Enhancing the Comprehensibility of Text Explanations via Unsupervised Concept Discovery

Yifan Sun[1,2] **Danding Wang**[1*] **Qiang Sheng**[1] **Juan Cao**[1,2] **Jintao Li**[1,2]
[1]Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
{sunyifan23z, wangdanding, shengqiang18z,caojuan,jtli}@ict.ac.cn

Abstract

Concept-based explainable approaches have emerged as a promising method in explainable AI because they can interpret models in a way that aligns with human reasoning. However, their adaption in the text domain remains limited. Most existing methods rely on predefined concept annotations and cannot discover unseen concepts, while other methods that extract concepts without supervision often produce explanations that are not intuitively comprehensible to humans, potentially diminishing user trust. These methods fall short of discovering comprehensible concepts automatically. To address this issue, we propose **ECO-Concept**, an intrinsically interpretable framework to discover comprehensible concepts with no concept annotations. ECO-Concept first utilizes an object-centric architecture to extract semantic concepts automatically. Then the comprehensibility of the extracted concepts is evaluated by large language models. Finally, the evaluation result guides the subsequent model fine-tuning to obtain more understandable explanations. Experiments show that our method achieves superior performance across diverse tasks. Further concept evaluations validate that the concepts learned by ECO-Concept surpassed current counterparts in comprehensibility.

1 Introduction

Deep neural language models lack explainability. A recent way to tackle this issue is concept-based explanations (Achtibat et al., 2023; Poeta et al., 2023), which map the inputs to a set of concepts and measure the importance of each concept to model predictions. By offering human-understandable attributes, concept-based explanations better resemble the way humans reason and explain.

Some existing concept-based methods apply post-hoc analysis to a trained model, providing concepts that either explain a model's prediction (Kim et al., 2018; Goyal et al., 2019; Ghorbani et al., 2019; Jourdan et al., 2023) or elucidate its internal network structure (Dalvi et al., 2022; Sajjad et al., 2022; Bills et al., 2023; Bricken et al., 2023). Since the model was not exposed to these concepts during training, it may lead to unfaithful interpretations or meaningless concept activations (Rudin, 2019). Self-explaining concept-based methods offer a more reliable approach by providing built-in explanations along with their predictions. However, most of these methods require abundant annotations (Tan et al., 2024a,b; De Santis et al., 2024) for predefined concepts, which shows their limitations in automatically discovering new concepts during training. Though there exist self-explaining methods that extract concepts without concept annotations (Rajagopal et al., 2021; Antognini and Faltings, 2021; Das et al., 2022), the extracted concepts often lack human comprehensibility due to the ignorance of such constraints to guide concept extraction. Consequently, the resulting concept explanations may highlight irrelevant or mislead-

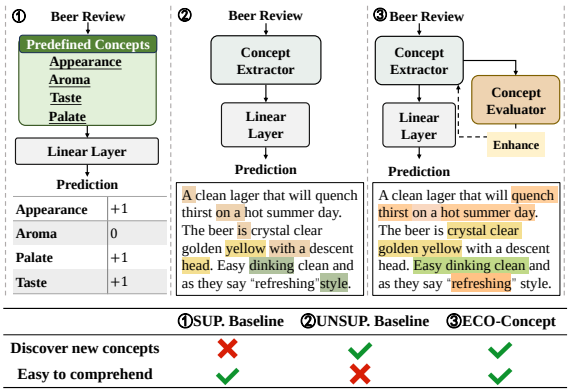


Figure 1: Comparison of explanations between our proposed ③ ECO-Concept and existing typical ① supervised and ② unsupervised concept-based methods. Supervised methods explain based on predefined concepts, while ECO-Concept and unsupervised methods explain via concept-related highlighted text. ECO-Concept eliminates the need for concept annotations and can discover unseen concepts with improved comprehensibility.

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2025, pages 14695–14713

ing information, bringing more confusion (demonstrated by previous research (Chaleshtori et al., 2024)) and less trust in humans (revealed by our human forward simulatability experiments). The comparison is illustrated in Fig. 1.

In this paper, we aim to build a model that can automatically discover comprehensible concepts with no need for concept supervision. Inspired by the recent application of the slot attention mechanism (Locatello et al., 2020) in concept-based interpretability for vision tasks (Wang et al., 2023; Hong et al., 2024), we develop a concept extractor that leverages slot attention, enabling each slot to learn a distinct task-specific concept. To ensure that the learned concepts are human-understandable, we leverage LLMs as human proxies to evaluate concept comprehensibility during model training and use the results to refine the concept extractor. Specifically, we propose a metric for concept comprehensibility that measures the capability of a concept's activation map to be summarized and reconstructed using natural language. Building on prior work that employs large language models to generate neuron explanations and simulate activations (Bills et al., 2023; Templeton et al., 2024), our approach introduces a novel feedback loop. This loop utilizes comprehensibility scores to refine the concept extractor, making the extracted concepts more intuitive and human-understandable. Experiments on seven tasks and human studies both show the superiority of our method compared with existing concept-based explanation methods. Our contributions are summarized as follows:

- We propose a method to evaluate the human interpretability of concepts by employing LLMs as human proxies, enabling real-time assessment of concept comprehensibility during training.
- We design a self-explaining model training mechanism without the need for concept annotations, where LLMs evaluate the discovered concepts and guide the subsequent model to learn concepts with greater interpretability.
- Our method demonstrates performance comparable to black-box models across various tasks, while the learned concepts are interpretable and comprehensible to humans.

2 Related Work

Concept-based Post-hoc Methods explain an existing model without modifying its internal architecture. Some studies predefine task-related concepts and assess them by quantifying contributions with linear probes (Alvarez Melis and Jaakkola, 2018; Crabbé and van der Schaar, 2022) or analyzing causal effects through proxy methods (Goyal et al., 2019; Wu et al., 2023). Methods without concept supervision extract concepts from intermediate layer representations using K-Means (Ghorbani et al., 2019), Non-Negative Matrix Factorization (Zhang et al., 2021; Fel et al., 2023; Jourdan et al., 2023) or concept completeness maximization (Yeh et al., 2020).

However, these post-hoc methods cannot guarantee that models truly comprehend or employ the adopted concepts, as the models were not exposed to these concepts during training (Rudin, 2019; Poeta et al., 2023). Additionally, some unsupervised post-hoc methods may extract concepts lacking semantic meaning.

Concept-based Self-explaining Methods aim to provide a built-in human-interpretable explanation along with the prediction. Some methods adopt a supervised paradigm with experts manually crafting a set of concepts. A representative framework is Concept Bottleneck Model (CBM) (Koh et al., 2020), where an intermediate concept bottleneck layer is introduced to break the standard end-to-end training paradigm. Recently, some studies have taken advantage of the capabilities of LLMs to generate concepts for each class as a replacement for manual annotation (Yang et al., 2023; Oikarinen et al., 2023). This insight has also been applied in the text domain (Tan et al., 2024b; De Santis et al., 2024). These methods rely on concepts that are predefined based on priors and cannot discover or adapt to new concepts during training. An incomplete or biased predefined concept set could severely compromise both the model's interpretability and performance. Text Bottleneck Model (TBM) (Ludan et al., 2023) utilized LLMs to automatically discover and measure concepts. However, in TBM, the mapping from samples to concepts is directly determined by LLMs, which remain largely opaque and function as a black box.

Unsupervised self-explaining concept models autonomously extract concepts during model training without the need for concept annotations. SENN (Alvarez Melis and Jaakkola, 2018) uses self-supervision by reconstruction loss for concept discovery. BotCL (Wang et al., 2023) and CCTs (Hong et al., 2024) utilize a slot attention-based mechanism to extract task-dependent concept slots. While in the text domain, SelfEx-

plain (Rajagopal et al., 2021) identifies concepts as the non-terminal leaves of the semantic tree parsing the text. Some methods classify samples based on their distance to learnable prototypes (Li et al., 2018; Das et al., 2022; Xie et al., 2023) and introduce distance-based losses to guide prototype learning. Without concept annotations, the concepts extracted by these methods pose challenges to human understanding. Although existing methods evaluate human interpretability through manual assessments during the experimental phase, no approach has yet incorporated human feedback into the training to optimize concept comprehensibility.

3 Method

3.1 Framework

Given a document consists of L tokens (words), we adopt a text encoder to encode tokens of the document as $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_L\} \in \mathbb{R}^{L\times D}$, where D is the dimension of the encoded space. ECO-Concept learns a set of M concept prototypes $\boldsymbol{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, ..., \boldsymbol{c}_M\} \in \mathbb{R}^{M\times D}$while learning the original classification task. Each concept prototype $\boldsymbol{c}_m$ is a trainable D-dimensional variable and is initialized randomly. These prototypes are continuously optimized during the training process to represent task-relevant concepts. Fig. 2 shows an overview of our framework. ECO-Concept consists of three modules: a concept extractor, a classifier, and a concept evaluator. The concept extractor takes the encoded text as input and generates a concept slot attention matrix by interacting with the concept prototypes. The slot attention scores across tokens are then summed and fed into the classifier for prediction. To enhance the comprehensibility of the concepts, we designed a concept evaluator that receives slot attention from the concept extractor, maps them back to the original text, and uses human proxies (here, LLMs)[1] to summarize the concepts, and highlights concept-related segments with additional exemplars. If a concept is easy to understand, the highlighted segments should be similar to the model's slot attention. We use the difference between the highlighted segments and the model's slot attention scores, considering the importance of each concept, as a comprehensibility loss. This feedback is used to make the learned concepts easier for human understanding.

[1]We conducted a human assessment of using LLMs to evaluate concept comprehensibility. The details are in Appendix B.

3.2 Concept Extractor

Concept Extractor uses slot attention to discover concepts automatically. Slot Attention (Locatello et al., 2020) was initially applied in object recognition, with its core mechanism focusing on utilizing slots to compete for explaining parts of the input features. Through multiple rounds of attention competition, these slots are gradually adjusted to represent distinct object features, i.e., the underlying concepts. In computer vision, several methods already utilize the slot attention mechanism to achieve concept-based interpretability (Wang et al., 2023; Hong et al., 2024). Inspired by these approaches, we implemented a concept extractor based on slot attention, where each slot learns a distinct task-dependent concept.

Our concept extractor takes the encoded features $\boldsymbol{X}$ as input and interacts with the concept prototypes $\boldsymbol{C}$ via the slot attention module, obtaining the concept features $\boldsymbol{U} \in \mathbb{R}^{M\times D}$ and the slot attention matrix $\boldsymbol{A} \in [0, 1)^{M\times L}$. For this, we apply linear projection $\boldsymbol{W}_q$ on the concept slots to obtain the queries and projections $\boldsymbol{W}_k$ and $\boldsymbol{W}_v$ on the inputs to obtain the keys and the values, all having the same dimension D. Then, we perform a dot product between the queries and keys to get the attention matrix $\boldsymbol{A}$.

$$\boldsymbol{A} = \phi(\frac{(\boldsymbol{W}_q\boldsymbol{C})(\boldsymbol{W}_k\boldsymbol{X})^T}{\sqrt{D}}). \quad (1)$$

In the slot attention matrix $\boldsymbol{A}$, each element $\boldsymbol{A}_{m,l}$ represents the attention weight of the concept slot m when attending to the input vector l. Unlike traditional attention mechanisms, we normalize $\boldsymbol{A}$ by applying a sparse softmax function ϕ across the concept slots (along the M axis). This normalization introduces competition among slots to attend to each input token. The sparsity normalization ensures that each input token is primarily associated with a limited number of concepts, facilitating a more focused and interpretable learning of conceptual representations. We then aggregate features in $\boldsymbol{X}$ corresponding to each concept into concept features $\boldsymbol{U} \in \mathbb{R}^{M\times D}$.

$$\boldsymbol{U} = \frac{\boldsymbol{A}}{\sum_l^L \boldsymbol{A}_{:,l}}(\boldsymbol{W}_v\boldsymbol{X}). \quad (2)$$

For better interpretability, we also employ concept regularizers to constrain the training of concept prototypes $\boldsymbol{C}$.

Consistency The extracted concepts should represent consistent semantics. That is, the concept

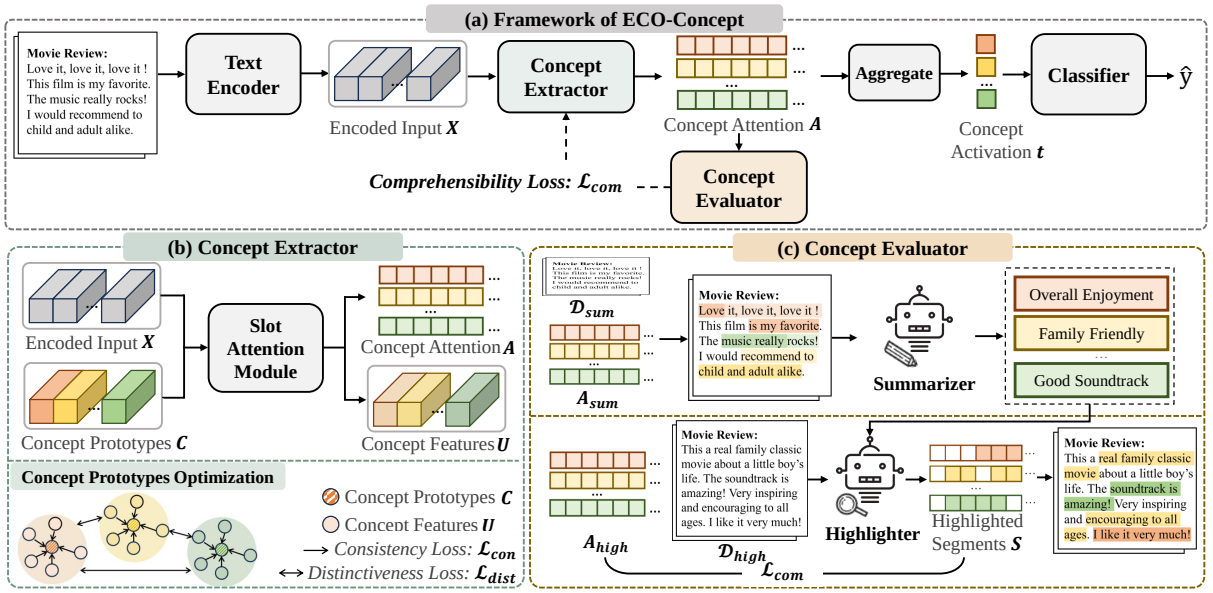


Figure 2: (a) Illustration of the proposed framework ECO-Concept. ECO-Concept consists of a concept extractor, a classifier, and a concept evaluator. (b) The concept extractor takes the encoded text $\boldsymbol{X}$ as input and interacts with the concept prototypes $\boldsymbol{C}$ to obtain a slot attention matrix $\boldsymbol{A}$ and concept features $\boldsymbol{U}$. The concept prototypes are optimized using consistency and distinctiveness loss. (c) The concept evaluator utilizes exemplars with the highest concept attention values to construct two sets, $\mathcal{D}_{sum}$ and $\mathcal{D}_{high}$. Using the corresponding slot attention matrices $\boldsymbol{A}_{sum}$ and $\boldsymbol{A}_{high}$, the evaluator highlights these exemplars to perform concept summarization and highlighting, thus getting the comprehensibility loss to guide model fine-tuning.

features activated by each concept prototype across the samples should not have large variations. The concept features $\boldsymbol{u}_m$ and $\boldsymbol{u}'_m$ under the same concept m of different documents should be similar to each other. Under each concept, we select the top k samples with the highest activation values in a mini-batch. We define the consistency loss as:

$$\mathcal{L}_{con} = \frac{1}{M} \sum_{m} \sum_{\boldsymbol{u}_m, \boldsymbol{u}'_m} \frac{\|\boldsymbol{u}_m - \boldsymbol{u}'_m\|_2^2}{k(k-1)}. \quad (3)$$

Distinctiveness To capture different aspects of documents, different concepts should cover different elements. This means that the average features of concept m calculated based on the top k samples with the highest activation values in a mini-batch, given by $\bar{\boldsymbol{u}}_m = (1/k) \sum \boldsymbol{u}_m$, should be different from any other. We encode this into a loss as:

$$\mathcal{L}_{dist} = - \sum_{\bar{\boldsymbol{u}}_m, \bar{\boldsymbol{u}}_{m'}} \frac{\|\bar{\boldsymbol{u}}_m - \bar{\boldsymbol{u}}_{m'}\|_2^2}{k(k-1)}. \quad (4)$$

3.3 Classifier

We use a fully connected layer for classification, and total concept activation $\boldsymbol{t} \in \mathbb{R}^M$ is the only input as the concept bottleneck (Koh et al., 2020):

$$\boldsymbol{t} = \sum_{l=1}^{L} \boldsymbol{A}_{:,l}. \quad (5)$$

Formally, letting $\boldsymbol{W} \in \mathbb{R}^{M \times \Omega}$ be a learnable matrix, prediction $\hat{y}$ is given by:

$$\hat{y} = \boldsymbol{W}\boldsymbol{t}. \quad (6)$$

We use softmax cross-entropy for the classification task, donated by $\mathcal{L}_{ce}$. The overall loss is

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_{con}\mathcal{L}_{con} + \lambda_{dist}\mathcal{L}_{dist}, \quad (7)$$

where λ_{con} and λ_{dist} are weight terms.

3.4 Concept Comprehensibility Enhancement

By optimizing the above loss function, our model achieved effective classification performance while extracting task-relevant concepts with semantic meanings. However, due to the lack of supervision from concept annotations, the identified concepts may still exhibit some degree of semantic ambiguity. To address this, we aim to further enhance the comprehensibility of the extracted concepts, prioritizing those of higher importance for improved human interpretability. Specifically, we evaluate each concept's importance and comprehensibility, combining these metrics accordingly to define a comprehensibility loss. This loss is integrated into the original loss function, and the model is further fine-tuned to improve concept comprehensibility.

Concept Importance. The weight matrix of the classifier $\boldsymbol{W}$ learns the correlation between the

class and concepts. For each concept m, $\boldsymbol{W}_{m,\omega}$ is the correlation from concept m to class ω. A positive value of $\boldsymbol{W}_{m,\omega}$ means that concept m co-occurs with class ω in the dataset, so its presence positively supports class ω. Meanwhile, a negative value means the concept m and the class ω rarely co-occurs. When concept m is present, the class is unlikely to be ω. Regardless of whether the value is positive or negative, $\boldsymbol{W}_{m,\omega}$ represents the relationship between concept m and class ω. Therefore, we take the absolute value of $\boldsymbol{W}_{m,\omega}$ here and compute the sum of the absolute values of all the weights connecting concept m as $\sum_{\omega=1}^{\Omega} |\boldsymbol{W}_{m,\omega}|$.

The importance of concept m should also take into account the concept activation. Given the average concept activation $\boldsymbol{t}_m$ in a mini-batch $\mathcal{B}$, the concept importance score β_m is:

$$\beta_m = \frac{1}{|\mathcal{B}|} \sum_{\boldsymbol{t} \in \mathcal{B}} \boldsymbol{t}_m \sum_{\omega=1}^{\Omega} |\boldsymbol{W}_{m,\omega}|. \tag{8}$$

Concept Comprehensibility. To measure the comprehensibility of concepts, we refer to the automated interpretability evaluation methods used for explaining neurons in LLMs (Bills et al., 2023; Templeton et al., 2024). In summary, the auto-interpretability procedure takes samples of text where the neurons activate, asks a language model to write a human-readable interpretation of the neuron features, and then prompts the language model to use this description to predict the neurons' activation on other samples of text. The correlation between the model's predicted and the actual activations is the feature's interpretability score.

We adopt a similar idea, measuring the comprehensibility of a concept by evaluating its ability to be summarized and reconstructed using natural language. For each concept, we select the exemplars with the highest attention values to form two sets, $\mathcal{D}_{sum}$ and $\mathcal{D}_{high}$, which are used for concept summarization and concept-related segment highlightings, respectively. We first present the exemplars $\mathcal{D}_{sum}$ with their corresponding slot attention values and ask the LLM to generate the interpretation. If it assesses that this concept has semantic meanings, we then prompt another LLM to highlight concept-related tokens on new exemplars from $\mathcal{D}_{high}$ in a 0-1 scale. For each highlighted exemplar, its slot attention matrix is $\boldsymbol{A} \in [0,1)^{M \times L}$ and the highlighted matrix obtained by LLM is $\boldsymbol{S} \in [0,1]^{M \times L}$. For concepts that are considered semantically meaningless by LLM, our goal is to minimize their corresponding activations. Specifically, we achieve this by setting the elements of $\boldsymbol{S}$ to zero for highlighted exemplars corresponding to these concepts.

The comprehensibility score of each concept is obtained by averaging the MSE loss between the slot attention matrix $\boldsymbol{A}$ and the highlighted matrix $\boldsymbol{S}$ for the highlighted samples within a mini-batch $\mathcal{B}$. To ensure that more important concepts are easier to understand, the comprehensibility loss is computed as a weighted combination of concept importance and comprehensibility:

$$\mathcal{L}_{com} = \frac{1}{M} \sum_{m=1}^{M} \frac{\beta_m \sum_{\boldsymbol{A} \in \mathcal{B}_m \cap \mathcal{D}_{high}} \|\boldsymbol{A}_{m,:} - \boldsymbol{S}_{m,:}\|_2^2}{|\mathcal{B}_m \cap \mathcal{D}_{high}|}. \tag{9}$$

Training Strategies. Based on the model trained in the first phase, we continue to train the concept prototypes $\boldsymbol{C}$ and the classifier using the following loss function:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_{con}\mathcal{L}_{con} + \lambda_{dist}\mathcal{L}_{dist} + \lambda_{com}\mathcal{L}_{com}. \tag{10}$$

At the end of each iteration, we conduct a concept re-summarization. If the meaning of a concept remains unchanged, the corresponding concept prototype parameters are frozen in the following iterations. Otherwise, we highlight concept-related segments based on the updated summarization. This process is repeated until the meanings of all concepts stabilize.

4 Experiments

In this section, we conduct extensive experiments to verify the task performance and interpretability of ECO-Concept. Due to space limitations, parameter sensitivity analysis and other experiments are included in Appendices E and F.

4.1 Experimental Settings

Datasets We conduct experiments on seven public datasets. To compare our results with supervised methods, we utilize three datasets with concept annotations: CEBaB (Abraham et al., 2022), Hotel (Wang et al., 2010), and Beer (McAuley et al., 2012). The rest four datasets, IMDB (Maas et al., 2011), AGnews (Gulli, 2004), Twitter (Sheng et al., 2021), and SciCite (Cohan et al., 2019), do not include concept annotations. More details about these datasets are in Appendix A.1.

Baselines The selected baselines include black-box, supervised concept-based, and unsupervised concept-based methods. The first group, black-box

Category	Method	CEBaB		Beer		Hotel		IMDB		AGnews		Twitter		SciCite	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Black-Box	RoBERTa	.682	.797	.882	.882	**.981**	**.981**	**.937**	**.937**	**.941**	.960	**.828**	.812	.858	.879
	BERT	.640	.770	.878	.878	.975	.975	.912	.912	.934	.956	.775	.745	.845	.861
SUP.	CBM	.669	.802	.883	**.885**	.979	.979	-	-	-	-	-	-	-	-
	SparseCBM	.644	.767	.883	.883	.981	.979	-	-	-	-	-	-	-	-
UNSUP.	SelfExplain	.683	.799	.873	.872	.978	.979	.936	.936	.925	.949	.817	.806	.856	.873
	PROTOTEX	.610	.728	.877	.877	.977	.977	.935	.935	.915	.943	.826	.812	.852	.871
	ECO-Concept	**.697**	**.808**	**.885**	**.885**	**.981**	**.981**	**.937**	**.937**	**.941**	**.961**	**.828**	**.813**	**.860**	**.881**

Table 1: Classification performance of ECO-Concept and other baselines. The best result is highlighted.

Method	CEBaB			Beer			Hotel			IMDB			AGnews			Twitter			SciCite		
	Sem.	Dist.	Con.	Sem.	Dist.	Con.	Sem.	Dist.	Con.	Sem.	Dist.	Con.	Sem.	Dist.	Con.	Sem.	Dist.	Con.	Sem.	Dist.	Con.
Cockatiel	.50	.40	.47	.55	.60	.42	.60	.65	**.49**	.35	.40	.41	.65	.60	.41	.50	.55	.53	.35	.35	.50
Concept-Shap	.25	.30	.42	.35	.15	.34	.65	.40	.35	.40	.30	.32	.35	.35	.35	**.80**	.30	.53	.40	.30	.44
ProtoTEx	.45	.45	.35	.20	.25	.38	.40	.30	.33	.25	.25	.36	.40	.45	.41	.45	.50	.47	.45	.45	.41
ECO-Concept	**.60**	**.60**	**.51**	**.85**	**.75**	**.49**	**.75**	**.70**	.48	**.65**	**.65**	**.52**	**.70**	**.65**	**.54**	.60	**.60**	**.55**	**.60**	**.50**	**.52**

Table 2: Concept comprehensibility evaluation of different concept-based methods. The best result is highlighted.

methods, directly tackles text classification tasks without interpretatability, including a **BERT-based classifier** (Devlin et al., 2019) and a **RoBERTa-based classifier** (Liu et al., 2019). The second group, supervised concept-based methods, leverages concept annotations to predict both the presence of concepts and the target class, including **CBM** (Kim et al., 2018) and an evolved CBM variant **SparseCBM** (Tan et al., 2024a). The third group, unsupervised concept-based methods, comprise two self-explaining methods **SelfExplain** (Rajagopal et al., 2021) and **PROTOTEX** (Das et al., 2022), and two post-hoc methods **COCKATIEL** (Jourdan et al., 2023) and **ConceptShap** (Yeh et al., 2020). The details of these methods are included in Appendix A.2.

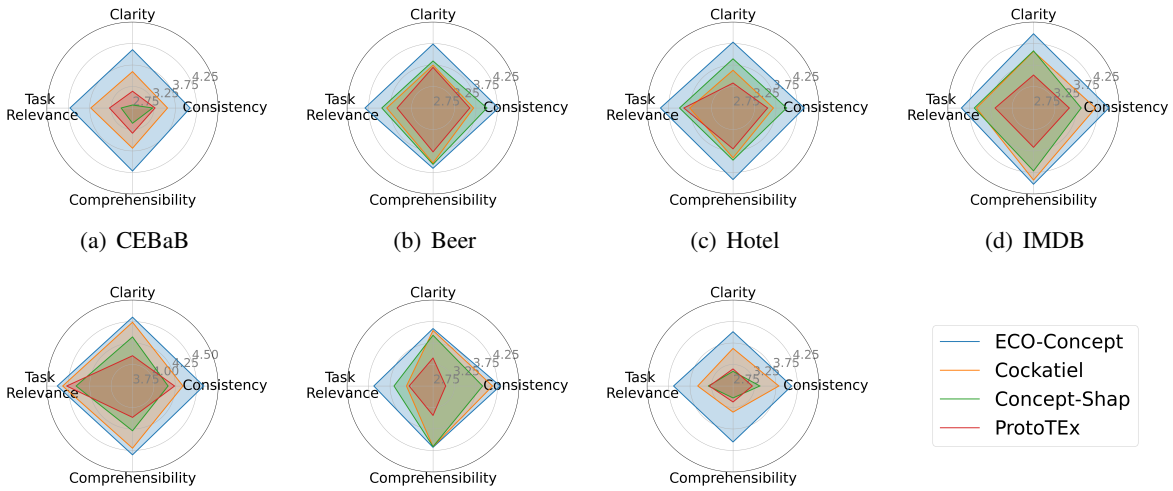
Implementation Details For all concept-based methods, we use RoBERTa as the text encoder and set the number of concepts to 20 across all tasks. For ECO-Concept, we provide the top 10 exemplars for each concept and utilize GPT-4o to summarize the concept interpretation. Additionally, we prompt GPT-4o-mini to highlight text segments on 100 exemplars per concept. The trade-off parameters λ_{con}, λ_{dist}, and λ_{com} are 0.1, -0.01, and 1, respectively. In our experiments, the concept comprehensibility enhancement stage ends within three iterations.

4.2 Task Performance

Table 1 shows the performance comparison of ECO-Concept and baselines. In general, our method achieves superior classification performance across various datasets. Compared to supervised methods, ECO-Concept achieves competitive results with no concept supervision, indicating its ability to automatically discover new concepts without compromising performance. Compared to unsupervised baselines, ECO-Concept shows significant improvements in both accuracy and F1 with pairwise t-tests at a 95% confidence level, validating the effectiveness of its conceptual representations. Moreover, ECO-Concept also has comparable or better performance compared with black-box models. This indicates that it effectively balances both task-discriminativity and concept comprehensibility, showing the potential to build interpretable models without performance trade-offs.

4.3 Concepts Evaluation

To evaluate the comprehensibility of the concepts extracted by our method, we first define three quantitative metrics. Then we conduct several human evaluations to further assess how easily these concepts can be understood. We compare our method's extracted concepts with those from three other global unsupervised concept extraction methods. Among these methods, Cockatiel and ConceptShap are post-hoc methods, and ProtoTEx is a self-explaining method. They extract concepts as representative training samples or text segments. For a fair comparison, we summarize the concept interpretation of these methods using the top 5 representative samples per concept, applying the same summary prompt as ECO-Concept. Note that we exclude SelfExplain in global concept comparisons, as it, differently, treats each concept as a single text segment, which cannot be further summarized.



(a) CEBaB (b) Beer (c) Hotel (d) IMDB

(e) AGnews (f) Twitter (g) SciCite (h) Legend

Figure 3: Human subjective ratings on concept quality

Quantitative Metrics. We assess the comprehensibility of concepts using the following metrics:

Semantics. (Sem.) This metric evaluates the proportion of extracted concepts that have clear semantic meaning. It is calculated as the ratio of concepts with identifiable semantics to the total number of extracted concepts.

Distinctiveness. (Dist.) This metric evaluates whether the extracted concepts are diverse and not redundant, which is the proportion of unique concepts identified by an LLM in all extracted ones.

Consistency. (Con.) This metric evaluates the internal consistency of the activation content within each concept. Specifically, we use the topic coherence score C_V (Röder et al., 2015), which measures whether the top 10 activated words within each concept collectively form a coherent topic.

As shown in Table 2, unlike post-hoc methods, our method is designed to balance task performance and interpretability simultaneously. However, we still achieve competitive performance in concept comprehensibility across various tasks. In the Hotel task, Cockatiel achieves a higher concept consistency. This is likely because Cockatiel extracts concepts from shorter sentences. However, in other tasks, our method demonstrates comparable or even superior consistency by directly extracting concepts from the original text. For the Twitter task, Concept-Shap achieves a higher proportion of semantically meaningful concepts. Nevertheless, the extracted concepts exhibit relatively low distinctiveness, suggesting that many of the concepts are redundant. In contrast, our method not only extracts semantically meaningful concepts but also ensures they are distinct from each other.

Human Evaluation. To further assess whether derived concepts are explainable to humans, we designed three human evaluation tasks, including intruder detection, subjective ratings, and forward simulatability. More details about the survey design and interface are in Appendix C.

To evaluate the comprehensibility of the extracted concepts, following Ghorbani et al. (2019); Fel et al. (2023), we designed an intruder detection experiment. In each question, participants were presented with four cases and tasked with identifying the one that was conceptually different from the others. A higher accuracy in identifying the intruder indicates that the concepts are more intuitive and easier for humans to understand. Table 3 summarizes the results. Our method achieves the highest intruder detection accuracy across all tasks, demonstrating that the concepts extracted by our approach are more comprehensible to humans.

In addition, we also conducted subjective rating experiments, where participants rated all the extracted concepts from multiple perspectives: *Consistency, Clarity, Task Relevance, Comprehensibility*. As shown in Fig. 3, our method consistently achieved the highest ratings across all perspectives for each task. This result indicates that our concepts are subjectively the most comprehensible to humans and are closely aligned with the tasks. In contrast, ProtoTEx received the lowest ratings in most tasks, suggesting that traditional unsupervised self-explaining methods often struggle to balance interpretability with task performance, resulting in less intuitive concepts for human understanding.

Method	CEBaB	Beer	Hotel	IMDB	AGnews	Twitter	SciCite
Cockatiel	.667	.633	.667	.833	.767	.850	.683
Concept-Shap	.550	.483	.483	.633	.683	.850	.433
ProtoTEx	.500	.500	.483	.683	.750	.733	.350
ECO-Concept	**.767**	**.750**	**.767**	**.900**	**.900**	**.867**	**.700**

Table 3: Accuracy of concept intruder detection

Method	Beer		AGnews	
	Accuracy	Confidence	Accuracy	Confidence
NE	.808	4.52	.708	4.44
Cockatiel	.967	4.13	.833	4.50
Concept-Shap	.833	4.22	.783	3.80
PROTOTEX	.950	4.28	.783	4.23
ECO-Concept	**.983**	**4.45**	**.867**	**4.55**

Table 4: Results of human forward simulatability

Finally, we validated the extracted concepts through a forward simulatability experiment to evaluate their effectiveness in helping users understand the model's behavior. This aligns with the definition of explainability proposed by Kim et al. (2016), which emphasizes a user's ability to "correctly and efficiently predict the method's results." In our experiment, participants received explanations of model outputs, then attempted to infer the model's outputs based on the explanations and rated their confidence. For comparison, we also measured the accuracy of human judgments without explanations (NE). We conducted forward simulatability experiments with explanations on the two most challenging tasks for human judgment (Beer and AGnews), as determined by the lowest human accuracy rates in no-explanation conditions. As shown in Table 4, our method significantly improved the simulatability accuracy for participants in inferring model outputs. This demonstrates that our explanations effectively enhance human understanding of the model's behavior and increase their trust in its predictions. Moreover, for other concept-based methods, although the simulatability accuracy is higher when their explanations are provided compared to when no explanations are given, humans generally tend to have lower confidence in these explanations. This further validates that existing methods often produce confusing explanations, which in turn reduces human trust. Besides, we also asked participants to rate the explanations for each example across three aspects. Detailed information and results are in Appendix C.4.

4.4 Ablation Study

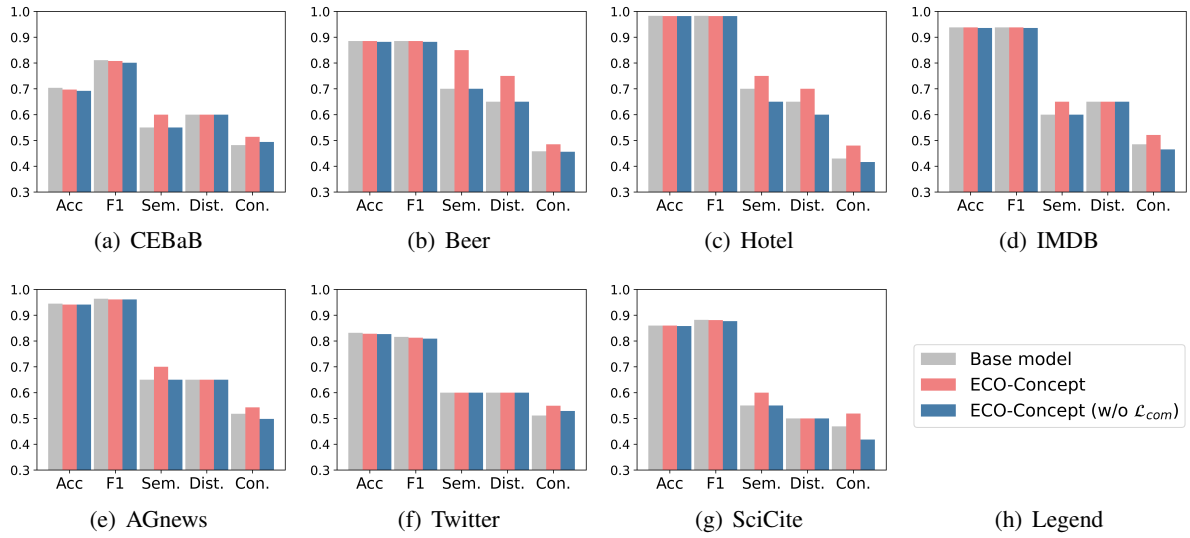
To further explore the impact of different model modules, we conducted several ablation studies.

Metric	Method	CEBaB	Beer	Hotel	IMDB	AGnews	Twitter	SciCite
Acc	w/o $\mathcal{L}_{con}$	.679	**.885**	.980	.937	.939	.823	.858
	w/o $\mathcal{L}_{dist}$	.699	**.885**	.977	.937	.940	.824	.852
	w/ all	**.704**	**.885**	**.981**	**.938**	**.945**	**.832**	**.860**
F1	w/o $\mathcal{L}_{con}$	.796	**.885**	.980	.937	.959	.811	.879
	w/o $\mathcal{L}_{dist}$	.806	**.885**	.977	.937	.960	.809	.875
	w/ all	**.811**	**.885**	**.981**	**.938**	**.964**	**.816**	**.882**
Sem.	w/o $\mathcal{L}_{con}$	.45	.65	**.70**	.40	.50	.40	.25
	w/o $\mathcal{L}_{dist}$	.50	**.70**	**.70**	.50	**.65**	.50	.15
	w/ all	**.55**	**.70**	**.70**	**.60**	**.65**	**.60**	**.55**
Dist.	w/o $\mathcal{L}_{con}$	.50	**.65**	.60	.40	.45	.50	.35
	w/o $\mathcal{L}_{dist}$	.30	.50	.55	.50	.55	.55	.20
	w/ all	**.60**	**.65**	**.65**	**.60**	**.65**	**.60**	**.50**
Con.	w/o $\mathcal{L}_{con}$	.43	.40	.41	.38	.47	.45	.42
	w/o $\mathcal{L}_{dist}$	**.48**	.41	.41	.40	.51	.45	.42
	w/ all	**.48**	**.46**	**.43**	**.49**	**.52**	**.51**	**.47**

Table 5: Task performance and concept metrics comparison between our method and its ablative variants

Impact of Concept Regularizers. To analyze the effects of the concept consistency and concept distinctiveness regularizers, we conduct an ablation study. We experiment on two types of variant models: w/o $\mathcal{L}_{con}$ and w/o $\mathcal{L}_{dist}$, which respectively remove the concept consistency regularizer and the concept distinctiveness regularizer. Experimental results are shown in Table 5. The results show that the best task performance and concept interpretability are achieved when both regularizers are applied together. By comparing the w/ all and w/o $\mathcal{L}_{con}$ variants, removing the consistency loss significantly reduces the consistency score, demonstrating the importance of the consistency regularizer. A similar trend is observed with the distinctiveness regularizer. These results highlight the importance of incorporating both types of regularization. Further analysis is provided in Appendix D.1.

Impact of Concept Comprehensibility Enhancement. To demonstrate the importance of the concept comprehensibility enhancement stage, we compare the model before enhancement (denoted as the Base model), ECO-Concept, and ECO-Concept (w/o $\mathcal{L}_{com}$). We introduce the ECO-Concept (w/o $\mathcal{L}_{com}$) variant to exclude the effect of longer training with concept regularizers, ensuring that any improvements in comprehensibility are due to $\mathcal{L}_{com}$. From the result in Fig. 4, we observe that after concept enhancement, classification performance remains largely unchanged (or experiences only a slight decline). Regarding concept comprehensibility, our ECO-Concept yields the best performance. After the enhancement, the semantics and consistency of concepts are further improved across most tasks. Besides, we also discovered more diverse concepts in two tasks. Fur-



(a) CEBaB (b) Beer (c) Hotel (d) IMDB

(e) AGnews (f) Twitter (g) SciCite (h) Legend

Figure 4: Classification performance and concept metrics comparison of the model before concept enhancement (the Base model), ECO-Concept, and ECO-Concept (w/o $\mathcal{L}_{com}$).

ther analysis is provided in Appendix D.2.

5 Conclusion

To automatically extract human-understandable concept explanations with no predefined concept annotations, we proposed ECO-Concept, an intrinsically interpretable framework. ECO-Concept employs an object-centric architecture based on the slot attention mechanism to extract task-specific semantic concepts. To ensure the extracted concepts are comprehensible to humans, ECO-Concept incorporates LLMs as human proxies to evaluate concept comprehensibility during model training, using their feedback to iteratively refine the concept extractor. ECO-Concept achieves classification performance on par with black-box models while offering enhanced interpretability. It outperforms existing concept-based approaches in both quantitative metrics and user studies, demonstrating its effectiveness in generating human-aligned, interpretable explanations.

Limitations

While our approach presents a significant step towards more interpretable models, several limitations warrant further exploration. First, in this work, we focus on enhancing pre-trained language models with self-explainable structures to improve interpretability in text classification tasks. Given the inherent challenges of unsupervised concept extraction, we employ a comparatively lightweight BERT-based model as the backbone. For future work, we plan to explore concept extraction methods compatible with larger models and extend our framework to encompass state-of-the-art architectures such as LLaMA and Mixtral. Second, to ensure the assessment of concept comprehensibility is close to the human level, we utilized two well-recognized API-based LLMs based on their strong instruction-following capabilities. However, due to cost limitations, we were only able to simulate concepts on a subset of samples. The current setting represents the best trade-off between performance and cost within our acceptable range. In the future, we plan to test more cost-effective and deployable open-source LLMs. Lastly, in our method, the number of concepts is fixed and cannot be adaptively adjusted during training. Currently, there are no well-suited approaches to address this challenge. We are actively exploring potential techniques (such as incremental learning) to adjust the number of concepts in a more flexible manner.

Acknowledgment

The authors would like to thank the anonymous reviewers for their insightful comments. This work is supported by the Innovation Funding of ICT, CAS under Grant (No. E561160), the Strategic Priority Research Program of the Chinese Academy of Sciences (No. XDB0680202), the National Natural Science Foundation of China (62406310), the Postdoctoral Fellowship Program of CPSF (GZC20232738), and the China Postdoctoral Science Foundation (2024M763336).

References

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

Reduan Achtibat, Maximilian Dreyer, Ilona Eisenbraun, Sebastian Bosse, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. 2023. From attribution maps to human-understandable explanations through concept relevance propagation. *Nature Machine Intelligence*, 5(9):1006–1019.

David Alvarez Melis and Tommi Jaakkola. 2018. Towards robust interpretability with self-explaining neural networks. In *Proceedings of the International Conference on Neural Information Processing Systems*, volume 31, pages 7786–7795.

Diego Antognini and Boi Faltings. 2021. Rationalization through concepts. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 761–775.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. 2018. Deriving machine attention from human rationales. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1903–1913.

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. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Fateme Hashemi Chaleshtori, Atreya Ghosal, Alexander Gill, Purbid Bambroo, and Ana Marasović. 2024. On evaluating explanation utility for human-ai decision making in nlp. *Preprint*, arXiv:2407.03545.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3586–3596.

Jonathan Crabbé and Mihaela van der Schaar. 2022. Concept activation regions: A generalized framework for concept-based explanations. In *Proceedings of the International Conference on Neural Information Processing Systems*, volume 35, pages 2590–2607.

Fahim Dalvi, Abdul Rafae Khan, Firoj Alam, Nadir Durrani, Jia Xu, and Hassan Sajjad. 2022. Discovering latent concepts learned in bert. *Preprint*, arXiv:2205.07237.

Anubrata Das, Chitrank Gupta, Venelin Kovatchev, Matthew Lease, and Junyi Jessy Li. 2022. Prototex: Explaining model decisions with prototype tensors. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 2986–2997.

Francesco De Santis, Philippe Bich, Gabriele Ciravegna, Pietro Barbiero, Danilo Giordano, and Tania Cerquitelli. 2024. Self-supervised interpretable concept-based models for text classification. *Preprint*, arXiv:2406.14335.

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

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. 2023. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2711–2721.

Amirata Ghorbani, James Wexler, James Zou, and Been Kim. 2019. Towards automatic concept-based explanations. In *Proceedings of the International Conference on Neural Information Processing Systems*, pages 9277–9286.

Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. 2019. Explaining classifiers with causal concept effect (cace). *Preprint*, arXiv:1907.07165.

Antonio Gulli. 2004. Ag's corpus of news articles. http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html.

Jinyung Hong, Keun Hee Park, and Theodore P Pavlic. 2024. Concept-centric transformers: Enhancing model interpretability through object-centric concept learning within a shared global workspace. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4880–4891.

Fanny Jourdan, Agustin Picard, Thomas Fel, Laurent Risser, Jean-Michel Loubes, and Nicholas Asher. 2023. Cockatiel: Continuous concept ranked attribution with interpretable elements for explaining neural net classifiers on nlp tasks. In *61st Annual Meeting of*

the Association for Computational Linguistics (ACL 2023), pages 5120–5136.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. In *Proceedings of the International Conference on Neural Information Processing Systems*, volume 29, pages 2288–2296.

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

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

Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. 2018. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pages 3530–3537.

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. *Preprint*, arXiv:1907.11692.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. 2020. Object-centric learning with slot attention. In *Proceedings of the International Conference on Neural Information Processing Systems*, volume 33, pages 11525–11538.

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. *Preprint*, arXiv:2310.19660.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 142–150.

Julian McAuley, Jure Leskovec, and Dan Jurafsky. 2012. Learning attitudes and attributes from multi-aspect reviews. In *2012 IEEE 12th International Conference on Data Mining*, pages 1020–1025. IEEE.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Tuomas Oikarinen, Subhro Das, Lam Nguyen, and Lily Weng. 2023. Label-free concept bottleneck models. In *International Conference on Learning Representations*.

Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. 2023. Concept-based explainable artificial intelligence: A survey. *Preprint*, arXiv:2312.12936.

Dheeraj Rajagopal, Vidhisha Balachandran, Eduard H Hovy, and Yulia Tsvetkov. 2021. Selfexplain: A self-explaining architecture for neural text classifiers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 836–850.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM International Conference on Web Search and Data Mining*, pages 399–408.

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

Hassan Sajjad, Nadir Durrani, Fahim Dalvi, Firoj Alam, Abdul Khan, and Jia Xu. 2022. Analyzing encoded concepts in transformer language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 3082–3101.

Qiang Sheng, Xueyao Zhang, Juan Cao, and Lei Zhong. 2021. Integrating pattern-and fact-based fake news detection via model preference learning. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1640–1650.

Zhen Tan, Tianlong Chen, Zhenyu Zhang, and Huan Liu. 2024a. Sparsity-guided holistic explanation for llms with interpretable inference-time intervention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21619–21627.

Zhen Tan, Lu Cheng, Song Wang, Bo Yuan, Jundong Li, and Huan Liu. 2024b. Interpreting pretrained language models via concept bottlenecks. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 56–74. Springer.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. 2024. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Bowen Wang, Liangzhi Li, Yuta Nakashima, and Hajime Nagahara. 2023. Learning bottleneck concepts in image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10962–10971.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 783–792.

Zhengxuan Wu, Karel D'Oosterlinck, Atticus Geiger, Amir Zur, and Christopher Potts. 2023. Causal proxy models for concept-based model explanations. In *International Conference on Machine Learning*, pages 37313–37334. PMLR.

Sean Xie, Soroush Vosoughi, and Saeed Hassanpour. 2023. Proto-lm: A prototypical network-based framework for built-in interpretability in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3964–3979.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. 2023. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19187–19197.

Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. 2020. On completeness-aware concept-based explanations in deep neural networks. In *Proceedings of the International Conference on Neural Information Processing Systems*, volume 33, pages 20554–20565.

Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A Ehinger, and Benjamin I.P. Rubinstein. 2021. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11682–11690.

A Experimental Details

A.1 Datasets

In this section, we provide detailed descriptions of the benchmark datasets used in our experiments.

CEBaB (Abraham et al., 2022) is a commonly used dataset for concept-based text analysis. It includes restaurant reviews, annotated with sentiment ratings (ranging from 1 to 5 stars) and four dining experience concepts: food quality, noise level, ambiance, and service. Following Tan et al. (2024a), we frame this as a five-class classification task.

Beer (McAuley et al., 2012) is a multi-aspect beer reviews dataset. Each review includes sentiment ratings across five aspects: appearance, aroma, palate, taste, and overall impression. Following Jourdan et al. (2023), we train the model to predict whether the overall score exceeds 3, indicating a positive review. The remaining four aspects are treated as concept labels for supervised concept-based methods.

Hotel (Wang et al., 2010) is a multi-aspect hotel reviews dataset, comprising review texts annotated with seven concept labels: value, rooms, location, cleanliness, check-in/front desk, service, and business service. Since the original labels are on a scale of 0 to 5, we utilize the binarized version proposed by Bao et al. (2018).

IMDB (Maas et al., 2011) is a movie reviews dataset, where each review is labeled with either positive or negative sentiment.

AGnews (Gulli, 2004) is a dataset of news articles categorized into one of four classes: business, science/technology, sports, or world/political.

Twitter (Sheng et al., 2021) is a dataset consisting of real and fake news collected from Twitter.

SciCite (Cohan et al., 2019) is a dataset designed for classifying citation intents in academic papers. Each citation is categorized into one of three classes: method, background, or result.

A.2 Baselines

The details of the baseline methods are shown as follows.

BERT-/RoBERTa-based classifier (Devlin et al., 2019; Liu et al., 2019): We utilize BERT/RoBERTa to encode tokens of text content and feed the extracted average embedding into an MLP to obtain the final prediction.

CBM (Kim et al., 2018) introduced an intermediate layer that first predicts the human-defined concepts and then uses the concepts to predict the final output.

SparseCBM (Tan et al., 2024a) enhances CBM by integrating unstructured pruning. SparseCBM constructs concept-specific sparse subnetworks within the backbone network, providing interpretability while retaining model performance.

SelfExplain (Rajagopal et al., 2021) is an unsupervised self-explaining model. It provides both global and local concept explanations for each sample while performing the classification tasks.

PROTOTEX (Das et al., 2022) is an unsupervised self-explaining classification architecture based on prototype networks. It explains model decisions based on prototype tensors that encode latent clusters of training examples.

Cockatiel (Jourdan et al., 2023) is an unsupervised post-hoc concept-based method. It generates meaningful concepts from the last layer of a neural net model trained on an NLP classification task by using Non-Negative Matrix Factorization. In our experiment, we use a RoBERTa-based classifier as the base model for concept extraction.

Concept-Shap (Yeh et al., 2020) is an unsupervised post-hoc concept-based method that aims to infer a complete set of concepts. In our experiment, we use a RoBERTa-based classifier as the base model for concept extraction.

It is worth noting that we did not include approaches that utilize LLMs to extract concepts. These methods generate concept sets using LLMs before training and follow a pipeline similar to that of CBM. Specifically, in the text domain, their primary distinction from CBM lies in the use of concept sets defined by LLMs or humans. As such, we use CBM as a representative method.

A.3 Computational Complexity Analysis

Compared to the baseline methods, incorporating LLM-based evaluation during training increases computational costs. Specifically, our method requires approximately 22.2% more resources than the best-performing baseline, Cockatiel. However, in real-time or large-scale applications, our approach does not rely on querying LLMs during deployment. The inference phase only involves the local model, ensuring that deployment incurs no significant additional computational overhead. Moreover, our method demonstrates superior performance in interpretability compared to the best baseline, achieving enhanced comprehensibility (semantics, distinctiveness, consistency) of 25%. We believe the limited additional computational cost during training is a worthwhile trade-off for these significant benefits, especially as it does not impact real-world deployment efficiency.

B Human Evaluation of using LLMs as Human Proxies

We conducted several human studies to assess the quality of using LLMs for concept comprehensibility evaluation. For each task, we presented the concepts extracted using our method to human evaluators and asked them to rate their level of agreement with the conceptual summaries and segment highlightings generated by LLMs. For the ratings of agreement with conceptual summaries, we presented all the concepts obtained using our method with three top-activated examples. For the ratings of agreement with concept-related segment highlightings, we focused on the concepts that LLM identified as having semantic meanings. For each concept, we randomly selected three cases with corresponding LLM-generated highlightings. Each concept was evaluated three times by different workers using a 5-point scale (1 indicating complete disagreement and 5 indicating complete agreement). We conducted our evaluation on Prolific[2], a platform for facilitating high-quality human surveys. Each worker is paid $10.5 per hour and must pass a screening test to take the survey. A total of 42 participants took part in the LLM-generated conceptual summaries evaluation, while 21 participants took part in the LLM-generated segment highlightings evaluation. The demographic information of the evaluators, as provided by Prolific, is shown in Tables 11 and 12.

	CEBaB	Beer	Hotel	IMDB	AGnews	Twitter	SciCite
Summaries	4.23	4.21	4.45	4.56	4.63	4.39	4.14
Highlightings	4.56	4.24	4.40	4.62	4.42	4.33	4.18

Table 6: Ratings of agreement with the LLM-generated conceptual summaries and segment highlightings

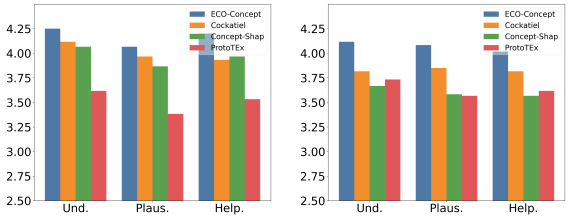
The rating results are shown in Table 6. Our human study, conducted with diverse participants, confirms that LLMs closely align with human judgment in concept evaluation without introducing much bias. This validates our method of employing LLMs as human proxies for concept evaluation and leveraging their feedback to refine the model.

C Human Evaluation of Concept Comprehensibility

C.1 Participants Information

Our human evaluation experiments are conducted on Prolific. Each worker is paid $10.5 per hour and must pass a screening test to take the survey. To ensure more reliable results, each question is evaluated three times by different workers. For the intruder detection and subjective rating experiments, we recruited 42 participants, while 24 participants were recruited for the forward simulatability experiments. The demographic information of the annotators is shown in Tables 11 and 13.

[2]https://app.prolific.com



(a) Beer (b) AGnews

Figure 5: Human ratings of provided explanations (Und. indicating understandability, Plaus. indicating plausibility, and Help. indicating helpfulness)

C.2 Intruder Detection Experiments Details

For the intruder detection experiments, we tested all extracted concepts of each method across seven tasks. Participants were asked to identify the example that did not belong to the same concept from a set of four highlighted examples. Their success rate in identifying the intruder was then calculated. The interface of the task is shown in Fig. 9(a).

C.3 Subjective Rating Experiments Details

For the subjective rating experiments, we present three top-activated examples for each concept along with its summary and ask participants to rate these examples based on the following criteria using a 1-5 scale:

Consistency: Do you think this concept formed by all these highlighted text parts has consistent semantic meanings?

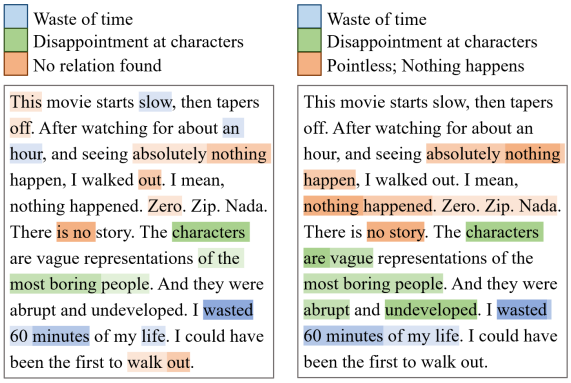
Clarity: Do you think the semantic meaning of this concept is clear and easy to identify?

Task Relevance: Do you think this concept is related to the task?

Comprehensibility: Do you think this concept is comprehensible and easy for humans to understand?

C.4 Forward Simulatability Experiments Details

For the forward simulatability experiments, we randomly selected 20 samples from each of the Beer and AGnews datasets, ensuring that the accuracy of all methods within these samples was between 80% and 90%. First, participants classified the samples without any model explanations. Then, one method's explanation was provided to the participants, who used this to predict the model's output and recorded their confidence. To facilitate understanding by non-experts, we normalized the contribution of each concept to the class comparison, as



(a) Concepts explanations provided by the Base model (b) Concepts explanations provided by ECO-Concept

Figure 6: Comparison of explanations before and after concept enhancement in a beer review case

illustrated in Fig. 9(c). Additionally, participants were asked to rate the understandability, plausibility, and helpfulness of the provided explanation, as Fig. 9(d) shows. The results of the human ratings are presented in Fig. 5.

D Further Analysis on Ablation Study

D.1 Further Analysis on Impact of Concept Regularizers

From Table 5, we observed that the decrease in the *Distinctiveness* score when removing $\mathcal{L}_{con}$ (from w/ all to w/o $\mathcal{L}_{con}$ variant) is greater than the decrease when removing $\mathcal{L}_{dist}$ (from w/ all to w/o $\mathcal{L}_{dist}$ variant). This occurs in the tasks (IMDB, AGnews, Twitter) where the *Semantics* score of w/o $\mathcal{L}_{con}$ variant is consistently lower than that of w/o $\mathcal{L}_{dist}$ variant. This indicates that removing the consistency regularizer in these tasks leads to fewer discovered concepts with clear semantic meaning, which in turn affects the calculation of concept distinctiveness. This further supports the rationale for using both regularizers simultaneously. The two regularizers mutually reinforce each other, and their combined use is crucial for discovering more comprehensible concepts.

D.2 Further Analysis on Impact of Concept Comprehensibility Enhancement

The ECO-Concept (w/o $\mathcal{L}_{com}$) variant is fine-tuned using the same parameter settings and number of iterations as ECO-Concept training, but without the comprehensibility loss. As shown in Fig. 4, for concept comprehensibility metrics, fine-tuning with the comprehensibility loss performs better than fine-tuning without this loss. Moreover, in

certain tasks, solely relying on concept regularizers for retraining can even result in a decline in the semantics, distinctiveness, and consistency of the concepts. The possible explanation for this phenomenon is that concept regularizers guide concept learning within the tensor space. During the early stages of training, they enhance the consistency of representations for the same concept while improving the distinctiveness between different concepts. As training progresses, most concepts gradually acquire certain semantic meanings. However, in the absence of explicit concept annotation guidance, semantic ambiguity becomes unavoidable. This semantic ambiguity is often difficult to resolve in high-dimensional tensor spaces. For semantically ambiguous concepts, some samples may exhibit inconsistencies in semantics, while their features remain relatively close within the tensor space. In such cases, continuing to rely solely on concept regularizers not only fails to resolve the ambiguity but may exacerbate it, thereby compromising the performance of concept metrics. To address this, we propose an enhancement method inspired by the human evaluation process. By incorporating an evaluation mechanism that aligns more closely with human cognitive dimensions, we can reduce the semantic ambiguity of concepts and enhance their comprehensibility. The above analysis demonstrates the significance and effectiveness of our proposed comprehensibility enhancement stage.

A case in Fig. 6 presents that after concept enhancement, the model not only achieves higher semantic consistency but also clarifies the meaning of concepts that were previously semantically ambiguous.

E Parameter Sensitivity Analysis

E.1 Impact of the Number of Concepts

To evaluate the impact of the number of concepts on task performance and concept-related metrics, we conducted experiments with 10, 20, 30, 40, and 50 concepts, respectively. The results, as shown in Fig. 7, indicate that when the number of concepts is 20, most metrics achieve their optimal values. All variants achieve similar results across different tasks, and the number of concepts has minimal effect. For concept interpretability, when the number of concepts exceeds 20, both the concept semantics and distinctiveness decline. This suggests that an excessive number of concepts may bring more semantically irrelevant or redundant concepts. Furthermore, regardless of the number of concepts, it is impossible to ensure that all extracted concepts have semantic meanings due to the unsupervised nature of our method. Nevertheless, when the number of concepts is set to 20, over half of the extracted concepts can be confidently considered semantically meaningful.

Parameter	Search Range	Optimal Value
λ_{con}	0.01 / 0.05 / 0.1 / 0.3 / 0.5	0.1
λ_{dist}	-0.001 / -0.005 / -0.01 / -0.05 / -0.1	-0.01
λ_{com}	0.1 / 0.5 / 1.0 / 1.5 / 2.0	1.0

Table 7: Details of the grid search range for optimal trade-off parameters

	CEBaB	Beer	Hotel	IMDB	AGnews	Twitter	SciCite
K=3	.895	.925	.934	.947	.897	.916	.919
K=5	.817	.826	.803	.821	.802	.844	.870

Table 8: Rank correlation scores calculated based on the ranks of the top K attributed concepts before and after noisy perturbations

E.2 Grid Search Details

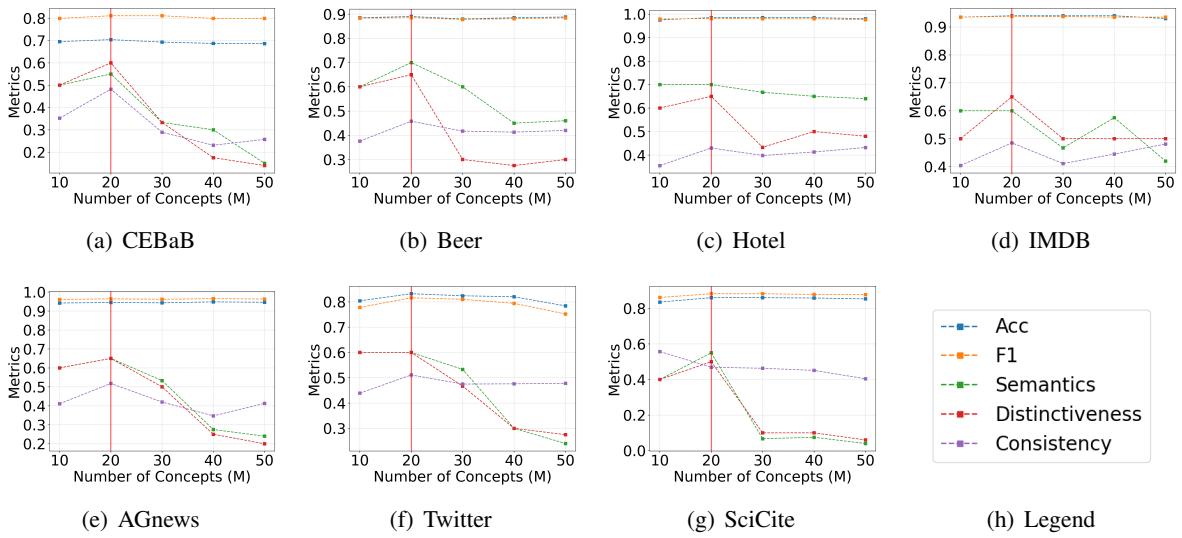
The trade-off parameters λ_{con}, λ_{dist}, and λ_{com} were selected through grid search, aiming to optimize the model's classification performance on the validation set. The detailed information about the grid search range is presented in Table 7.

Our grid-search results indicated that other parameter combinations had lower performance compared to the optimal combination. Table 9 shows the classification performances of the top 4 performing parameter combinations across all tasks.

F Model Robustness Evaluation

To test whether our method can yield reliable explanations for noisy samples, we performed experiments using seven datasets. We randomly chose 100 test samples and applied noisy perturbations to 5% of the words in each sample. These words were chosen based on their frequency in the training set (excluding stop words), with higher-frequency words being more likely to be perturbed. For each sample, we randomly applied one of these adversarial strategies: synonym replacement or spelling error. For synonym replacement, we selected a synonym from WordNet (Miller, 1995), prioritizing the one with the lowest frequency in the training set. For spelling errors, we randomly altered a single character in the word.

To evaluate the robustness of our model's expla-



(a) CEBaB (b) Beer (c) Hotel (d) IMDB

(e) AGnews (f) Twitter (g) SciCite (h) Legend

Figure 7: Classification performance and concept metrics of our method with different concept numbers.

Parameter Combination	CEBaB		Beer		Hotel		IMDB		AGnews		Twitter		SciCite	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
(0.1, -0.01, 1)	**.697**	**.808**	**.885**	**.885**	**.981**	**.981**	**.937**	**.937**	**.941**	**.961**	**.828**	**.813**	**.860**	**.881**
(0.05, -0.01, 1)	.677	.793	.882	.882	.977	.977	.934	.934	.927	.939	.824	.801	.856	.879
(0.1, -0.01, 0.5)	.683	.799	.883	.883	.979	.979	.931	.931	.936	.951	.826	.804	.856	.879
(0.05, -0.01, 0.5)	.686	.799	.883	.883	.977	.977	.929	.929	.929	.948	.823	.799	.858	.880

Table 9: Classification performances of the top 4 performing parameter combinations (λ_{con}, λ_{dist}, and λ_{com}) across all tasks. The best results are bolded.

nations, we compared the top concepts attributed to the same sample before and after the noisy perturbations. We used Kendall's rank correlation (Kendall, 1938) as the evaluation metric, which accounts for the concept importance rankings. The rank correlation scores, calculated based on the ranks of the top K attributed concepts before and after noisy perturbations, are shown in Table 8. After noisy perturbations, our model is still able to attribute the test samples to concept explanations that closely match those before the attack. This demonstrates the robustness of the explanations provided by our method under noisy conditions, highlighting its potential for practical applications.

G Prompts

The prompts we used are shown in Table 10.

H Comparison of Explanations

The comparison of explanations of different concept-based methods is shown in Fig. 8.

Prompt for Concept Summary
Prompt: We're studying the concepts used to determine whether the sentiment of movie reviews is positive or negative. Each concept focuses on some specific elements. First, pay attention to all highly activated text parts in the following example sentences. Then think deeply to find the relations between all the highly activated text parts. Finally, determine whether all these highly activated text parts can represent a consistent concept. If not, please output 'No relation found'; if yes, provide a summary in no more than ten words. The activation format is token<tab>activation. Activation values range from 0 to 1. A concept finding what it's looking for is represented by a non-zero activation value. The higher the activation value, the stronger the match.
Prompt for Concept-related segment highlightings
Prompt: We're studying the concepts used to determine whether the sentiment of movie reviews is positive or negative. Each concept focuses on specific elements in a short document. Look at the summary of what the concept is, and try to determine whether each token in the document is related to the concept. Return 1 for related tokens and 0 for unrelated tokens. You should carefully consider whether each token is related to the concept and not return 1 for all the tokens in the document.

Table 10: Prompt for concept summary and concept-related segment highlightings in the movie review sentiment classification task

Highest education level completed				Avg. Age
High School	Undergraduate	Graduate	Doctorate	
5	21	11	5	36.4
Ethnicity				Female:Male
White	Black	Asian	Mixed	
21	12	5	4	1:1.62

Table 11: Demographic information of participants in LLM-generated conceptual summaries evaluation, intruder detection experiments, and subjective rating experiments

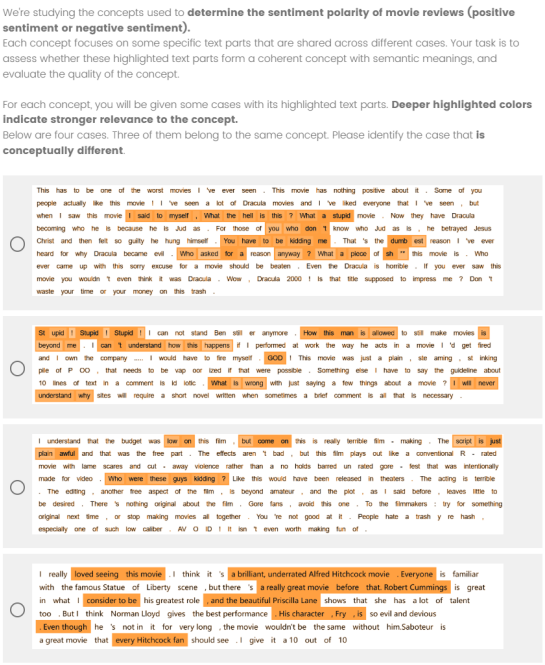
Highest education level completed				Avg. Age
High School	Undergraduate	Graduate	Doctorate	
3	10	7	1	29.1
Ethnicity				Female:Male
White	Black	Asian	Mixed	
8	6	5	2	1:1.33

Table 12: Demographic information of participants in LLM-generated concept-related highlightings evaluation

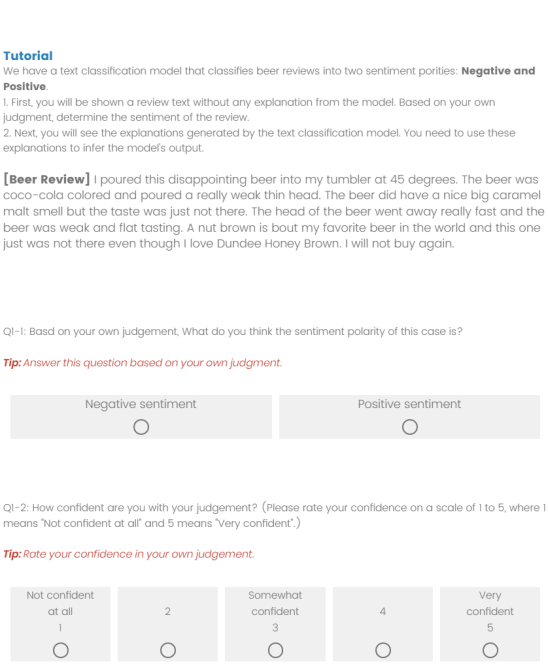
Highest education level completed				Avg. Age
High School	Undergraduate	Graduate	Doctorate	
3	11	8	2	40.2
Ethnicity				Female:Male
White	Black	Asian	Mixed	
11	6	4	3	1:1.18

Table 13: Demographic information of participants in forward simulatability experiments

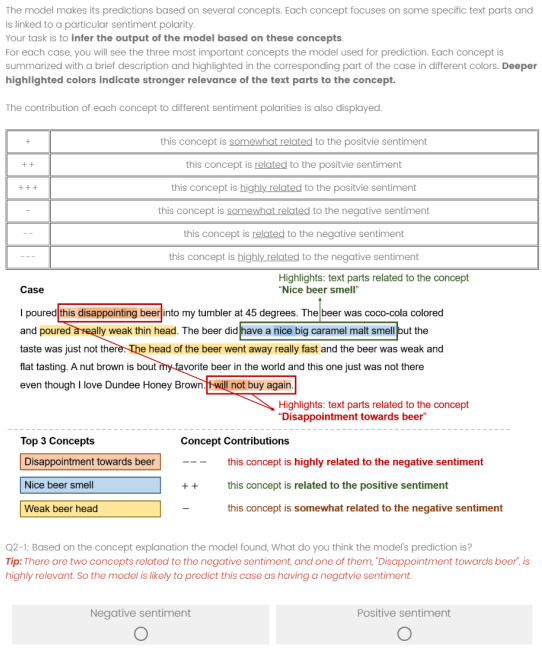
Figure 8: Comparison of concept explanations provided by ECO-Concept and other concept-based methods



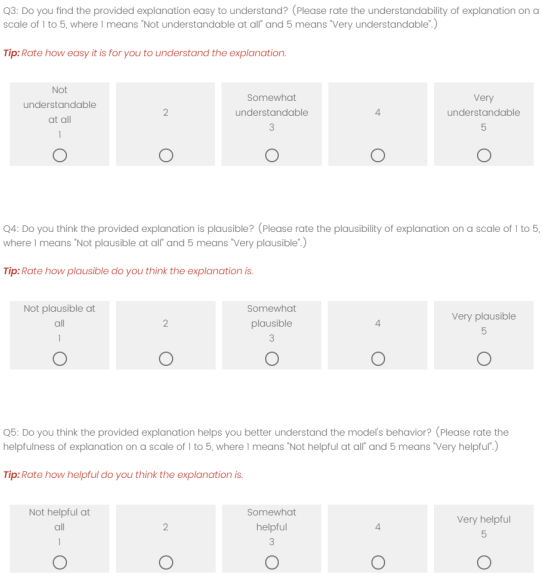
(a) The survey interface of intruder detection



(b) The tutorial of forward simulatability part (1)



(c) The tutorial of forward simulatability part (2)



(d) The tutorial of forward simulatability part (3)

Figure 9: Screenshots of the survey interface