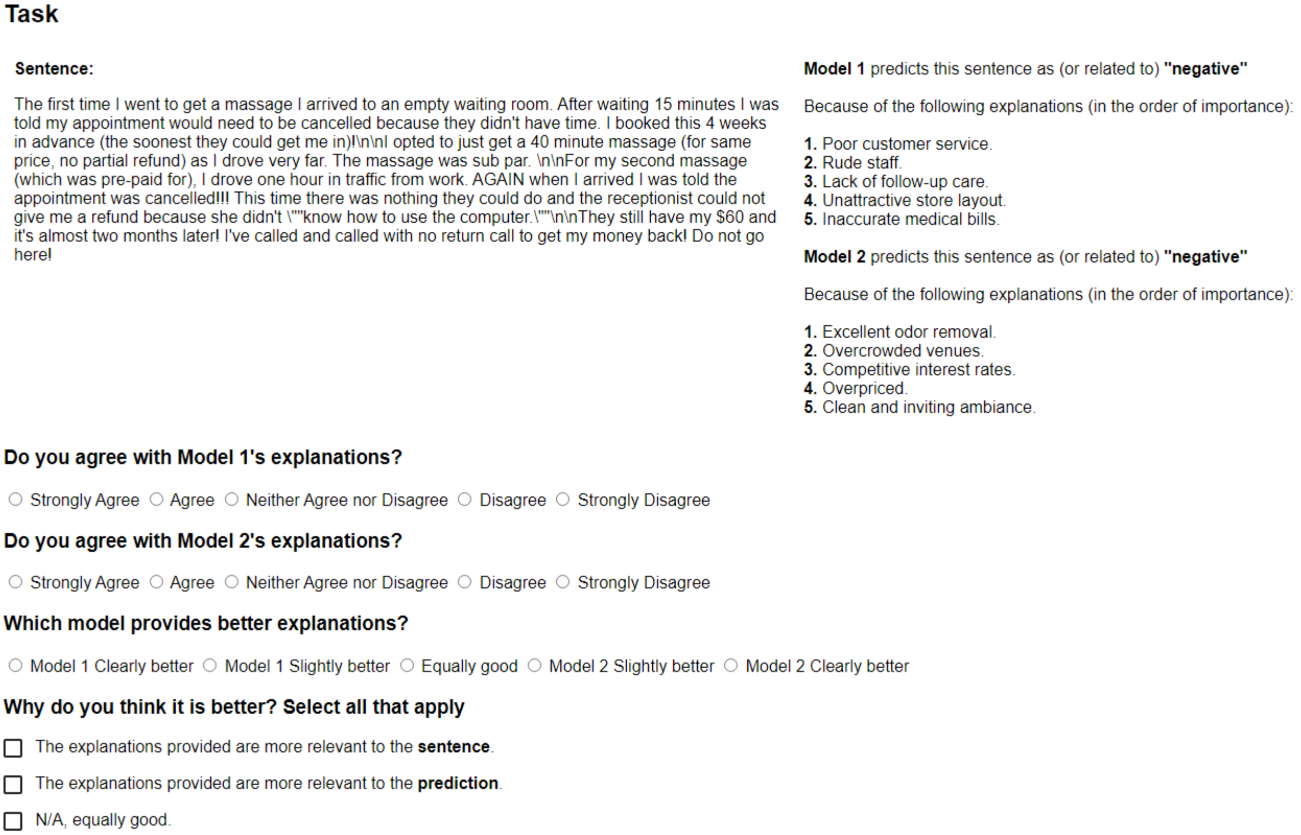
3. there are a few stabs at absurdist comedy ... but mostly the humor is of the sweet , gentle and occasionally cloying kind that has become an iranian specialty .

4. it 's laughing at us .

5. a great comedy filmmaker knows great comedy need n't always make us laugh .

Description: **"Clever and unexpected humor."**. Does the description accurately describe most of the above 5 sentences?

Figure 7. The interface for task 1 — Activation faithfulness.



Task

Sentence:

The first time I went to get a massage I arrived to an empty waiting room. After waiting 15 minutes I was told my appointment would need to be cancelled because they didn't have time. I booked this 4 weeks in advance (the soonest they could get me in)!\n\nI opted to just get a 40 minute massage (for same price, no partial refund) as I drove very far. The massage was sub par. \n\nFor my second massage (which was pre-paid for), I drove one hour in traffic from work. AGAIN when I arrived I was told the appointment was cancelled!!! This time there was nothing they could do and the receptionist could not give me a refund because she didn't \""know how to use the computer.\""\n\nThey still have my $60 and it's almost two months later! I've called and called with no return call to get my money back! Do not go here!

Model 1 predicts this sentence as (or related to) **"negative"**

Because of the following explanations (in the order of importance):

1. Poor customer service.
2. Rude staff.
3. Lack of follow-up care.
4. Unattractive store layout.
5. Inaccurate medical bills.

Model 2 predicts this sentence as (or related to) **"negative"**

Because of the following explanations (in the order of importance):

1. Excellent odor removal.
2. Overcrowded venues.
3. Competitive interest rates.
4. Overpriced.
5. Clean and inviting ambiance.

Do you agree with Model 1's explanations?

○ Strongly Agree ○ Agree ○ Neither Agree nor Disagree ○ Disagree ○ Strongly Disagree

Do you agree with Model 2's explanations?

○ Strongly Agree ○ Agree ○ Neither Agree nor Disagree ○ Disagree ○ Strongly Disagree

Which model provides better explanations?

○ Model 1 Clearly better ○ Model 1 Slightly better ○ Equally good ○ Model 2 Slightly better ○ Model 2 Clearly better

Why do you think it is better? Select all that apply

□ The explanations provided are more relevant to the **sentence**.

□ The explanations provided are more relevant to the **prediction**.

□ N/A, equally good.

Figure 8. The interface for task 2 — Contribution faithfulness.

A.2. More details for ablation study

We also compare the Vanilla CB-LLM without ACC or a sparse final layer with the random baseline. Vanilla CB-LLM still provides more favorable explanations compared to the random explanations.

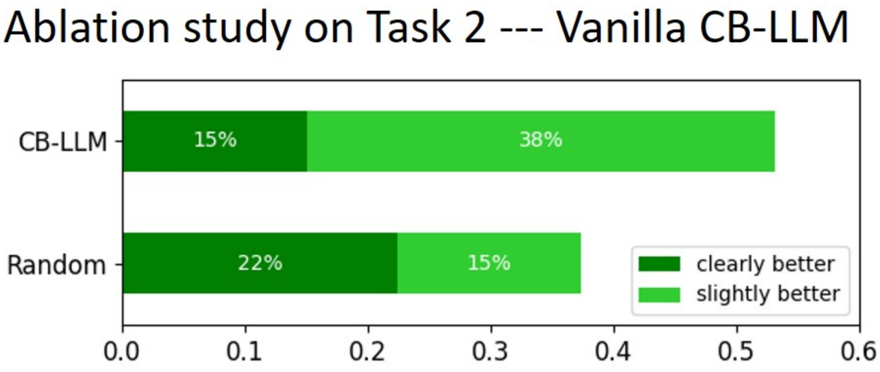
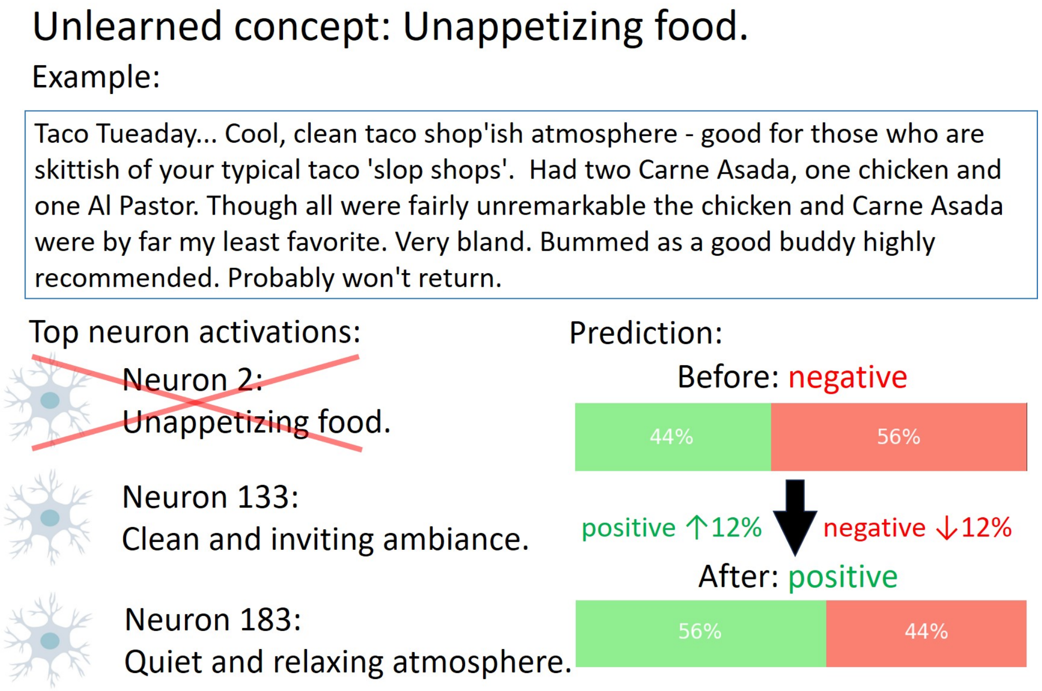


Figure 9. The evaluation of vanilla CB-LLM on task 2. Workers still prefer the explanations from Vanilla CB-LLM more than the explanations from random baseline.

A.3. More examples for Concept Unlearning

Figure 10 demonstrates another example of Concept Unlearning. The concept "Unappetizing food" is unlearned. After unlearning, the predictions of 370 samples changed from negative to positive, with 313 of them (85%) strongly entailing "Unappetizing food". This suggests that most of the samples now predicting positive were initially classified as negative due to the presence of the "Unappetizing food" concept.


Unlearned concept: Unappetizing food.

Example:

Taco Tueaday... Cool, clean taco shop'ish atmosphere - good for those who are skittish of your typical taco 'slop shops'. Had two Carne Asada, one chicken and one Al Pastor. Though all were fairly unremarkable the chicken and Carne Asada were by far my least favorite. Very bland. Bummed as a good buddy highly recommended. Probably won't return.

Top neuron activations:
~~Neuron 2: Unappetizing food.~~
Neuron 133: Clean and inviting ambiance.
Neuron 183: Quiet and relaxing atmosphere.
Prediction:
Before: negative
44%
56%
positive ↑12%
negative ↓12%
After: positive
56%
44%

Figure 10. Another example of concept unlearning. This example is initially classified as negative due to the customer complaining about the bland food, despite the cool and clean atmosphere. After unlearning the concept "Unappetizing food" the concepts "Clean and inviting ambiance" and "quiet and relaxing atmosphere" dominate the prediction, resulting in a positive prediction.

A.4. Visualization of neurons in CB-LLM

In this section, we provide more visualizations of the neurons in our CB-LLM. We select 3 neurons that have the highest activations across samples for each dataset.

Table 6: The neurons of CB-LLM w/ ACC and corresponding highly activated samples for each dataset. We show the top 3 neurons with the largest activations for each dataset.

Dataset	Neuron	Highly activated samples
SST2	Neuron 184: Clever and unexpected humor.	1. the humor is hinged on the belief that knees in the crotch , elbows in the face and spit in the eye are inherently funny . 2. it ’s a sly wink to the others without becoming a postmodern joke , made creepy by its “ men in a sardine can ” warped logic . 3. there are a few stabs at absurdist comedy ... but mostly the humor is of the sweet , gentle and occasionally cloying kind that has become an iranian specialty . 4. it ’s laughing at us . 5. a great comedy filmmaker knows great comedy need n’t always make us laugh .
SST2	Neuron 170: Great chemistry between actors.	1. when your leading ladies are a couple of screen-eating dominatrixes like goldie hawn and susan sarandon at their raunchy best , even hokum goes down easily . 2. binoche and magimel are perfect in these roles . 3. hugh grant and sandra bullock are two such likeable actors . 4. interacting eyeball-to-eyeball and toe-to-toe , hopkins and norton are a winning combination – but fiennes steals ‘ red dragon ’ right from under their noses . 5. without resorting to hyperbole , i can state that kissing jessica stein may be the best same-sex romance i have seen .
SST2	Neuron 34: Lack of humor or wit.	1. frenetic but not really funny . 2. but here ’s the real damn : it is n’t funny , either . 3. francophiles will snicker knowingly and you ’ll want to slap them . 4. beyond a handful of mildly amusing lines ... there just is n’t much to laugh at . 5. do not , under any circumstances , consider taking a child younger than middle school age to this wallow in crude humor ."

YelpP	Neuron 184: Good breakfast options.	1. I'm obsessed with the breakfast here. There's a huge smorgasbord of options to choose from on the brekkie menu, and the hardest part is actually picking something to order because they all sound so good! I couldn't resist ordering the eggs benedicto. What a cute twist on your typical eggs benedict dish! The eggs were perfectly poached on toasty slabs of english muffin and accented with the rich and savory sundried tomato hollandaise. The bits of candied prosciutto added a nice meatiness to the benedict without making it too heavy. And while I don't normally reach for mixed greens for breakfast.... I did like it in this dish because my usual gripe with eggs benedict is that there's just wayyy too much going on. But the greens were a light alternative that kinda balanced everything out in a way that potatoes don't do it for me. I also picked up the horchata latte. I'm a huge fan of horchata (which is pretty hard to find in Hawaii where I'm from) and a coffee lover, so this was a must try for me! It's totally sweet, creamy, and probably chock full of calories, but worth every single tasty sip. If you're not feeling in a benedicto mood, that's OK because there's a ton of other food options to choose from. All of which resemble your standard breakfast fare, with a little bit of a twist. Mexican, southern, classic american breakfasts... You name it. If I had more stomach room and a little more time in Madison, I'd wanna try a little bit of every dish on the menu. One of each, please! 2. Half order of Mashed Potatoes Omelet and an ice tea is how everyone should start their day! 3. Quite delicious for brunch. I am not normally a sweet breakfast food person, however the buckwheat waffle with a mimosa seems to be a perfect combination. 4. The breakfast took a long time but when it finally did it was good! But a little pricey for eggs and bacon! 5. Great breakfast.

YelpP	Neuron 159: Engaging performances.	1. I saw LOVE yesterday, my first Las Vegas show. It was mind-bogglingly fantastic. I was totally swept away and mesmerized for over two hours. The sheer creativity, imagination, music, engineering, intricate choreography left me in a state of deep admiration for the entire effort. It was superb beyond words. See it before you die. 2. If you're a huge Beatles fan, you will love this show. If you're a huge Cirque du Soleil fan, you might feel a lil' bit disappointed? But I guarantee this, you will definitely appreciate the artistic value of the show and what it's goal was..and that was to pay homage to one of the most influential bands in the history of music. Since I have been a life long Beatles fan, I was very curious as what I should expect? And then my wife literally said, """Let It Be!""" , and I did...I just relaxed and let go of any expectations from any other show that had seen in the past. Once the music began, I was knocked into the back of my seat. The audio and visual presentation is awesome. In addition to the theaters dynamic audio system, you have high end audio speakers that are embedded in the headrest of your very comfortable theater chairs, plus the seats that are in front of you have the speakers directed towards you as well! The main body of the show starts off very solemn, and as the crescendo builds....it EXPLODES on to the scene with a re-mastered version of """Get Back""". With performers dancing and running around the middle of the stage, skaters skating, and acrobats literally falling and flying from the sky.....Whew! 3. this show was great!! if you love fire and acrobatic stuff you will love this show!! its good for families as well. this was the 3rd cirque du soleii show they never dissapoint me. the set was awesome and costumes! 4. Great show! Great acts! Wally Eastwood was awesome and funny. Had 2 finalists from America got talent show. I would see it again - great for all ages! Acrobatics were cool. The magic show was ok but still good to see as 1 of the many acts. Wally Eastwood is on YouTube. I highly recommend this show. The only missing star is for people expecting great props and scenery but for the great show, you wouldn't care. 5. what a really fun show! It was really well paced and had a great selection of Beatles music. The story line runs you through the decades. The use of multi-media is really great and any seat in the house would be an incredible show. it's a circular stage so there is stuff going on everywhere - it's hard to know where to look!! I thought the Cirque stuff was a little less insane than some of their other shows. don't get me wrong - stunning and fun to watch but it didn't seem as over the top/awe-inspiring as some of the others I have seen when it comes to the athleticism and """never before seen""" type stuff. But the show was packed with a great story, amazing costumes, graphics, dancing, etc., and I loved every minute of it!

YelpP	Neuron 104: Unattractive store layout.	1. I totally agree with Tina S. for such a large and beautiful store to be quite honest......the selection in a word.....SUCKS. The only reason I didn't give this store one star was because it is a very spacious store....but I think they waste a lot of space......and the customer service was excellent. However when you go into a Nike Store of any kind.....exception being the outlets.......there should be more than 7 or 8 NFL team Jerseys and T-Shirts in the place. I was extremely disappointed with that....and for that fact that is why I have never been a huge fan of Nike products or stores. Eat, Drink, and be Merry my Friends!!!!! 2. This mall- eh It's not horrible, but it's a waste of time. I visited from out of town and it was not worth my while. The stores were your typical """upscale""" shops, but good luck finding anything with the pacs of shoppers looking to score """deals""". The only stores worth going to are Gap outlet and J Crew factory. I was excited when I saw H&M but don't be fooled, it's not an outlet store so no """special""" deals there. Avoid the crowds, save the gas $ and go elsewhere. Pros: - I got 2 dresses at Gap outlet for less than $20 Cons: - Crowded - Lack of selection - Not all stores are outlets even though this is an outlet mall - No food courts and when you put your credit card in the vending machine good luck getting your drink 3. I made a few trips to this mall during our week in the Phoenix area. The Nike Outlet was great, but otherwise, there weren't that many quality outlet stores. Most (or so it seemed to me) of the stores in this mall are not outlets and there just weren't the deals that I was expecting. 4. I hate to say it, but this mall is kind of ghetto. The layout is somewhat bizarre, and depending on which side you enter, you'd never know about the other side if you didn't look at a map and just decide to wander. The stores are really nothing special and if you seek high end stores, you're better off hitting the strip. What's really weird is the women's stores in there–they're either plus sizes, clubby looking stuff, or outright hooker uniforms. There is also a Macy's, JC Penny and Sears. Three stores I never buy anything from anyway. There is, however, a Cinnabon, and I LOVE Cinnabon.... 5. This mall is sad. You will actually feel bad for this mall. Only a couple shops are open and they are either shoe stores, clothing or cell phones. The food court doesn't make any sense and not very inviting. Also there wasn't a mrs. Fields cuz I was craving cookies. Lol Your better off going to the flea market for better stuff and cheaper prices!

AGnews	Neuron 20: sports events and achievements.	1. Maddux Wins No. 302, Baker Wins No. 1,000 Greg Maddux pitched the Chicago Cubs into the lead in the NL wild-card race and gave Dusty Baker a win to remember. Maddux threw seven shutout innings for his 302nd career win, Baker got his 1,000th victory as a manager and Chicago beat the Montreal Expos 5-2 on Monday night... 2. Colts Lead Pats Early in Third Quarter FOXBORO, Mass. - Peyton Manning reached the 25,000-yard passing mark faster than anyone but Dan Marino, and the Indianapolis Colts shredded the New England Patriots for a 17-13 halftime lead Thursday night... 3. Davenport Advances at U.S. Open NEW YORK - Lindsay Davenport's summer of success stayed on course Thursday when the fifth-seeded former U.S. Open champion defeated Arantxa Parra Santonja 6-4, 6-2 and advanced to the third round of the season's final Grand Slam event... 4. U.S. Men's Hoops Team Finally Gets a Rout ATHENS, Greece - The Americans got a taste of what it was like in the good ol' days. They finally played an opponent they were able to beat easily, routing Angola 89-53 Monday in their final preliminary game of the Olympic men's basketball tournament... 5. U.S. Softball Team Wins, Closes in on Gold ATHENS, Greece - Right now, the Americans aren't just a Dream Team - they're more like the Perfect Team. Lisa Fernandez pitched a three-hitter Sunday and Crystl Bustos drove in two runs as the Americans rolled to their eighth shutout in eight days, 5-0 over Australia, putting them into the gold medal game...

AGnews	Neuron 16: human rights violations and advocacy.	1. US soldier convicted of torture in Iraq A US military intelligence soldier in Iraq has been sentenced to 8 months in prison for taking part in torturing detainees in Abu Ghraib prison. 2. Pinochet is ordered to stand trial for murder Augusto Pinochet, the former Chilean dictator, was ordered under house arrest yesterday, charged with kidnapping and murder dating back to his 17-year rule. 3. Trial Date Set for Soldier at Abu Ghraib (AP) AP - A military judge ordered a U.S. Army reservist on Friday to stand trial Jan. 7 in Baghdad for allegedly abusing Iraq inmates at the Abu Ghraib prison outside Baghdad. 4. Afghan court convicts US trio of torture KABUL, Afghanistan – Three Americans – led by a former Green Beret who boasted he had Pentagon support – were found guilty yesterday of torturing Afghans in a private jail and were sentenced to prison. 5. Soldier to Plead Guilty in Iraq Abuse Case (AP) AP - An Army reservist charged with abusing Iraqi prisoners plans to plead guilty at a court martial to four counts arising from the Abu Ghraib prison abuse scandal in a plea deal in which eight other counts will be dropped, his lawyer has said.
AGnews	Neuron 10: terrorism and security threats.	1. Pakistan's top wanted terrorist killed Pakistani security forces Sunday killed the country's most wanted terrorist allegedly involved in an assassination attempt on President Pervez Musharrafand indicted in the murder of a US journalist. 2. Al-Qaeda Group Kills a Second US Hostage in Iraq (Update3) An Iraqi group linked to al-Qaeda killed a second US hostage, Jack Hensley, and threatened to kill a British hostage unless Iraqi women detainees are freed, the group said on its Web site. 3. Pakistan arrests key Al-Qaeda operative (AFP) AFP - Pakistani security forces have arrested a key Al-Qaeda operative wanted in connection with attacks on Christian targets and a failed bid to kill President Pervez Musharraf, an official said. 4. Pakistan al-Qaeda suspect killed Pakistan says it has dealt a major blow to al-Qaeda's operations after its security forces shot dead the country's most wanted terror suspect. 5. Seven suspected terrorists arrested in Spain Spain's Interior Minister says police have broken up a radical Muslim cell, plotting to bomb the country's National Court."

DBpedia	Neuron 174: words related to ship, car, train.	1. USS England (DE-635) a Buckley-class destroyer escort of the United States Navy was named in honor of Ensign John C. England (1920–1941) who was killed in action aboard the battleship Oklahoma during the Japanese attack on Pearl Harbor on 7 December 1941. 2. HMS Siren (most often referred to as Syren in contemporary records) was a sixth-rate post ship of the British Royal Navy in commission between 1745 and 1763 seeing action during the War of the Austrian Succession and the Seven Years' War. 3. HMS Benbow was a Victorian era Admiral-class battleship of the British Royal Navy named for Admiral John Benbow. 4. HMS Rackham was one of 93 ships of the Ham-class of inshore minesweepers. Their names were all chosen from villages ending in -ham. The minesweeper was named after Rackham in West Sussex. 5. HMS Captain was a 74-gun third-rate ship of the line of the Royal Navy launched on 26 November 1787 at Limehouse. She served during the French revolutionary and Napoleonic Wars before being placed in harbour service in 1799. An accident caused her to burn and founder in 1813. Later that year she was raised and broken up.
DBpedia	Neuron 71: the artist's born date.	1. Joanna Taylor (born 24 July 1978) is an English actress and former model. 2. Jody Miller (born November 29 1941) is an American country music singer. Born as Myrna Joy Miller she was born in Phoenix Arizona and raised in Oklahoma. 3. Priscilla Mitchell (born September 18 1941 in Marietta Georgia) was an American country music singer. 4. Geoffrey Davies (born 15 December 1942 Leeds West Riding of Yorkshire) is a British actor. 5. He was born in Asunción Paraguay on March 27 1950. Son of Carmen Emategui and Rodolfo Barreto.

DBpedia	Neuron 469: the publisher and imprint of the work.	1. The Tameside Advertiser is a weekly newspaper which serves the Metropolitan Borough of Tameside Greater Manchester England. It is owned by Trinity Mirror plc. The paper has a sister paper The Glossop Advertiser which is also a freesheet but covers the bordering town of Glossop in Derbyshire. The main competitors to both papers are the Tameside Reporter and Glossop Chronicle which are both paid-for newspapers. 2. Independent Tribune is a newspaper and based in Concord North Carolina covering Cabarrus County North Carolina. The newspaper is owned by Berkshire Hathaway. The Independent Tribune was formed with the merger of The Concord Tribune and The (Kannapolis) Daily Independent.It was originally a daily newspaper but changed to 3 days a week in 2009. 3. The Livingston County Daily Press & Argus is a daily newspaper published in Howell Michigan and owned by Gannett. 'As its name implies it covers news and sports within Livingston County and had offices in both Howell and Brighton. The Brighton office closed in December 2008. Its printing facility is located in Howell Township. It publishes every day except Saturday. 4. The Anchorage Press is a free alternative weekly newspaper based in Anchorage Alaska and owned by Wick Communications.Established in 1992 by Bill Boulay Barry Bialik and Nick Coltman as the Anchorage Bypass it was renamed the Anchorage Press in 1994. It is published and distributed every Thursday with a circulation of approximately 25000. The paper was sold to Wick Communications Company in August 2006. 5. The Imperial Valley Press (originally known as the Imperial Press) is a daily newspaper published in El Centro California. It has been owned by Schurz Communications of South Bend Indiana since 1965.The Imperial Valley Press features local news from all communities of the Imperial Valley and the Mexicali Baja California area as well as San Diego County and portions of southwestern Arizona. The newspaper focuses on local news sports and opinion pieces.

A.5. explanations from CB-LLM

In this section, we provide more explanations generated by our CB-LLM. We randomly select 3 samples and show the top 5 explanations for each dataset.

Table 7: The explanations generated by CB-LLM w/ ACC for a given text sample. We show 3 random samples for each dataset.

Dataset	Sample	Explanations
SST2	Sample 260: a very witty take on change , risk and romance , and the film uses humour to make its points about ACC eptance and growth .	1. Clever and unexpected humor. 2. Charming characters. 3. Stellar and diverse ensemble cast. 4. Unique and well-developed characters. 5. Captivating and layered character backstories.
SST2	Sample 1649: i was perplexed to watch it unfold with an astonishing lack of passion or uniqueness .	1. Lack of tension-building scenes. 2. Unexplained or unresolved mysteries. 3. Uninspiring character deaths. 4. Poorly executed voice-over narration. 5. Lack of authentic cultural representation.
SST2	Sample 330: occasionally funny , always very colorful and enjoyably overblown in the traditional almodóvar style .	1. Charming characters. 2. Clever and unexpected humor. 3. Stunning and exotic locations. 4. Stellar and diverse ensemble cast. 5. Unique and well-developed characters.
YelpP	Sample 21864: These guys are money grubbing. What WAS a $25 haircut just jumped up to a $32 haircut. It's just a haircut for God's sake! I'm going elsewhere.	1. Poor customer service. 2. Unattractive store layout. 3. Rude staff. 4. Hidden fees. 5. Overpriced.
YelpP	Sample 34857: This place has something for everyone. My wife and I started going there out of convenience before attending a movie at the South Pointe. But then we continued going back because we liked the food and the staff is very helpful. This most recent visit I had sushi for the first time and it was very good - and reasonably priced. We have company coming and are going to make it one of our stops on their visit.	1. Welcoming and friendly staff. 2. Clean and inviting ambiance. 3. Amazing flavors. 4. Great warranty and support. 5. Delicious food.

YelpP	Sample 10736: One of the few Cirque du Soleil that follow a story line, so if you are looking for a Cirque du Soleil show and a story this is the one to see. Although it strays a bit from the traditional style of Cirque du Soleil, it is still sure to please. We were fortunate enough to be able to purchase front section tickets for 50% off AMAZING deal! (End of summer special). KA is the show which it is the stage that is at the center of attention. It uses a sectional stage that is fully mobile it rotates and moves on a 3D axis it really adds another level of excitement to the show. I would not recommend this as anyone's first Cirque du Soleil show but for a any repeat or veteran Cirque du Soleil viewer this must make it onto your "Seen it" list.	1. Engaging performances. 2. Clean and inviting ambiance. 3. Interactive experiences. 4. Engaging podcasts. 5. Welcoming and friendly staff.
AGnews	Sample 3058: Mobile phone network reaches last of China's ethnic minorities (AFP) AFP - China has brought its mobile phone network to the last of its ethnic minority regions previously cut off from communication with the outside world, state media reported.	1. telecommunications and 5G technology. 2. tech giants and major industry players. 3. consumer electronics and gadgets. 4. emerging technologies and startups. 5. words related to technical devices.
AGnews	Sample 6124: Van Gogh's murder brings out Holland's contradictions The murder of Dutch filmmaker Theo van Gogh by a young Muslim of Moroccan descent has shaken Holland to its very foundations. To most people, including the Dutch, the killing and its violent	1. human rights violations and advocacy. 2. terrorism and security threats. 3. words related to war, conflict. 4. international aid and humanitarian efforts. 5. public health crises and pandemics.
AGnews	Sample 1035: Orioles 8, Devil Rays 0 Javy Lopez drove in four runs, Daniel Cabrera became the first rookie to win 10 games this season, and the Baltimore Orioles held the Tampa Bay Devil Rays to two hits in an 8-0 victory.	1. team rankings and standings. 2. fan reactions and opinions. 3. record-breaking performances. 4. athlete comebacks after injury. 5. name of sports stars.
DBpedia	Sample 52170: Narthecium is a genus of flowering plants. This genus was traditionally treated as belonging to the family Liliaceae but the APG II system of 2003 placed it in the family Nartheciaceae.The global distribution of the genus is widely disjunct - 1 species in Asia 1-5 species in Europe (see Narthecium ossifragum and 2 species in North America. Narthecium americanum is a candidate for listing under the federal Endangered Species Act in the United States.	1. The botanical classification of the plant. 2. the name of the plant. 3. The native habitat of the plant. 4. the genus or family of plant. 5. The plant's contribution to biodiversity.

DBpedia	Sample 32678: Pemberton's Headquarters also known as Willis-Cowan House is a two-story brick house that served as the headquarters for Confederate General John C. Pemberton during most of the 47 day siege of Vicksburg and the site where he decided to surrender the city to Union General Ulysses S. Grant on July 4 1863.During the 1960s the building housed a kindergarten associated with Vicksburg Catholic School (St.	1. the location of the building. 2. The historical significance of the building. 3. the name of the building. 4. the built date of the building. 5. The cultural or artistic significance of the building.
DBpedia	Sample 12750: Disma Fumagalli (born Inzago September 8 1826 - died Milan March 9 1893) was an Italian composer and teacher of music. He was a graduate of the Milan Conservatory where he began teaching piano in 1853. He composedmore than 300 études for piano as well as other exercises; he also wrote a concerto for piano and string orchestra. Fumagalli's brothers Carlo Polibio Adolfo and Luca were all composers.	1. the artist's born date 2. The artist's cultural significance. 3. The artist's famous collaborations. 4. The artist's notable achievements. 5. The artist's early influences.

A.6. Details of prompting ChatGPT

In this section, We provide the details of how we prompt ChatGPT to acquire the concept set. We use four human-designed concepts as examples for in-context learning. This prompting style requires only n queries to ChatGPT to obtain the full concept set and can be done efficiently through the web interface provided by OpenAI. The full prompts are shown in 8.

Table 8: The designed prompts for each dataset and class.

Dataset	Class	Prompt
SST2	negative	Here are some examples of key features that are often present in a negative movie rating. Each feature is shown between the tag <example></example>. <example>Flat or one-dimensional characters.</example> <example>Uninteresting cinematography.</example> <example>Lack of tension-building scenes.</example> <example>Lack of emotional impact.</example> List 100 other different important features that are often present in a negative movie rating. Need to follow the template above, i.e. <example>features</example>.
SST2	positive	Here are some examples of key features that are often present in a positive movie rating. Each feature is shown between the tag <example></example>. <example>Engaging plot.</example> <example>Strong character development.</example> <example>Great humor.</example> <example>Clever narrative structure.</example> List 100 other different important features that are often present in a positive movie rating. Need to follow the template above, i.e. <example>features</example>.
YelpP	negative	Here are some examples of key features that are often present in a negative Yelp review with lower star ratings (e.g., 1 or 2 stars). Each feature is shown between the tag <example></example>. <example>Overpriced.</example> <example>Unappetizing food.</example> <example>Unprofessional service.</example> <example>broken products.</example> The reviews fall into the following categories: Food, Automotive, Home Services, Entertainment, Medical, Hotels, Financial Services, Media, Parking, Clothing, Electronic devices, and Cleaning. List 100 other different important features that are often present in a negative Yelp review with lower star ratings (e.g., 1 or 2 stars). Need to follow the template above, i.e. <example>features</example>.
YelpP	positive	Here are some examples of key features that are often present in a positive Yelp review with higher star ratings (e.g., 4 or 5 stars). Each feature is shown between the tag <example></example>. <example>Delicious food.</example> <example>Outstanding service.</example> <example>Great value for the price.</example> <example>high quality products.</example> The reviews fall into the following categories: Food, Automotive, Home Services, Entertainment, Medical, Hotels, Financial Services, Media, Parking, Clothing, Electronic devices, and Cleaning. List 100 other different important features that are often present in a positive Yelp review with higher star ratings (e.g., 4 or 5 stars). Need to follow the template above, i.e. <example>features</example>.

AGnews	world	Here are some examples of key features that are often present in worldwide news. Each feature is shown between the tag <example></example>. <example>words related to country and place.</example> <example>political stunts taken by governments.</example> <example>global issues.</example> <example>words related to war, conflict.</example> List 50 other important features that are often present in worldwide news. Need to follow the template above, i.e. <example>features</example>.
AGnews	sports	Here are some examples of key features that are often present in sport news. Each feature is shown between the tag <example></example>. <example>name of sports stars.</example> <example>words related to game, competition.</example> <example>ball games like baseball, basketball.</example> <example>name of sport teams.</example> List 50 other important features that are often present in sport news. Need to follow the template above, i.e. <example>features</example>.
AGnews	business	Here are some examples of key features that are often present in business and financial news. Each feature is shown between the tag <example></example>. <example>words related to currency, money.</example> <example>the numerical amount of dollars.</example> <example>the symbol like $.</example> <example>words related to stock, Portfolio.</example> List 50 other important features that are often present in business and financial news. Need to follow the template above, i.e. <example>features</example>.
AGnews	science/ technology	Here are some examples of key features that are often present in news related to science and technology. Each feature is shown between the tag <example></example>. <example>name of scientists or the word scientists.</example> <example>words related to technical devices.</example> <example>words related to universe, space, planet.</example> <example>words related to the natural landscape.</example> List 50 other important features that are often present in news related to science and technology. Need to follow the template above, i.e. <example>features</example>.
DBpedia	company	Here are some examples of key features that are often present when introducing a company. Each feature is shown between the tag <example></example>. <example>the name of the company.</example> <example>the location of the company</example> <example>the founding year of the company</example> <example>words related to organization, group.</example> List 30 other important features that are often present when introducing a company. Need to follow the template above, i.e. <example>features</example>.

DBpedia	educational institution	Here are some examples of key features that are often present when introducing an educational institution. Each feature is shown between the tag <example></example>. <example>the name of the school.</example> <example>the location of the school</example> <example>the founding year of the school</example> <example>words related to college, university.</example> List 30 other important features that are often present when introducing an educational institution. Need to follow the template above, i.e. <example>features</example>.
DBpedia	artist	Here are some examples of key features that are often present when introducing an artist. Each feature is shown between the tag <example></example>. <example>the artist's name.</example> <example>the artist's works</example> <example>the artist's born date</example> <example>words related to music, painting.</example> List 30 other important features that are often present when introducing an artist. Need to follow the template above, i.e. <example>features</example>.
DBpedia	athlete	Here are some examples of key features that are often present when introducing an athlete or sports star. Each feature is shown between the tag <example></example>. <example>the athlete's or sports stars' name.</example> <example>the sport the athlete plays (e.g. football, basketball).</example> <example>the athlete's or sports stars' born date</example> <example>words related to ball games, competition.</example> List 30 other important features that are often present when introducing an athlete or sports star. Need to follow the template above, i.e. <example>features</example>.
DBpedia	office holder	Here are some examples of key features that are often present when introducing an office holder. Each feature is shown between the tag <example></example>. <example>the office holder's name.</example> <example>the office holder's position.</example> <example>the office holder's born date</example> <example>words related to politician, businessman.</example> List 30 other important features that are often present when introducing an office holder. Need to follow the template above, i.e. <example>features</example>.
DBpedia	transportation	Here are some examples of key features that are often present when introducing transportation. Each feature is shown between the tag <example></example>. <example>the model type of the transportation or vehicle.</example> <example>the production date of the transportation or vehicle.</example> <example>the functions of the transportation or vehicle.</example> <example>words related to ship, car, train.</example> List 30 other important features that are often present when introducing transportation. Need to follow the template above, i.e. <example>features</example>.

DBpedia	building	Here are some examples of key features that are often present when introducing a building. Each feature is shown between the tag <example></example>. <example>the name of the building.</example> <example>the built date of the building.</example> <example>the location of the building.</example> <example>words related to the type of the building (e.g. church, historic house, park, resort).</example> List 30 other important features that are often present when introducing a building. Need to follow the template above, i.e. <example>features</example>.
DBpedia	natural place	Here are some examples of key features that are often present when introducing a natural place. Each feature is shown between the tag <example></example>. <example>the name of the natural place.</example> <example>the length or height of the natural place.</example> <example>the location of the natural place.</example> <example>words related to mountain, river.</example> List 30 other important features that are often present when introducing a natural place. Need to follow the template above, i.e. <example>features</example>.
DBpedia	village	Here are some examples of key features that are often present when introducing a village. Each feature is shown between the tag <example></example>. <example>the name of the village.</example> <example>the population of the village.</example> <example>the census of the village.</example> <example>words related to district, families.</example> List 30 other important features that are often present when introducing a village. Need to follow the template above, i.e. <example>features</example>.
DBpedia	animal	Here are some examples of key features that are often present when introducing a kind of animal. Each feature is shown between the tag <example></example>. <example>the species of the animal.</example> <example>the habitat of the animal.</example> <example>the type of the animal (e.g. bird, insect, moth).</example> <example>words related to genus, family.</example> List 30 other important features that are often present when introducing a kind of animal. Need to follow the template above, i.e. <example>features</example>.
DBpedia	plant	Here are some examples of key features that are often present when introducing a kind of plant. Each feature is shown between the tag <example></example>. <example>the name of the plant.</example> <example>the genus or family of plant.</example> <example>the place where the plant was found.</example> <example>words related to grass, herb, flower.</example> List 30 other important features that are often present when introducing a kind of plant. Need to follow the template above, i.e. <example>features</example>.

DBpedia	album	Here are some examples of key features that are often present when introducing an album. Each feature is shown between the tag <example></example>. <example>the name of the album.</example> <example>the type of music, instrument.</example> <example>the release date of the album.</example> <example>words related to band, studio.</example> List 30 other important features that are often present when introducing an album. Need to follow the template above, i.e. <example>features</example>.
DBpedia	film	Here are some examples of key features that are often present when introducing a film. Each feature is shown between the tag <example></example>. <example>the name of the film.</example> <example>the maker or producer of the film.</example> <example>the type of the film (e.g. drama, science fiction, comedy, cartoon, animation).</example> <example>words related to TV, video.</example> List 30 other important features that are often present when introducing a film. Need to follow the template above, i.e. <example>features</example>.
DBpedia	written work	Here are some examples of key features that are often present when introducing a written work. Each feature is shown between the tag <example></example>. <example>the name of the written work.</example> <example>the author of the film.</example> <example>the type of the written work (e.g. novel, manga, journal).</example> <example>words related to book.</example> List 30 other important features that are often present when introducing a written work. Need to follow the template above, i.e. <example>features</example>.