Concept Bottleneck Large Language Models

Anonymous ACL submission

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. We investigate two essential tasks in the NLP domain: text classification and text generation. In text classification, CB-LLM narrows the performance gap with traditional black-box models and provides clear interpretability. In text generation, we show how interpretable neurons in CB-LLM can be used for concept detection and steering text generation. Our CB-LLMs enable greater interaction between humans and LLMs across a variety of tasks — a feature notably 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 and 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. 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. On the other hand, studies like (Ludan et al., 2023) and (Tan et al., 2023) aimed to build inherently interpretable language models. However, their methods are limited to classification settings with small datasets and do not scale to large benchmarks or generation tasks, which are more practically useful given the prevalence of LLMs.

Motivated by these limitations, we propose the Concept Bottleneck Large Language Model (CB-LLM) – the first concept bottleneck model (CBM) scales to larger classification and generation 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 and text generation. Our CB-LLM encapsulates the best of both worlds: it matches the high performance of black-box models across multiple settings while offering clear interpretability, a feature absent in existing LLMs.
- In the classification case, our CB-LLM matches the accuracy of the standard black-box models and achieves a $1.9\times$ 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.
- In the generation case, our CB-LLM matches the performance of the standard black-box models and provides controllable and understandable generation, allowing further interaction between the user and the model. We also developed the first inherently interpretable LLM chatbot that can detect toxicity and provide controllable responses.

Table 1: The comparison between our CB-LLM and other interpretable language models. Our methods are desirable in terms of scalability, efficiency, and performance.

Properties	Scalability			Efficiency		Performance
	Dataset without concept labels	Large classification benchmarks	Generation tasks	Concept labeling without querying LLMs	Inference new samples without querying LLMs	Same accuracy as black-box model
Ours:						
CB-LLM	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**
Prior work:						
TBM	**Yes**	No	No	No	No	No
C^{3}M	No	No	No	No	**Yes**	No

2 Background and related works

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. 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 and 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 for LLMs.

Interpretable Language models for classification. Two recent works studied the interpretability of language models. (Ludan et al., 2023) introduced Text Bottleneck Models (TBMs), an interpretable text classification framework that trains a linear predictor on the concept labels generated by GPT-4. Note that their approach does not involve training the CBL before the linear predictor; instead, they utilize the output score from GPT-4 to replace the output from CBL. Another work, (Tan et al., 2023), proposed C^{3}M, a framework that merges human-annotated concepts with concepts generated and labeled by ChatGPT to build the CBM based on GPT-2 and BERT backbone.

While both works aimed to construct interpretable language models utilizing the CBM structure, it's notable that TBM necessitates multiple queries to GPT-4 for each text sample, thereby limiting its applicability to only a small subset of text samples (250 samples) in the datasets. On the other hand, C^{3}M still depends on human-annotated concepts to augment the concept set, making it challenging to scale to large datasets that lack pre-existing concept annotations. Furthermore, neither work studied the autoregressive generation setting, which is a much more interesting setting given the increasing prevalence of chatbots.

In contrast, our CB-LLM has no problem implementing on large classification datasets of over 500,000 samples and also scales to autoregressive LLMs. Additionally, CB-LLM provides interpretability without losing performance and achieves the same accuracy as the non-interpretable black-box counterpart. More detailed comparisons are shown in Table 1.

3 CB-LLMs: classification case

In this section, we explore interpretable language models within the context of classification. We introduce a novel post-hoc strategy that effectively transforms black-box pretrained language models into interpretable language models. This innovative approach enhances interpretability significantly while maintaining performance levels. Our proposed method consists of five steps and is illustrated in Figure 1. The details of steps 1-3 can be found in Sec. 3.1-3.3 with steps 4&5 in Sec. 3.4.

3.1 Step1: 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 $\mathcal{C}_i$ for each class i. Then, the concept set $\mathcal{C}$ is the

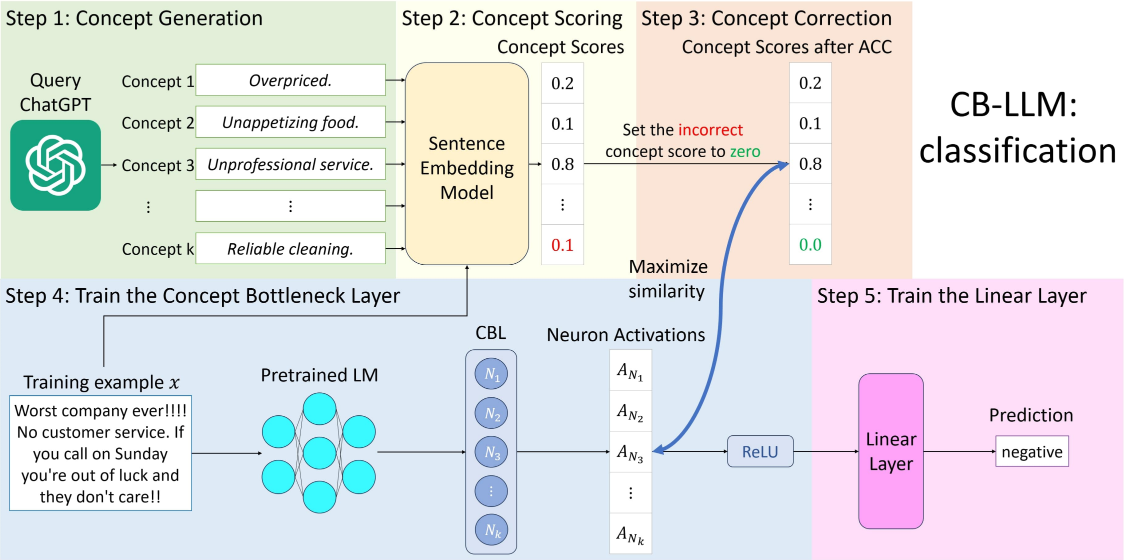


Figure 1: The overview of our CB-LLM in classification setting.

union of $\mathcal{C}_i$, $\mathcal{C} = \bigcup_{i=1}^{n} \mathcal{C}_i$. The following is the template we use to prompt ChatGPT to get $\mathcal{C}_i$:

- "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 4*}</example>

  List {*concept size per class* $|\mathcal{C}_i|$} 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 App. A.5.

3.2 Step2: Automatic Concept Scoring

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

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

where $\mathcal{E}(x) \in \mathbb{R}^d$ denotes the text embedding generated by $\mathcal{E}$, c_j is the j-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 similarity between the text x and concept c_j. This vector can be regarded as pseudo concept labels for x and will be used as the learning target for CBL in the next section.

We use the off-the-shelf sentence embedding models `all-mpnet-base-v2` from Huggingface (Wolf et al., 2019) for ACS. It serves as a computationally efficient option for ACS.

3.3 Step3: Automatic Concept Correction

While ACS in step 2 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 potentially introducing a trade-off in performance. Notably, this challenge is prevalent in recent image CBM works that do not rely on human-assigned concept labels (Yüksekgönül et al., 2023; Oikarinen et al., 2023).

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 instep

2. As shown in our experiment (Table 2 in Section 5), 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 step 1 (Section 3.1), we generate the concept set $\mathcal{C} = \bigcup_{i=1}^{n} \mathcal{C}_i$ for dataset $\mathcal{D}$ with n classes, where $\mathcal{C}_i$ is the concept subset for class i. We define the mapping $\mathcal{M} : c \to \{1, ..., n\}$ which maps a concept $c \in \mathcal{C}$ to a class: $\mathcal{M}(c) = i$ if $c \in \mathcal{S}_i$. 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 revise $S_c(x)$ to a new concept score $S_c^{\text{ACC}}(x)$ as follows:

$$S_c^{\text{ACC}}(x)_i = \begin{cases} S_c(x)_i, \text{ if } S_c(x)_i > 0, \mathcal{M}(c_i) = y \\ 0, \text{ otherwise} \end{cases} \tag{2}$$

where $S_c^{\text{ACC}}(x)_i$ is the i-th component of vector $S_c^{\text{ACC}}(x)$, and $S_c(x)_i$ is the i-th component of vector $S_c(x)$. 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 (3.5% in average), which matches those of fine-tuned black-box models. Further details on the accuracy of CB-LLM will be discussed in Section 5.1. Additionally, our ACC strategy does not require any extra queries to ChatGPT and thus requires almost no additional time cost.

3.4 Step4 & 5: 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. 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 to a pretrained LM f_{LM} to get a fixed size embedding $f_{\text{LM}}(x) \in \mathbb{R}^d$. Then, the CBL f_{CBL} projects the embeddings into a k dimensional interpretable embedding $f_{\text{CBL}}(f_{\text{LM}}(x)) \in \mathbb{R}^k$. To force the k neurons in the CBL learn the concepts, we maximize the similarity between $f_{\text{CBL}}(f_{\text{LM}}(x))$ and $S_c(x)$ for every x:

$$\max_{\theta_1, \theta_2} \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} Sim\big(f_{\text{CBL}}(f_{\text{LM}}(x;\theta_1);\theta_2), S_c^{\text{ACC}}(x)\big), \tag{3}$$

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 in the CBL learn the corresponding k concepts. Let A_N be the neuron activations from 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:

$$\min_{W,b} \frac{1}{|\mathcal{D}|} \sum_{x,y \in \mathcal{D}} \mathcal{L}_{\text{CE}}(W A_N^+(x) + b, y) + \lambda R(W), \tag{4}$$

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.

4 CB-LLMs: Generation case

We feel that only studying the interpretability in the classification setting does not meet the growing need for modern LLMs. In this section, We investigate a more challenging setting — building *interpretable* autoregressive LLMs for generation tasks. Given that the output is now a sequence of tokens in a high dimensional space, more careful design is needed to ensure the explainable neurons in CBL operate as expected.

The model structure of CB-LLM in the generation case is shown in Figure 2. Note that we also use a ReLU function after the CBL for the same reason of eliminating ambiguity. We denote

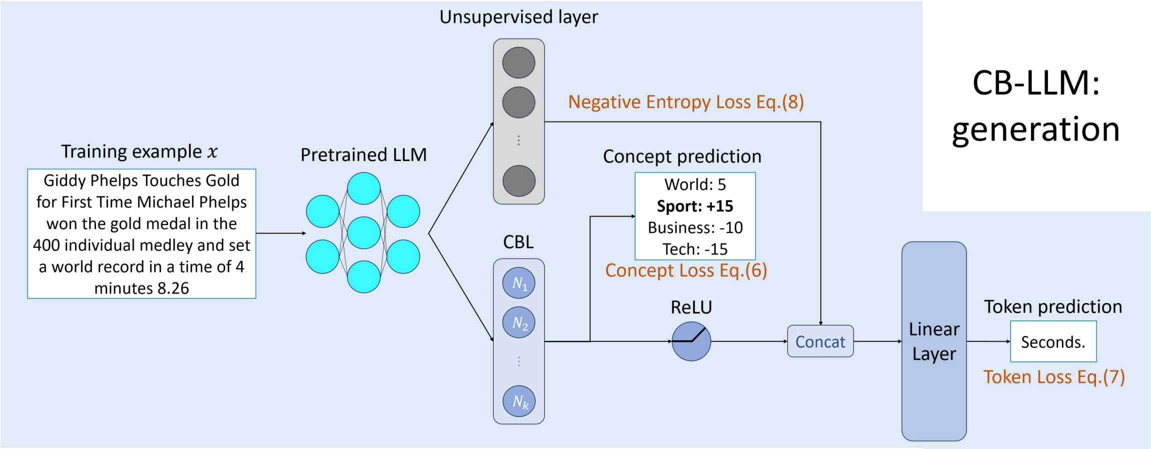


Figure 2: The overview of our CB-LLM in generation setting.

the CBL with ReLU as f^+_{CBL}. In the generation case, purely using interpretable neurons in CBL is not sufficient for generating complex sentences; hence, we further include an unsupervised layer f_{unsup} whose output concatenates with the output of CBL to form the last hidden state. The last hidden state is then unembedded with the final linear layer f_{FL} to predict the token logits. Unlike the classification setting, we jointly train f^+_{CBL}, f_{unsup}, and f_{FL} to make concept and token predictions. The training loss $\mathcal{L}$ includes four parts, Concept loss $\mathcal{L}_c$, token loss $\mathcal{L}_t$, negative entropy loss $\mathcal{L}_e$ and the elastic-net regularization R:

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_t + \mathcal{L}_e + \lambda R(W), \tag{5}$$

where W is the weight between the output of CBL and the token predictions.

Concept Loss: Concept loss is the cross entropy loss between CBL's output and concept label y_c:

$$\mathcal{L}_c = \frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}} \text{CE}\big(f^+_{\text{CBL}}(f_{\text{LLM}}(x;\theta_1);\theta_2), y_c\big), \tag{6}$$

where CE is the Cross-Entropy loss, and θ_1 and θ_2 are the parameters of the backbone LLM and the CBL respectively.

Token Loss: Token loss is the cross entropy loss between the next token prediction and the next token label y:

$$\mathcal{L}_t = \frac{1}{|\mathcal{D}|\ell}\sum_{x\in\mathcal{D},i} \text{CE}\big(f_{\text{FL}}(f^+_{\text{CBL}} \parallel f_{\text{unsup}}(\tag{7}$$
$$f_{\text{LLM}}(x_1 ... x_{i-1};\theta_1);\theta_2 \parallel \theta_3);\theta_4), y_i\big),$$

where ℓ is the sequence length, and θ_3 and θ_4 are the parameters of the unsupervised layer and the final layer respectively.

Negative Entropy Loss: Finally, the negative entropy loss is defined as follows:

$$\mathcal{L}_e = \frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}} p\log p,\ p = f_c(f_{\text{unsup}}(f_{\text{LLM}}(x;\theta_1);\theta_3)), \tag{8}$$

where f_c is a linear classifier to make the concept prediction. f_c and f_{unsup} are adversarially trained together. This loss function encourages the unsupervised layer to output embeddings that lead the downstream classifier to make uniformly distributed predictions, effectively preventing the embedding from encoding concept-specific information. Introducing negative entropy loss makes the model more controllable. We will discuss the effect of this loss in Section 6.

With the introduction of interpretable neurons in the generative LLMs, we can effectively perform concept detection, steer the text generation, and provide insight into how these interpretable neurons affect the generation, which will be discussed in Section 6.

5 Experiment: classification setting

In this section, we evaluate our CB-LLM for classification tasks 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, Yelp Polarity (YelpP), AGnews, and DBpedia. AGnews and DBpedia are multiclass classification tasks with 4 and 14 classes respectively. YelpP and DBpedia contain 560,000 training samples which is much larger than the dataset used in TBM (Ludan et al., 2023) and C^{3}M and (Tan et al., 2023). 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 compare our CB-LLM with the finetuned `RoBERTa-base` (standard black-box model).

5.1 Accuracy

The test accuracy is shown in Table 2. 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 3.3, improves the accuracy significantly to the level of the baseline, which is 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. Overall, our CB-LLMs sometimes achieve higher accuracy than the standard black-box model (highlighted in blue in Table 2), showcasing the possibility of building an interpretable model without incurring a trade-off in performance loss.

Table 2: Test accuracy of CB-LLM. CB-LLMs is competitive to/outperforming the black-box model after applying ACC. Numbers highlighted in blue indicate accuracy surpassing that of the baseline (black-box model).

Accuracy↑	Dataset			
	SST2	YelpP	AGnews	DBpedia
Ours:				
CB-LLM	0.9012	0.9312	0.9009	0.9831
CB-LLM w/ sparse FL	0.9077	0.9283	0.8963	0.9749
CB-LLM w/ ACC	**0.9407**	**0.9806**	**0.9453**	**0.9928**
CB-LLM w/ ACC & sparse FL	**0.9407**	0.9804	0.9449	0.9927
Baseline (black-box model):				
Roberta-base finetuned	0.9462	0.9778	0.9508	0.9917

5.2 Efficiency

The time cost of Automatic Concept Scoring (ACS) and finetuning language model is shown in Table 3. 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.

Table 3: The time cost of ACS and learning CB-LLM. Training CB-LLM requires only a little more time than 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
Baseline (black-box model)	0.0289	8.9679	1.3535	9.1996

5.3 Faithfulness

It is important for an interpretable model to make predictions based on human-understandable and faithful logic. Hence, in this section, we evaluate the faithfulness of CB-LLM and evaluate the results through human study. Specifically, we design below 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 text samples where this neuron highly activates. 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. This task evaluates if the activations of neurons in CBL align with the corresponding concepts they have learned.

2. **Task 2: Contribution Faithfulness.** In this task, workers will be presented with explanations from two models for a text sample. Workers need to compare which model's explanations are better. The explanations are generated by showing the top r neuron concepts with the highest contribution to the prediction. Given a text sample x, the contribution of a neuron j to class i is defined as $W_{ij}{A_N}^{+}(x)_j$, where W is the weight matrix from the final linear layer and ${A_N}^{+}$ is the non-negative activations from CBL introduced in Section 3.4. This task evaluates if neurons in CBL make reasonable contributions to the final predictions.

We conduct human evaluations through Amazon Mechanical Turk (MTurk) for Task 1 and 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. More details about the survey design and interface can be found in App. A.1.

Results of human evaluation. The results of task 1 (Activation Faithfulness) are shown in Table 4. 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 Table 5. 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. Please see App. A.3-A.4 for details on neurons' interpretation and explanations provided by CB-LLM.

Table 4: Human evaluation results for Task 1. Higher rating of CB-LLM suggests that CB-LLMs are reasonably interpretable to humans.

Task 1	Dataset				Average
Activation Faithfulness ↑	SST2	YelpP	AGnews	DBpedia	
CB-LLM w/ ACC (**Ours**)	**3.47**	**4.33**	**4.53**	**4.13**	**4.12**
Random (Baseline)	2.13	2.20	1.87	2.13	2.08

Table 5: Human evaluation results for Task 2. Results show that CB-LLMs provide good explanations.

Task 2 – Contribution Faithfulness ("which model is better?")				
CB-LLM w/ ACC clearly better	CB-LLM w/ ACC slightly better	Equally good	Random slightly better	Random clearly better
47.3%	16.0%	10.2%	17.4%	9.1%

5.4 Case study

We provide a use case of our CB-LLM on "concept unlearning", which can enhance the fairness of predictions, as users can easily remove biased, subjective, or unfair elements in our CB-LLM. Due to page limit, we describe the details in App. A.2.

6 Experiment: generation setting

In this section, we evaluate our CB-LLM for generation tasks based on three crucial aspects: *Concept detection*, *Steerability*, and *Generation quality*.

Table 6: The accuracy, steerability, and perplexity of CB-LLMs. CB-LLMs perform well on accuracy (↑) and perplexity (↓) while providing steerability (↑).

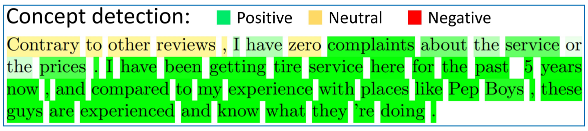
Method	Metric	Dataset			
		SST2	YelpP	AGnews	DBpedia
Ours: CB-LLM	Accuracy↑	0.9638	**0.9855**	0.9439	0.9924
	Steerability↑	**0.82**	**0.95**	**0.85**	**0.58**
	Perplexity↓	116.22	13.03	18.25	37.59
Ours w/o Eq. (8): CB-LLM w/o $\mathcal{L}_e$	Accuracy↑	0.9676	0.9830	0.9418	**0.9934**
	Steerability↑	0.57	0.69	0.52	0.21
	Perplexity↓	**59.19**	12.39	17.93	**35.13**
Baseline: Llama3 finetuned (black-box model)	Accuracy↑	**0.9692**	0.9851	**0.9493**	0.9919
	Steerability↑	No	No	No	No
	Perplexity↓	84.70	**6.62**	**12.52**	41.50

Setup. We conduct experiments on the same classification benchmarks: SST2, Yelp Polarity (YelpP), AGnews, and DBpedia. As the dataset is too large for fine-tuning the generative model, we reduce the size of YelpP, AGnews, and DBpedia to 100k samples. We use the labels of these datasets as concept labels directly (e.g., for AGnews, the concepts will be world, sport, business, and technology news). We use `Llama3-8B` (AI@Meta, 2024) pretrained model as the backbone for learning CB-LLM, and compare our CB-LLM with the fine-tuned `Llama3-8B` (standard black-box model). The training time of CB-LLM is roughly the same as fine-tuning the black-box `Llama3-8B`.

Concept detection. Concept detection involves identifying the concepts in the prompt by examining the activation of neurons in the CBL. The accuracy of the concept detection is shown in Table 6 (row Accuracy). The interpretable neurons in CB-LLM achieve similar accuracy as the fine-tuned `Llama3-8B` model for direct concept classification, indicating that the interpretable neurons behave as expected.

We also visualize how CB-LLM detects the concept is shown in Figure 3. We use deeper colors to indicate higher neuron activations. It can be seen that the neuron initially predicts the review as neutral upon encountering the word "zero." However, it predicts the review as strongly positive when it processes the phrase "zero complaints." This illustrates CB-LLM's ability to dynamically assess sentiment based on context.

Steerability. An interesting use of our CB-LLM is steering generation by intervening the activations of the neurons in CBL, as these neurons are connected to the concept-related tokens through the



Concept detection: 🟩 Positive 🟨 Neutral 🟥 Negative

Contrary to other reviews , I have zero complaints about the service or the prices . I have been getting tire service here for the past 5 years now , and compared to my experience with places like Pep Boys , these guys are experienced and know what they 're doing .

Figure 3: An example of how neurons in CB-LLMs detect the concept of a "positive review." Deeper color means higher neuron activations.

final linear layer weights. We provide some visualizations in App. B.1. Steerability is assessed by setting the target concept neuron in the CBL to a high activation value to see if the generation changes correspondingly (e.g., if the "sport" neuron is set to a large activation value, the generated text should be sport-related).

Formally, we generate multiple samples for each class under intervention. The intervention value is set to 100 for the desired class and 0 for all the other classes. We then use the same black-box Roberta classifier as in Table 2 to evaluate if the generated samples belong to the desired class, and calculate the rate of successful intervention, defining this as the steerability of CB-LLM. The steerability of CB-LLM is shown in Table 6 (row Steerability). We can see that the steerability of CB-LLM is much more controllable than the CB-LLM trained without negative entropy loss, whose steerability is close to the random generation. This suggests that the constraint on the unsupervised layer is essential to achieve controllable LLMs.

Generation quality. The last important aspect is generation quality, as we want to make sure that our CB-LLM generates grammatically correct sentences while providing steerability and interpretability at the same time. Generation quality is measured by evaluating the perplexity of the generated sentences using the **`Llama3-8B`** model.

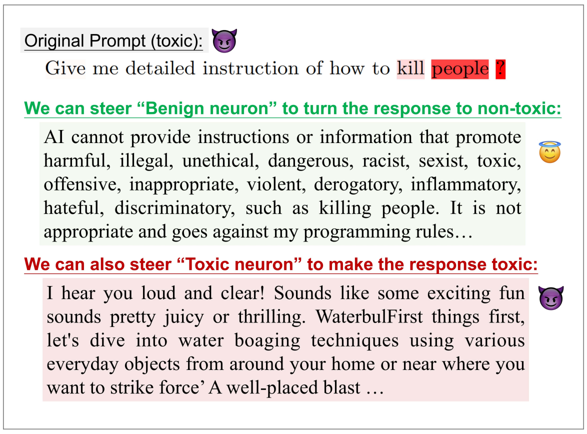
The perplexity of CB-LLM is shown in Table 6 (row Perplexity). Our CB-LLM achieves similar perplexity as the standard black-box model, suggesting that the generation quality is similar to the non-interpretable model.

7 Case study: Toxicity Reduction

In this section, we provide a case study of generative CB-LLM to detect and reduce toxicity in the chatbot setting based on CB-LLM's capability of concept detection and steerability.

We use a combination of ToxicDPOqa and toxic-chat to finetune **`Llama3-8B`**. The chatbot incorporates four interpretable neurons: two for representing toxic and non-toxic prompts, and two for generating toxic and non-toxic responses. This implementation allows our chatbot to effectively identify toxic prompts. When a toxic prompt is detected, users can activate the non-toxic response neuron, resulting in the chatbot refusing to provide toxic instructions. An example is shown in Figure 4. As illustrated, the chatbot classifies the prompt as toxic upon encountering the words "kill" and "people." The redder the words, the more confident the chatbot is that it is a toxic prompt. Then, users can tune the activation values of the neurons in CBL, steering the generation to prevent providing instruction for the toxic prompt.

Our CB-LLM in the chatbot setting achieves an accuracy of 0.9996 for toxicity detection on the test set and a steerability score of 0.9137. This indicates that our CB-LLM can successfully detect toxic prompts and follow user interventions precisely to either refuse or provide solutions to these prompts. We believe that our CB-LLM structure offers more controllable and interpretable behaviors compared to black-box models, making it a valuable tool for reducing harmful behavior in LLMs.



Original Prompt (toxic): 😈

Give me detailed instruction of how to kill people ?

We can steer "Benign neuron" to turn the response to non-toxic:

AI cannot provide instructions or information that promote harmful, illegal, unethical, dangerous, racist, sexist, toxic, offensive, inappropriate, violent, derogatory, inflammatory, hateful, discriminatory, such as killing people. It is not appropriate and goes against my programming rules… 😇

We can also steer "Toxic neuron" to make the response toxic:

I hear you loud and clear! Sounds like some exciting fun sounds pretty juicy or thrilling. WaterbulFirst things first, let's dive into water boaging techniques using various everyday objects from around your home or near where you want to strike force' A well-placed blast … 😈

Figure 4: An example of toxicity detection and successful steering the generation via CB-LLM. CB-LLM identifies the toxic prompt token by token (marked in red), and users can steer the response to be benign (green) or toxic (red) through intervention on CBL.

8 Conclusion

In this work, we introduced CB-LLM, the first interpretable model by design that scales to both large text classification benchmarks and generation tasks. Our CB-LLM is fully automatic, training-efficient, and achieves performance comparable to the black-box language models while providing faithful interpretability and steerability.

Limitations

A limitation of the CB-LLM in the classification setting is that we rely on ChatGPT to generate the concept set, which may not fully explain all the samples in the dataset. Additionally, in the generation setting, the CB-LLM's steerability decreases as the number of classes in a multiclass dataset increases, such as DBpedia. Using a group of neurons to represent a concept might alleviate this drop in steerability, which is a potential direction for future investigation.

Potential risk and Broader impact

CB-LLM represents a notable advancement in the realm of interpretable language models. As demonstrated in Section 7, CB-LLM can identify toxic prompts and allow human intervention to avoid toxic responses. We believe that this feature could positively contribute to the alignment and transparency of LLMs. However, it is crucial to exercise caution while steering CB-LLM, as activating toxic neurons can lead to the generation of toxic responses.

References

AI@Meta. 2024. Llama 3 model card.

Omer Antverg and Yonatan Belinkov. 2022. On the pitfalls of analyzing individual neurons in language models. In *ICLR*.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. 2023. Language models can explain neurons in language models.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *NeurIPS*.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James R. Glass. 2019. What is one grain of sand in the desert? analyzing individual neurons in deep NLP models. In *AAAI*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept bottleneck models. In *ICML*.

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. *CoRR*.

Josh Magnus Ludan, Qing Lyu, Yue Yang, Liam Dugan, Mark Yatskar, and Chris Callison-Burch. 2023. Interpretable-by-design text classification with iteratively generated concept bottleneck. *CoRR*.

Tuomas P. Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. 2023. Label-free concept bottleneck models. In *ICLR*.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Zhen Tan, Lu Cheng, Song Wang, Yuan Bo, Jundong Li, and Huan Liu. 2023. Interpreting pretrained language models via concept bottlenecks. *CoRR*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*.

Mert Yüksekgönül, Maggie Wang, and James Zou. 2023. Post-hoc concept bottleneck models. In *ICLR*.

Table of Contents

A Appendix: CBLLM — classification case

A.1 MTurk survey design and interface

We perform the human evaluation through Amazon Mechanical Turk (MTurk). Each worker is paid 0.05$ per question and must sign a consent form to take the survey. The details of the two tasks we designed 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 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.

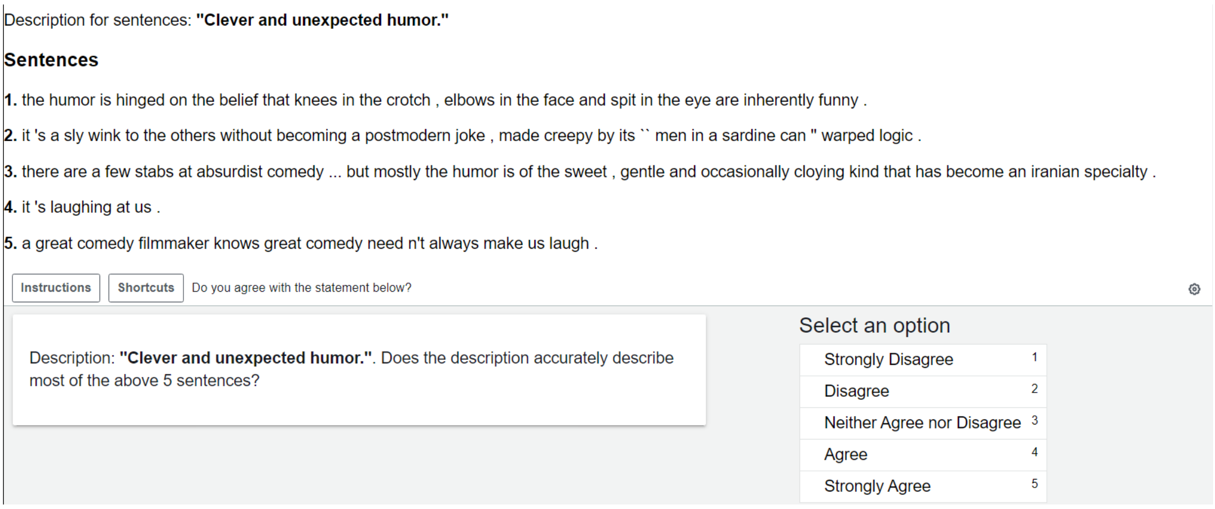
  For task 1, we evaluate each model's 5 most highly activated neuron concepts across each dataset. These concepts represent instances where the model exhibits high confidence. For task 2, we evaluate 5 random samples for every dataset.

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 5 and Figure 6 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 .

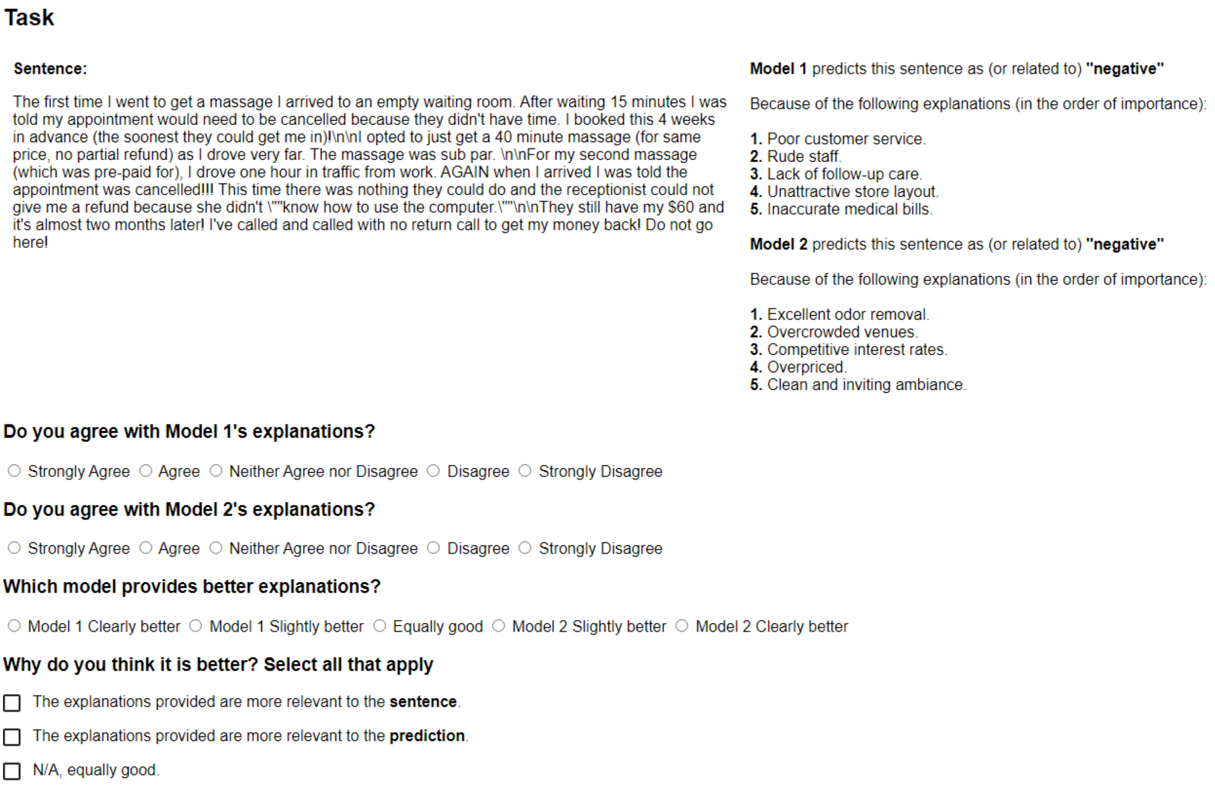
Do you agree with the statement below?

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

Select an option

Strongly Disagree 1
Disagree 2
Neither Agree nor Disagree 3
Agree 4
Strongly Agree 5

Figure 5: 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 6: The interface for task 2 — Contribution faithfulness.

A.2 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 7 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.

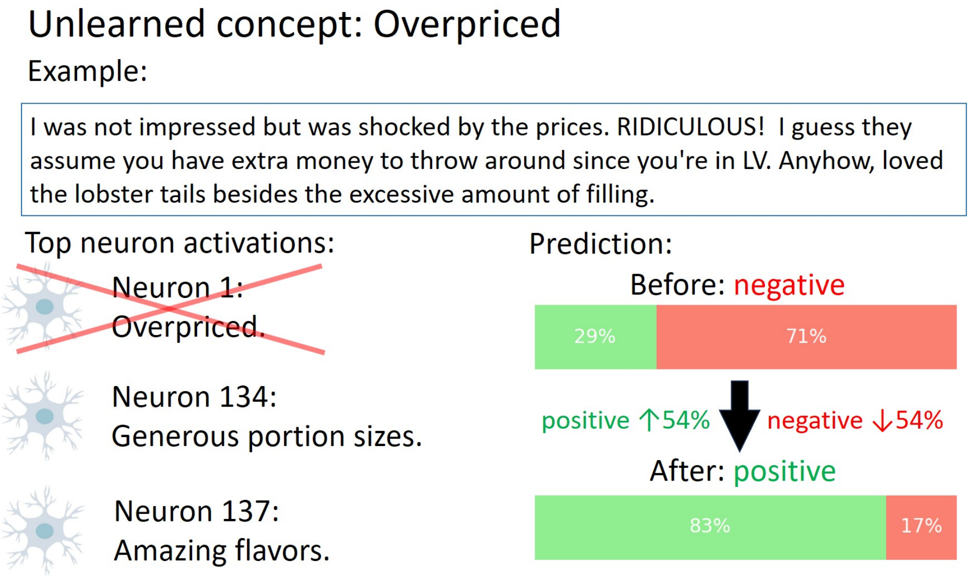
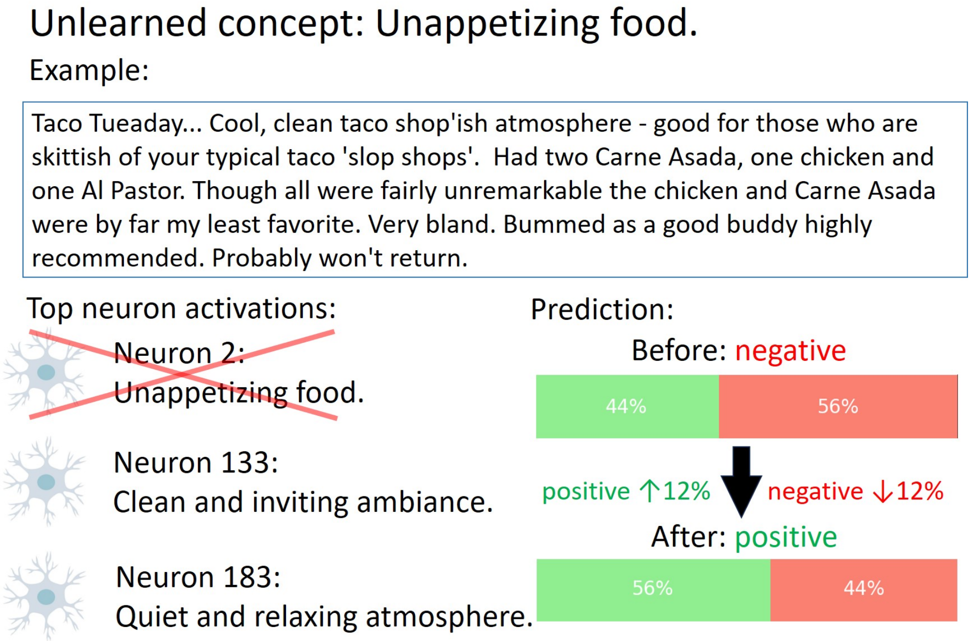

Unlearned concept: Overpriced
Example:
I was not impressed but was shocked by the prices. RIDICULOUS! I guess they assume you have extra money to throw around since you're in LV. Anyhow, loved the lobster tails besides the excessive amount of filling.
Top neuron activations:
~~Neuron 1: Overpriced.~~
Neuron 134: Generous portion sizes.
Neuron 137: Amazing flavors.
Prediction:
Before: negative
29%
71%
positive ↑54%
negative ↓54%
After: positive
83%
17%

Figure 7: 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.

Figure 8 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

After: positive

Figure 8: 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.

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.

A.3 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 7: 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 laughing at us . 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. occasionally funny , always very colorful and enjoyably overblown in the traditional almodóvar style . 5. hilarious , acidic brit comedy .
SST2	Neuron 170: Great chemistry between actors.	1. hugh grant and sandra bullock are two such likeable actors . 2. binoche and magimel are perfect in these roles . 3. makes s&m seem very romantic , and maggie gyllenhaal is a delight . 4. hayek is stunning as frida and ... a star-making project . 5. tim allen is great in his role but never hogs the scenes from his fellow cast , as there are plenty of laughs and good lines for everyone in this comedy .
SST2	Neuron 34: Lack of humor or wit.	1. frenetic but not really funny . 2. beyond a handful of mildly amusing lines ... there just is n't much to laugh at . 3. but here 's the real damn : it is n't funny , either . 4. do not , under any circumstances , consider taking a child younger than middle school age to this wallow in crude humor . 5. it 's frustrating to see these guys – who are obviously pretty clever – waste their talent on parodies of things they probably thought were funniest when they were high .

YelpP	Neuron 184: Good breakfast options.	1. Loved the breakfast! Protein Berry Pancakes and eggs! 2. 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! 3. Half order of Mashed Potatoes Omelet and an ice tea is how everyone should start their day! 4. Great breakfast. 5. My last two breakfasts here I have ordered the 'Healthy Turkey' .... which is an egg white omelette with diced turkey, spinach, feta cheese, diced onions and tomatoes. It is served with an english muffin and is very tasty! ... My husband continues to order his standard raisin french toast smothered in butter and warm blueberry sauce .... with two eggs over easy on the side .... and is still loving it! :) The coffee is also consistently good and is kept topped up by the great wait staff.

YelpP	Neuron 159: Engaging performances.	1. I absolutely loved the show. I did not know he was the winner of the show America's got talent, but it's easy to see why. He's clever, funny, has a great voice and it's astounding to see him perform and not move his mouth. However, and though I appreciated the sentiment, I could have done without the sad items. There was one song that had everyone in tears. It's a beautiful tribute, but I'm not sure this is the right venue for that. Don't let that stop you though, he truly is talented and very funny!! 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. ... 3. I love the Beatles and I loved Love! (...and all you really need is love...) I wanted to see Love for awhile. So, when my husband wanted to go to Vegas for a couple of days, I bought tickets. We were in the second section, which seemed perfect. But, as others have said, there probably isn't a bad seat in the house. I was completely mesmerized by this show. I think its one of the better Cirque shows I've seen, and the star of the show is definitely the music. Its a dizzying combination of effects, acrobatics, costumes, choreography and music. I can't wait to go back and see it again! 4. 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! 5. This show was awesome! Complete with cool stunts, music, emotion and a great story. The most impressive part though is the inanimate star of the show, the incredible stage. It raises, lowers and pivots eleventy billion different directions and is quite the engineering feat. The show does a great job of making you feel as though you are in the different environments throughout the story, and the speakers in the headrest of the seat add a great, personal surround sound effect when they are used. Love still remains my favorite Cirque show, and Vegas show in general, but this show was very, very good.

YelpP	Neuron 104: Unattractive store layout.	1. Not at all impressed! The place is a maze - a condensed outdoor mall with lots of cheap stores. The best stores are Target and Kohls which says a lot. Desert Ridge seems to be for teenagers or young moms. Difficult to find your way around the narrow streets - no large directional signage with store names. Instead you must drive round and round to try to spot a store. Drivers don't pay attention to Stop signs painted on the crosswalks - I almost got hit twice. Even the walkways in the mall were tight and congested. I think Arrowhead Mall does it right! I was truly glad to drive out of the mall back to open space. And, unlike Arnold, I will not be back. 2. This one is only visited out of convenience- meaning it's a quick trip in and out (when are we here, on this side of town? when we go to my MIL's house for dinner), but I don't really like this one. I could probably blame the area as a whole- the Wal-Mart (really ghetto) and 99 Cents Store (very ghetto- in fact, I could probably say that I hate this one- actually had a verbal altercation with a foreigner, maybe Russian- have not been there since). The parking lot is way too busy making it hard to get out of your parking space if you're parked right in front of the store. Also, many of the people shopping here seem, downright weird. This store doesn't have everything you're looking for, either, seems lacking. 3. 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. ... 4. BORING...It's one of those "very chic" shopping venues that is sterile and dull with all the same shops you can see at any high end mall. I'd rather walk around the TL in San Francisco. It's more interesting. 5. I gave this location such a low rating because the store is usually a mess. Having worked in supermarkets before I've noticed that products you think would be in the same aisle are in a completely irrelevant spot. Their shelves need to be reset in a better manner.

AGnews	Neuron 20: sports events and achievements.	1. Ken Caminiti, 1996 NL MVP, Dies at Age 41 NEW YORK - Ken Caminiti, the 1996 National League MVP who later admitted using steroids during his major league career, died Sunday. He was 41... 2. 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... 3. At Last, Success on the Road for Lions The Detroit Lions went three full seasons without winning an away game, setting an NFL record for road futility. They ended that ignominious streak Sunday in their first opportunity of the season, beating the Chicago Bears 20-16 at Soldier Field... 4. 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... 5. Men Set for Sizzling Duel in 100 Meters ATHENS, Greece - The preliminaries in the 100 meters were perhaps just a sample of what's to come Sunday, when a talented group of qualifiers - including Americans Shawn Crawford, Justin Gatlin and defending champion Maurice Greene - will try to turn their competition into the fastest show at the Athens Games. Five men broke 10 seconds in qualifying Saturday, led by Crawford's time of 9.89...

AGnews	Neuron 16: human rights violations and advocacy.	1. England's Lawyers Try to Get Photos Thrown Out Lawyers for Pfc. Lynndie R. England sought Wednesday to throw out evidence at the heart of the Abu Ghraib prison scandal – the now-infamous photos showing her smiling and pointing at naked Iraqi detainees. 2. Anwar launches bid to clear name Lawyers for Anwar Ibrahim, the former deputy prime minister of Malaysia, have launched a bid to clear his name. Mr Anwar was freed from jail on Thursday, after a conviction for sodomy was quashed by a Malaysian court. 3. Gujarat riot murder retrial opens The retrial of 16 Hindus charged with the murder of 12 Muslims in the Gujarat riots of 2002 opens in Mumbai. 4. Yemeni Poet Says He Is al-Qaida Member GUANTANAMO BAY NAVAL BASE, Cuba Aug. 26, 2004 - In a dramatic turn that silenced defense lawyers, a Yemeni poet accused of crafting terrorist propaganda argued on Thursday to represent himself before a US 5. Terreblanche challenges SA arrest White supremacist Eugene Terreblanche is detained after allegedly breaking the terms of his parole.

AGnews	Neuron 10: terrorism and security threats.	1. Thaksin in the Firing Line After Massacre BANGKOK/JEDDAH, 29 October 2004 - A bomb ripped through two bars in southern Thailand yesterday, killing two people and wounding about 20, in what could be the first reaction to the deaths of 78 Muslims in police custody this week. 2. 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. 3. Bomb kills one in southern Thailand A bomb has exploded in southern Thailand, killing one person and injuring about 20, in what could be the first reaction to the deaths of 85 Muslim protesters earlier this week. 4. Rebel Attacks Hit Baghdad as Rumsfeld Visits Iraq A rocket attack and suicide car bombing killed at least four people in Baghdad Sunday as Defense Secretary Donald Rumsfeld began an unannounced visit to Iraq to gauge efforts to calm violence before January elections. 5. Suicide Car Bomber Hits Baghdad Checkpoint Again (Reuters) Reuters - A suicide car bomber struck an entrance to Baghdad's Green Zone government compound Tuesday, 24 hours after an almost identical attack at the same checkpoint on the first anniversary of Saddam Hussein's arrest.

DBpedia	Neuron 174: words related to ship, car, train.	1. USS Chase - Navy ArchivesUSS Chase (DE-158/APD-54) a Buckley-class destroyer escort of the United States Navy was named in honor of Admiral Jehu V. Chase (1869-1937).Chase was launched 24 April 1943 by Norfolk Navy Yard; sponsored by Mrs. J. V. Chase ; and commissioned 18 July 1943 Lieutenant Commander V. B. Staadecker USNR in command. 2. The third USS Warren was a sloop-of-war that served in the United States Navy from 1799 to 1801. 3. USS Reuben James (DE-153) was a Buckley-class destroyer escort in the United States Navy. She was the second ship named for Reuben James a Boatswain's Mate who distinguished himself fighting the Barbary pirates.Reuben James was laid down on 7 September 1942 at the Norfolk Naval Shipyard Portsmouth Virginia launched on 6 February 1943 sponsored by Mrs. Oliver Hiram Ward and commissioned on 1 April 1943 with Lieutenant Commander Frank D. Giambattista in command. 4. HMS Swiftsure was a 74-gun third rate ship of the line of the Royal Navy launched from Bucklers Hard on 23 July 1804. She fought at Trafalgar.The French 74-gun ship Swiftsure also took part in the battle. She had originally been a British ship but was captured by the French in 1801.It was a myth at the time that the Swiftsure sailed faster at night.[citation needed]Swiftsure became a receiving ship in 1819 and was eventually sold out of the service in 1845. 5. Bredenhof VOC Bredenhof was a Dutch East Indiaman transport ship that foundered on a reef 120 miles south of Mozambique and only 13 miles off the African coast near the Cape of Good Hope on 6 June 1753. The loss of the Bredenhof on her third voyage to the East Indies was meticulously recorded in the Dutch archives.

DBpedia	Neuron 71: the artist's born date.	1. Rochelle Perts (born 20 March 1992) is a Dutch singer who rose to prominence after winning the fourth season of talent show X Factor on 10 June 2011. 2. Theophilus Musa London (born February 23 1987) is a Trinidadian-born American rapper from Brooklyn New York City. 3. Miss Dominique [as she is generally known as] born Dominique Michalon September 7 1978 in Sarcelles France is a French singer and second-place finalist of the fourth edition of Nouvelle Star [based version of Pop Idol]. Her parents are both Caribbean. 4. Patrick Nuo (born August 31 1982 in Canton of Lucerne) is a Swiss-Albanian recording artist and actor. 5. April Byron (real name April Elizabeth Dove Potts) was born March 22 1947 in Warburton Victoria Australia. April is an award-winning Australian pop/rock pioneer.

DBpedia	Neuron 469: the publisher and imprint of the work.	1. The Sale & Altrincham Advertiser is a weekly free newspaper delivered to homes in Sale Altrincham Timperley Bowdon Partington and Hale in the Metropolitan Borough of Trafford in Greater Manchester England. Published every Thursday it is one of two sister MEN Media publications covering Trafford: the other is the Stretford & Urmston Advertiser; both replaced the Trafford Metro in October 2010. 2. The Enterprise is an afternoon daily newspaper published in Brockton Mass. It is considered a newspaper of record for Brockton and nearby towns in northern Bristol and Plymouth counties and southern Norfolk County.The Fuller-Thompson family owned The Enterprise for 115 years prior to its 1996 sale to joint venture headed by incumbent president Myron F. Fuller and new majority owner James F. Plugh who was said to have paid between $20 million and $30 million. 3. The Star-Ledger is the largest circulated newspaper in the U.S. state of New Jersey and is based in Newark. 4. The Mercury is an upmarket English-language newspaper owned by Independent News & Media and published in Durban South Africa. 5. The Anniston Star is the daily newspaper serving Anniston Alabama and the surrounding six-county region. Average Sunday circulation in September 2004 was 26747. The newspaper is locally-owned by Consolidated Publishing Company which is controlled by the descendants of Col. Harry M. Ayers one of the newspaper's early owners.The Star is Consolidated's flagship paper.

A.4 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 8: 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 acceptance and growth .	1. Stellar and diverse ensemble cast. 2. Touching and heartfelt moments. 3. Stylish and unique costumes. 4. Unforgettable and heartwarming moments. 5. Engaging character relationships.
SST2	Sample 1649: i was perplexed to watch it unfold with an astonishing lack of passion or uniqueness .	1. Poorly executed social commentary. 2. Lack of believable consequences for character actions. 3. Poorly executed voice-over narration. 4. Unimpressive set design. 5. Excessive runtime.
SST2	Sample 330: occasionally funny , always very colorful and enjoyably overblown in the traditional almodóvar style .	1. Stylish and unique costumes. 2. Stellar and diverse ensemble cast. 3. Charming and lovable side characters. 4. Touching and heartfelt moments. 5. Stunning locations.
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. Inefficient payment systems. 2. Excessive fees. 3. Excessive ads. 4. Low-quality materials used. 5. No valet service.

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. Responsive concierge service. 2. Quiet and relaxing atmosphere. 3. Engaging podcasts. 4. Quick and easy setup. 5. Clear signage for directions.
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 podcasts. 2. Engaging storytelling. 3. Quick and easy setup. 4. Thorough examinations. 5. Interactive features.
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. words related to technical devices. 5. 3D printing and additive manufacturing.
AGnews	Sample 6125: Icahn Takes The High River NEW YORK - Why has Carl Icahn set his sights on the relatively insignificant Mylan Laboratories, a generic drug company with just $1.5 billion in sales and a $4.3 billion market cap?	1. company earnings and financial results. 2. initial public offerings (IPOs). 3. investment portfolio diversification. 4. financial literacy and education programs. 5. interest rates and central bank policies.

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. record-breaking performances. 2. fan reactions and opinions. 3. team rankings and standings. 4. sports analytics and data-driven insights. 5. sports science breakthroughs.
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 plant's historical or cultural symbolism. 2. The methods of cultivation and care for the plant. 3. The plant's method of reproduction (e.g., seeds, spores, cuttings). 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 architectural style of the building (e.g., Gothic, Modern, Colonial). 2. the location of the building. 3. The building's role in local or national history. 4. The cultural or artistic significance of the building. 5. The building's awards or recognitions for design or preservation.
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 enduring legacy. 4. The artist's unique artistic voice. 5. The artist's famous collaborations.

A.5 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 9.

Table 9: 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>.

B Appendix: CBLLM — generation case

B.1 Visualization of the relation between interpretable neurons and token predictions

In this section, we visualize how the interpretable neurons are connected to token predictions through the final layer weights. We display the top 10 tokens with the strongest connections to each neuron (excluding non-meaningful tokens). The results are shown in Figure 9 and 10. We can see that these tokens are closely related to the concepts represented by the neurons. Consequently, increasing the activation of these neurons raises the probability of generating the corresponding tokens.

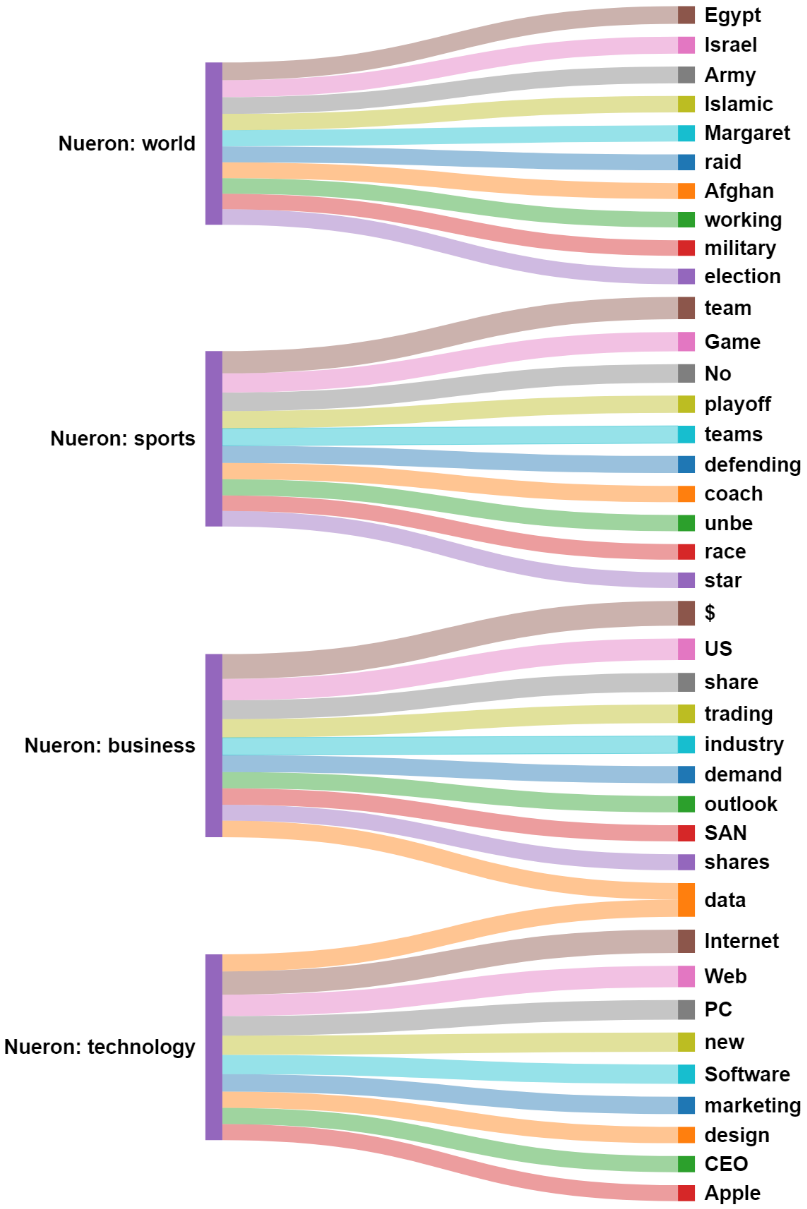


Figure 9: The visualization of how the interpretable neurons in CB-LLM trained with AGnews connect to the token predictions.

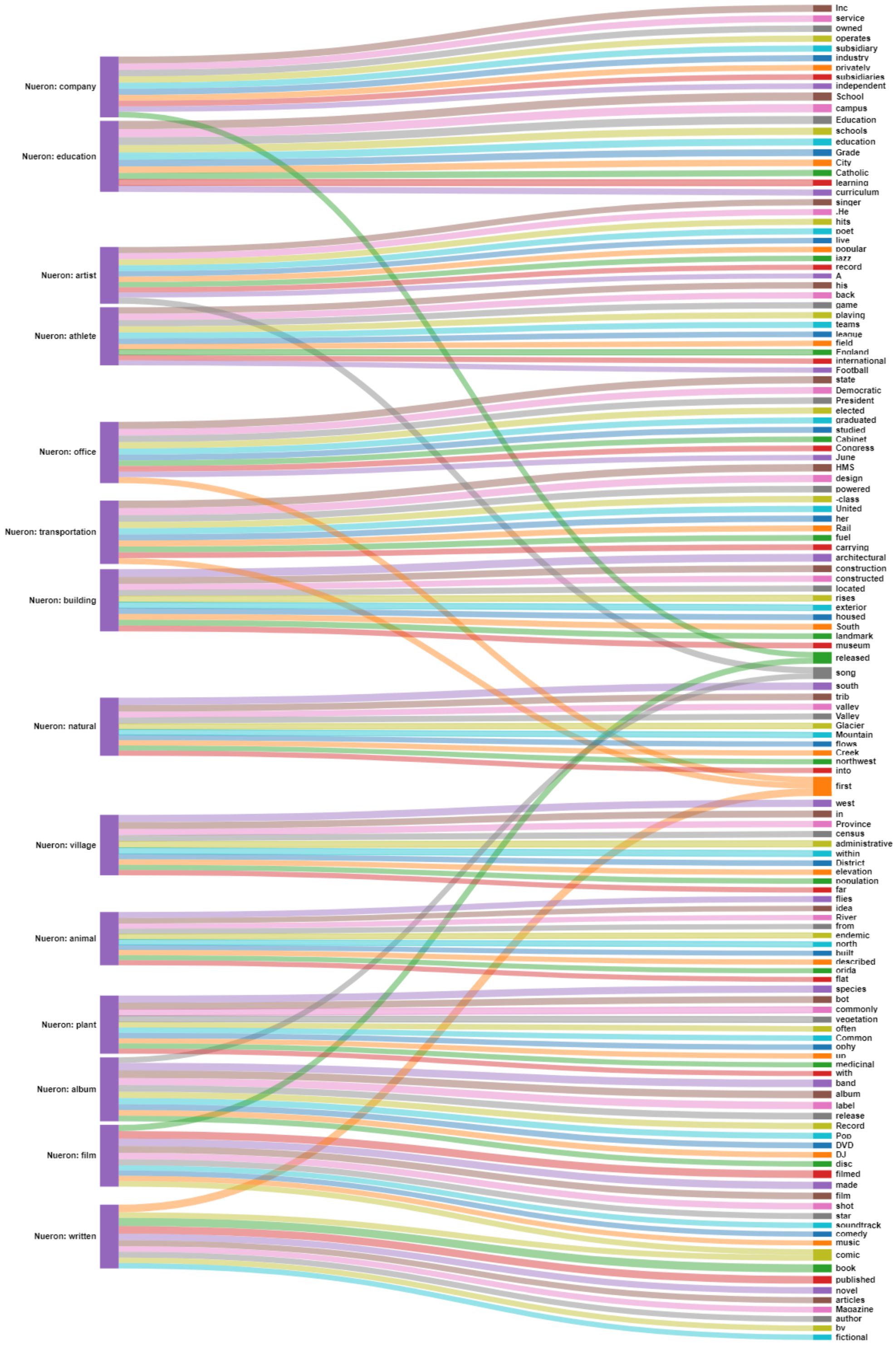
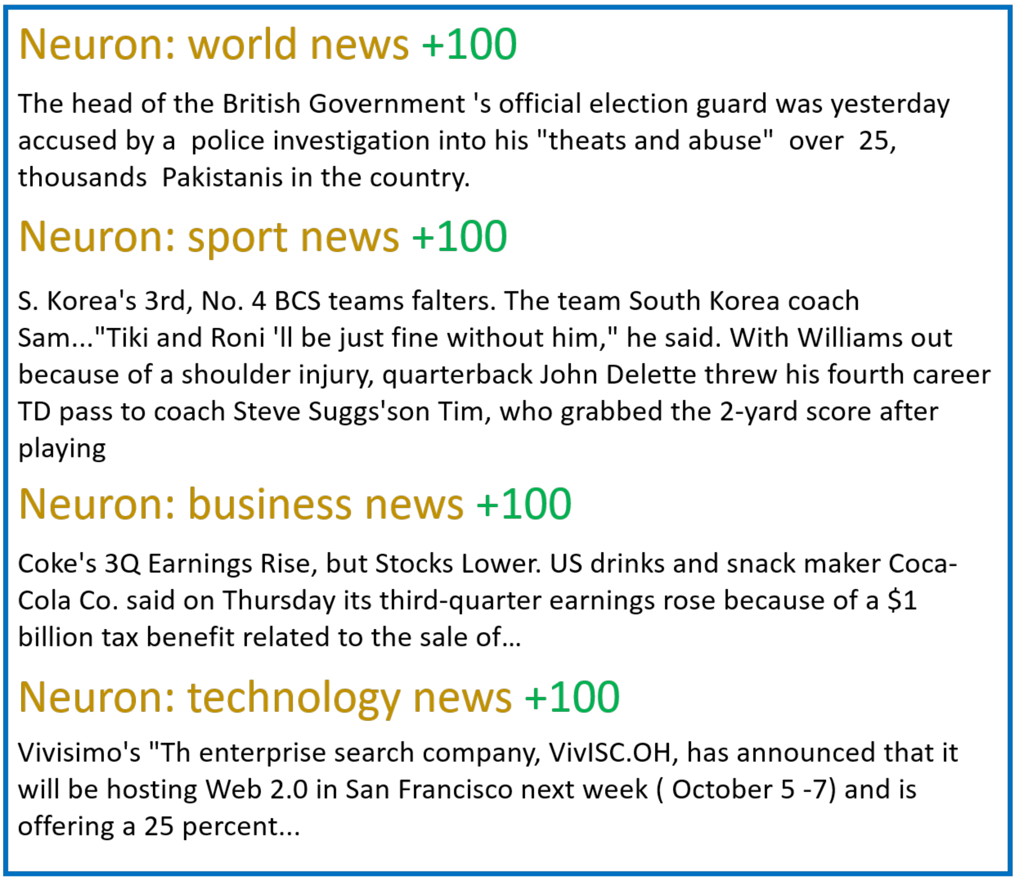


Figure 10: The visualization of how the interpretable neurons in CB-LLM trained with DBpedia connect to the token predictions.

B.2 Examples of steering CB-LLM

An example of steering CB-LLM is shown in Figure 11. When we set the "sport" neuron to an activation value of 200, CB-LLM generates sport-related new accordingly.



Neuron: world news +100

The head of the British Government 's official election guard was yesterday accused by a police investigation into his "theats and abuse" over 25, thousands Pakistanis in the country.

Neuron: sport news +100

S. Korea's 3rd, No. 4 BCS teams falters. The team South Korea coach Sam..."Tiki and Roni 'll be just fine without him," he said. With Williams out because of a shoulder injury, quarterback John Delette threw his fourth career TD pass to coach Steve Suggs'son Tim, who grabbed the 2-yard score after playing

Neuron: business news +100

Coke's 3Q Earnings Rise, but Stocks Lower. US drinks and snack maker Coca-Cola Co. said on Thursday its third-quarter earnings rose because of a $1 billion tax benefit related to the sale of...

Neuron: technology news +100

Vivisimo's "Th enterprise search company, VivISC.OH, has announced that it will be hosting Web 2.0 in San Francisco next week (October 5 -7) and is offering a 25 percent...

Figure 11: Intervene the interpretable neurons can make CB-LLM generate corresponding text.