# Explaining Language Model Predictions with High-Impact Concepts

**Ruochen Zhao**[1] **Tan Wang**[1] **Yongjie Wang**[1] **Shafiq Joty**[1,2]
[1]Nanyang Technological University, Singapore
[2]Salesforce AI
{ruochen002, tan317, yongjie002}@e.ntu.edu.sg
srjoty@ntu.edu.sg

## Abstract

To encourage fairness and transparency, there exists an urgent demand for deriving reliable explanations for large language models (LLMs). One promising solution is concept-based explanations, *i.e.* human-understandable concepts from internal representations. However, due to the compositional nature of languages, current methods mostly discover *correlational* explanations instead of *causal* features. Therefore, we propose a novel framework to provide impact-aware explanations for users to understand the LLM's behavior, which are robust to feature changes and influential to the model's predictions. Specifically, we extract predictive high-level features (concepts) from the model's hidden layer activations. Then, we innovatively optimize for features whose existence causes the output predictions to change substantially. Extensive experiments on real and synthetic tasks demonstrate that our method achieves superior results on predictive impact, explainability, and faithfulness compared to the baselines, especially for LLMs.

## 1 Introduction

Over the past few years, large language models (LLMs) have achieved tremendous progress, leading them to be widely applied in sensitive applications such as personalized recommendation bots and recruitment. However, Explainable AI (XAI) has not witnessed the same progress, making it difficult to understand LLMs' opaque decision processes (Mathews, 2019). Therefore, many users are still reluctant to adopt LLMs in high-stake applications due to transparency and privacy concerns. In this work, we aim to increase user trust and encourage transparency by deriving explanations that allow humans to better predict the model outcomes.

To understand what happens inside an LLM, previous studies (Dalvi et al., 2021) show that dense vector representations in high layers of a language model tend to capture semantic meanings that are useful for solving the underlying task. However, such vector representations are not understandable to humans. To solve it, concept-based explanations attempt to map the hidden activation space to human-understandable features. For example, Koh et al. (2020) provides the concept bottleneck model, which first predicts an intermediate set of human-specific concepts, then uses them to predict the target. As illustrated by purple boxes in Fig. 1, for the movie review classification task, concept-based explanations are semantically meaningful word clusters (Dalvi et al., 2021) corresponding to abstract features such as "acting" and "directing".

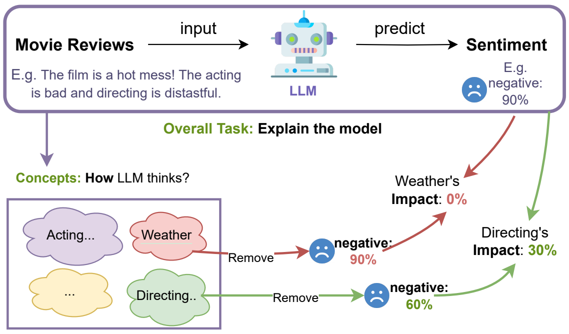


Figure 1: Illustration of concept-based explanations that result in high impact (green line) or not (red line) when explaining the LLMs in a sentiment classification task.

However, existing concept-based methods do not consider of the *explanation impact* on output predictions, leading to inferior explanations. By *impact*, we mean the causal effect of removing a feature on output predictions (Goyal et al., 2019; Abraham et al., 2022). As Moraffah et al. (2020) points out, these non-impact-aware methods derive correlational explanations that cannot answer questions about decision-making under alternative situations and are thus unreliable. An example is illustrated in Fig. 1. Due to the conventional expression "hot mess", the word "hot" often co-occurs with "mess", which is usually used to classify nega-

*Findings of the Association for Computational Linguistics: EACL 2024*, pages 995–1012
March 17-22, 2024

tive sentiment. Traditional concept-based methods that do not consider impact may falsely use the correlational feature "weather" (*i.e.*, "hot") to explain why the model classifies something as negative. However, excluding the "weather" concept does not cause the output prediction to change at all, resulting in zero impact (red line). Thus, low-impact explanations such as "weather" are less valid as users cannot utilize them to consistently predict the model's behaviors when a feature changes.

To tackle this bottleneck and incorporate impact into traditional concept-based models, in this work, we propose High-Impact Concepts (*HI-concept*), a complete concept explanation framework with causal impact optimization (§3.2). Specifically, We design a *causal* loss objective, stemming from the treatment effects in the causality literature (Pearl, 2009). Moreover, previous causality evaluations (Goyal et al., 2019; Feder et al., 2021b) primarily focused on assessing the causal effect via *local* (*i.e.*, instance-level) change and *removal* intervention (*i.e.*, eliminating words/concepts from the source), leading to potentially biased evaluation results. To this end, we further propose a novel *global* (*i.e.*, model-level) accuracy change metric and *insertion* operation to effectively diagnose the causality measurement (§3.4).

As a result, our method can consistently prioritize more influential features (green line in Fig. 1) while disregarding correlational ones. Extensive experiments with multiple language models, both established and newly proposed evaluation metrics, and rigorous human studies fully validate the effectiveness of *HI-concept* in finding high-impact concepts compared to baselines, especially for LLMs. Our contributions are summarized as follows[1]:

- To alleviate the problem of correlational explanations, we propose *HI-concept*, a framework for deriving explanatory features with high impacts by innovatively optimizing a causal objective.
- Towards comprehensive evaluations, we propose a theoretically grounded metric, namely reconstruction accuracy change, and devise an insertion study, which serves as a complement to the traditional removal intervention.
- Extensive experiments show that *HI-concept* is impactful, explainable, and faithful, with especially outstanding improvements on LLMs (*e.g.*, improving the causal effect on accuracy from **2.83**% to **27.79**% on Llama-7B).

## 2 Preliminaries

We first introduce what concept-based explanations are, what properties they should satisfy, and our key baseline, concept bottleneck models.

### 2.1 Concept-based Explanations

Concept-based explanations is a well-established method (Kim et al., 2018; Koh et al., 2020; Yeh et al., 2020) that extracts human-understandable concepts from the model's hidden space. As stated in Kim et al. (2018), the activation space of an ML model can be seen as a vector space $E_m$ spanned by basis vectors $e_m$ which correspond to input features. Humans work in a different vector space $E_h$ spanned by implicit vectors $e_h$ corresponding to an unknown set of human-understandable concepts. Then, concept-based explanations $g : \mathbb{E}_m \rightarrow E_h$ aim to translate from high-level representations into task-relevant and human-understandable concepts.

Ideally, concept-based explanations should satisfy the following properties (Doshi-Velez and Kim, 2017). *Faithfulness:* The explanations can be able to accurately mimic the original model's prediction process (Ribeiro et al., 2016). *Causality:* When the feature is perturbed in real life, the output predictions should change accordingly. This causal impact ensures that explanations are reliable under alternative situations. *Explainability:* The explanations should be understandable to humans and able to assist users in real-life tasks. These three properties will be the guiding principles for our model design and evaluation.

### 2.2 Concept Bottleneck Models

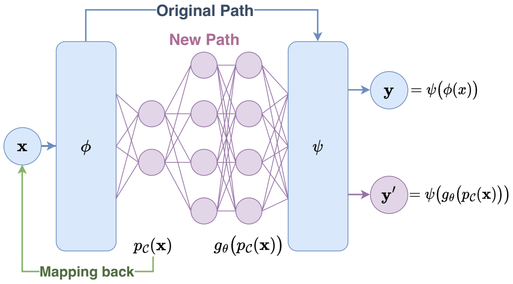

Original Path
New Path
$\mathbf{x}$
$\phi$
$\psi$
$\mathbf{y}$ $= \psi(\phi(x))$
$\mathbf{y}'$ $= \psi(g_\theta(p_C(\mathbf{x})))$
$p_C(\mathbf{x})$
$g_\theta(p_C(\mathbf{x}))$
Mapping back

Figure 2: The overall concept generation process of a concept bottleneck model.

To derive concept-based explanations, one classic architecture is concept bottleneck models (Yeh et al., 2020), shown in Fig. 2. The pretrained

[1] Our codebase is available at https://github.com/RuochenZhao/HIConcept.

model $f$ can be viewed as a composite of two functions, divided at an intermediate layer: $f = \psi \circ \phi$. After initializing the concepts $\mathcal{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_n\} \in \phi(\cdot)$ uniformly, $\phi(\mathbf{x})$ is encoded into concept probabilities $p_{\mathcal{C}}(\mathbf{x})$, calculated as $p_c^i(\mathbf{x}) = \text{TH}((\phi(\mathbf{x})^\top \mathbf{c}_i), \beta)$[2] Then, the bottleneck-shaped network reconstructs $\phi(\mathbf{x})$ with a 2-layer perceptron $g_\theta$ such that $g_\theta(p_{\mathcal{C}}(\mathbf{x})) \approx \phi(\mathbf{x})$. Intuitively, hidden space $\phi(\cdot)$ corresponds to the vector space $E_m$. The concept probability space $p_{\mathcal{C}}(\cdot)$ corresponds to the human-understandable space $E_h$. To train the concept model in an end-to-end way, two losses are used:

• *Reconstruction loss:* To faithfully recover the original model's predictions, a surrogate loss with cross-entropy (CE) is optimized[3]:

$$\begin{aligned}\mathcal{L}_{\text{rec}}(\theta, \mathcal{C}) &= \text{CE}\Big(\psi(\phi(\mathbf{x})), \psi(g_\theta(p_{\mathcal{C}}(\mathbf{x})))\Big) \\ &= -\sum_{b \in \mathcal{B}} \psi(\phi(\mathbf{x}))_b \log\left(\psi(g_\theta(p_{\mathcal{C}}(\mathbf{x})))_b\right).\end{aligned} \tag{1}$$

• *Regularization loss:* To make concepts more explainable, a regularization loss forces each concept vector to correspond to actual examples and concepts to be distinct from each other[4]:

$$\begin{aligned}\mathcal{L}_{\text{reg}}(\mathcal{C}) = &- \lambda_1 \frac{\sum_{i=1}^{n} \sum_{\mathbf{x}_t \in \mathcal{T}_{\mathbf{c}_i}} \mathbf{c}_i^\top \phi(\mathbf{x}_t)}{nN} \\ &+ \lambda_2 \frac{\sum_{i_1 \neq i_2} \mathbf{c}_{i_1}^\top \mathbf{c}_{i_2}}{n(n-1)}.\end{aligned} \tag{2}$$

## 3 Methodology

Then, we propose *HI-concept*, which aims to fill the current research gap on explanatory impact.

### 3.1 Defining Impact

As stated earlier, not considering impact could result in confounding and correlational explanations. The failure cases can be theoretically explained by causality analysis in Fig. 3. To achieve sentiment prediction $Y$ , the hidden activation space in pre-trained LLMs consists of both correlated features $E$ and predictive features $Z$. Although only $Z$ truly affects prediction $Y$, $E$ and $Z$ may be correlated due to the confounding effects brought by input $X$. However, a traditional concept mining model does not differentiate between $E$ and $Z$ and considers both as valid. Thus, it may easily use the confounding association as an explanation instead of the true causal path. The resulting concepts would be problematic as they do not facilitate a robust understanding of the model's behaviors.

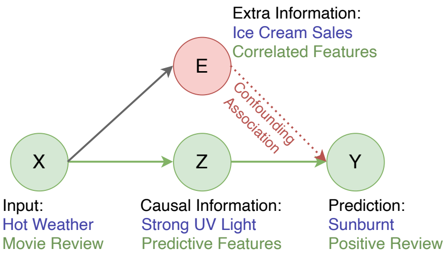


Figure 3: Illustration of the causal graph indicating the confounding association in explanation models. Blue is a real-life example. Green is the correspondence in a movie review classification task.

To tackle this challenge, we enforce explanations to be predictive by considering their "impact". To formally define the *impact* of a feature, we utilize two important definitions in causal analysis: Individual Treatment Effect (ITE) and Average Treatment Effect (ATE), which measure the effect of interventions in randomized experiments (Pearl, 2009). Given a binary treatment variable $T$ that indicates whether a *do-operation* is performed (*i.e.*, perturb a feature), ATE and ITE are defined as the change in expected outcome with treatment $T = 1$:

$$\begin{aligned}\text{ITE}(x) :=\; &\mathbb{E}[y|\mathbf{X} = x, \text{do}(T = 1)] \\ &- \mathbb{E}[y|\mathbf{X} = x, \text{do}(T = 0)]; \\ \text{ATE} :=\; &\mathbb{E}[\text{ITE}(x)].\end{aligned} \tag{3}$$

In our case, a concept $\mathbf{c}_i$ is discovered as a direction in the latent space, corresponding to a feature in the input distribution. As $f$ is fixed, its prediction process is deemed deterministic and reproducible, allowing us to conduct experiments with treatments (Koh et al., 2020). Therefore, we propose to remove a specific concept (Goyal et al., 2019)[5] as the do-operation and define *impact* $I$ of a concept $\mathbf{c}_i$ on an instance $(\mathbf{x}, y)$ as:

$$\text{I}(\mathbf{c}_i, \mathbf{x}) = \mathbb{E}[y|\mathbf{X} = \mathbf{x}, \mathbf{c}_i = \mathbf{0}] - \mathbb{E}[y|\mathbf{X} = \mathbf{x}, \mathbf{c}_i = \mathbf{c}_i]. \tag{4}$$

### 3.2 Optimizing for Impact

In order to incorporate consideration for impact into the concept discovery process, we introduce two new losses to the original framework:

• *Auto-encoding loss:* To guarantee that the intervened representations are still meaningful, we

[2] TH is a threshold function that forces all inputs smaller than $\beta$ to be 0.

[3] $\mathcal{B}$ is the set of class labels and $\psi(.)_b$ denotes the prediction score corresponding to label $b$.

[4] $\mathcal{T}_{\mathbf{c}_i}$ as the set of top-k neighbors of $\mathbf{c}_i$

[5] We assume that, as the concept vectors coexist in the hidden embedding space, there is no causal relationship among the concepts $\{\mathbf{c}_1, \ldots, \mathbf{c}_n\}$ themselves.

optimize an auto-encoding loss to learn a proxy task that reconstructs the hidden representations. With this loss, the concept model becomes Auto-encoder-like and can mimic a generation process of the real distribution of $\phi(\mathbf{x})$. Therefore, concept vectors can then be seen as key factors in the generation process of $\phi(\mathbf{x})$. Then, we can perform valid interventions on the concept vectors, such as the removal intervention. Formally:

$$\begin{aligned}\mathcal{L}_{\text{enc}}(\theta, \mathcal{C}) &= \text{MSE}\Big(\phi(\mathbf{x}), g_\theta(p_\mathcal{C}(\mathbf{x}))\Big) \\ &= \frac{1}{d}||\phi(\mathbf{x}) - g_\theta(p_\mathcal{C}(\mathbf{x})||_2^2.\end{aligned} \tag{5}$$

• *Causality loss:* Directly optimizing for causality is a challenging objective as causal impact is difficult to estimate during training. Therefore, we approximate impact (Eq. (4)) by randomly removing a set of concepts $\mathcal{S} \subseteq \mathcal{C}$ and calculating the expectation of impact on the training set. Then, we could disentangle concept directions that have a greater impact by optimizing the following loss:

$$\begin{aligned}\mathcal{L}_{\text{cau}}(\theta, \mathcal{C}) = -\sum_{\mathbf{c}_i \in \mathcal{S}} \sum_{\mathbf{x}_j \in \mathcal{D}} \Big| \psi\Big(g_\theta(p_\mathcal{C}(\mathbf{x}_j)|\mathbf{c}_i = \mathbf{0})\Big) \\ - \psi\Big(g_\theta(p_\mathcal{C}(\mathbf{x}_j)|\mathbf{c}_i = \mathbf{c}_i)\Big)\Big| \approx -|I_{\text{avg}}(\mathcal{C})|.\end{aligned} \tag{6}$$

As all inputs $\mathbf{x}_j \in \mathcal{D}$ are perturbed, the training dataset $\mathcal{D}$ serves both as the treatment group and the nontreatment group, ensuring no divergence.

Finally, the overall loss function becomes:

$$\begin{aligned}\mathcal{L}(\theta, \mathcal{C}) =& \mathcal{L}_{\text{rec}}(\theta, \mathcal{C}) + \mathcal{L}_{\text{reg}}(\mathcal{C}) \\ &+ \lambda_e \mathcal{L}_{\text{enc}}(\theta, \mathcal{C}) + \lambda_c \mathcal{L}_{\text{cau}}(\theta, \mathcal{C}),\end{aligned} \tag{7}$$

where $\lambda_e$, $\lambda_c$ are the weights for the auto-encoding loss and the causal loss respectively. In practice, the hyperparameters require minimal tuning. Specifically, we recommend fixing $\lambda_1 = 0.1$ and $\lambda_2 = 0.5$ for regularizer loss in Eq. (2), and $\lambda_e = 1$ for reconstruction loss. The only hyperparameter to tune is $\lambda_c$, whose optimal level can be found within a few steps. Futher details on implementation and the training process could be found in Appendix A.

### 3.3 Visualizing Concepts via Impact

As a concept $c_i \in \phi(\cdot)$ is a hidden space vector, previous concept discovery methods face difficulties in mapping concept vectors to semantic meanings. They mainly relied on naively clustering the high-frequency words (Dalvi et al., 2021; Yeh et al., 2020). To address this issue, we use established visualization techniques to translate it to human-understandable concepts (*i.e.*, word clusters and highlights).

For models where the hidden representation is token-level, we simply use the individual token's concept probability $p_\mathcal{C}(x_i)$ as token importance scores. For models with sequence-level representations such as BERT, we employ the well-established transformer visualization method proposed in Chefer et al. (2021) to map back from the [CLS] activation concepts to input tokens. As an adaption of Grad-CAM (Selvaraju et al., 2017) to transformers, it visualizes classifications with layer-wise propagation, gradient backpropagation, and layer aggregation with rollout. As a result, for each sample $\mathbf{x}$ with tokens $x_1, \ldots, x_T$, we go from having only one concept similarity score $p_c^i(\mathbf{x})$ to a list of normalized token importance scores $s_1(\mathbf{c}_i), \ldots, s_T(\mathbf{c}_i)$. Therefore, we derive both global/model-level concepts (*i.e.*, word clusters) and their corresponding local/instance-level explanations (*i.e.*, token importance scores for an instance) that result in high impact. Both forms of generated explanations can complement each other while conforming to the model's 'mindset'.

### 3.4 Evaluating Impact of Concepts

Quantitatively, traditional causality evaluation metrics focus on local (*i.e.*, instance-level) perturbations (Feder et al., 2021b), which may be biased to global (*i.e.*, model-level) performance evaluations. Thus, we innovatively propose *Recovering Accuracy Change ($\Delta$Acc)*. Following the causality definition Doshi-Velez and Kim (2017) and human intuition, if a concept $\mathbf{c}_i$ is a crucial factor used by the model to make predictions, omitting it should disrupt the ability to faithfully recover predictions. Formally, it is defined as:

$$\Delta\text{Acc}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{\mathbf{c}_i \in \mathcal{C}} |\text{Acc}(\mathcal{C}) - \text{Acc}(\mathcal{C} \setminus \{\mathbf{c}_i\})|,$$

where Acc denotes the recovering accuracy (Yeh et al., 2020).

Moreover, we follow previous work to use *Causal Concept Effect* (CACE) (Goyal et al., 2019) to evaluate the causal effect of the set of concepts $\mathcal{C}$. Formally, it is defined as:

$$\begin{aligned}\text{CACE}(\mathbf{c}_i) := \sum_{\mathbf{x}_j \in \mathcal{D}_{\text{test}}} &|\psi(g_\theta(p_\mathcal{C}(\mathbf{x}_j))) \\ &- \psi(g_\theta(p_{\mathcal{C}\setminus\{i\}}(\mathbf{x}_j)))|; \\ \text{CACE}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} &\sum_{\mathbf{c}_i \in \mathcal{C}} \text{CACE}(\mathbf{c}_i)\end{aligned}$$

Qualitatively, existing evaluations mostly assess concepts' impact $\mathcal{C}$ via feature *removal* (Goyal et al., 2019). We argue that obtained concepts should also be generalizable to cases of *insertion*. Thus, we propose a novel insertion operation. Intuitively, when inserting explanation features one by one, gradual improvement of recovering accuracy should be observed, indicating incremental impact of each concept.

## 4 Experiment Setup

### 4.1 Datasets and Metrics

We test the effectiveness of our method with two standard text classification datasets: IMDB (Maas et al., 2011) and AG-news (Zhang et al., 2015). IMDB consists of movie reviews labeled with positive or negative sentiments, while AG-news is a dataset of news articles categorized into 4 topics. Appendix B gives a dataset summary. We explain four classification models: (*i*) a 6-layer transformer encoder trained from scratch, (*ii*) a pre-trained BERT with finetuning, (*iii*) a pre-trained T5 model (Raffel et al., 2020) with finetuning, (*iv*) 7B Llama (Touvron et al., 2023) with in-context learning.

We evaluate the explanation methods quantitatively and qualitatively with comprehensive metrics based on the three important considerations described in §2.1. **Faithfulness.** To ensure that the surrogate model can accurately mimic the original model's prediction process, we evaluate whether the captured concept probabilities $p_{\mathcal{C}}(\mathbf{x})$ can recover the original model's predictions $\psi(\phi(\mathbf{x}))$ quantitatively with *Recovering Accuracy (Acc)* (Yeh et al., 2020), *Precision*, *Recall*, *F1*, and *Completeness* (Yeh et al., 2020). Please check the details of the metric calculation in Appendix C. **Causality** is the key of the XAI model evaluation. As mentioned in §3.4, we use the CACE metric (Goyal et al., 2019), a novel accuracy change metric (ΔAcc), and insertion operations to provide a more comprehensive overview. **Explainability.** With the concepts generating a high impact on predictions, we expect that it can allow end-users to better understand the model's decisions. We include visualizations and human studies to test it qualitatively.

### 4.2 Baselines and Hyperparameters

For baselines, we use other unsupervised dimension reduction methods to discover concepts on the hidden space: (*i*) PCA (F.R.S., 1901) and K-means (Likas et al., 2003) are popular non-parametric clustering techniques that reduce high-dimensional datasets into key features to increase interpretability. (*ii*) $\beta$-TCVAE (Chen et al., 2018) is a disentanglement VAE method that explicitly considers causal impact while reducing dimensionality. (*iii*) ConceptSHAP (Yeh et al., 2020) represents the traditional concept bottleneck models that do not consider impact.

The full list of hyperparameters used for training *HI-concept* can be found in Appendix B. Briefly, we use the causal coefficient $\lambda_c \in [1, 3]$, depending on the level of confounding within the dataset. During training, perturbation is performed on the most similar concept to the input. All experiments are conducted on the penultimate layer. The hyperparameters are chosen as an optimal default through grid search. To make the comparison fair, all methods use 10 dimensions to encode.

## 5 Results and Analysis

| $p_{\text{cor}}$ | Cls.Acc | Method | Acc | CACE | ΔAcc |
|---|---|---|---|---|---|
| 0.50 | 95.4% | ConceptSHAP | 97.6% | 0.070 | 6.1% |
| | | *HI-concept* | **98.4%** | **0.102** | **9.4%** (+3.3%) |
| 0.65 | 99.0% | ConceptSHAP | **99.7%** | 0.038 | 3.5% |
| | | *HI-concept* | 99.3% | **0.084** | **6.8%** (+3.4%)) |
| 0.75 | 96.1% | ConceptSHAP | 98.3% | 0.069 | 6.0% |
| | | *HI-concept* | **98.9%** | **0.123** | **12.2%** (+6.2%) |

Table 1: Faithfulness (Acc) and Causality (CACE, ΔAcc) evaluation on the toy dataset. Cls.Acc denotes the original classification model's accuracy.

### 5.1 Sanity Check

To first provide a sanity check for our method, we follow the toy experiment design in Yeh et al. (2020), which explains a CNN model trained on a synthetic graphic dataset. To mimic the confounding effects ($X \rightarrow E$) as in Fig. 3, we add correlations (controlled by $p_{cor}$) among ground truth concepts. Then, we compared discovered concepts by *HI-concept* with ConceptSHAP. Appendix D gives details of the experiment. In Table 1, results show that our method discovers concepts that consistently outperforms the baseline by deriving more impactful features. As confounding levels ($p_{\text{cor}}$) in the dataset increase, the performance gap (ΔAcc) also widens. Therefore, *HI-concept* successfully improves explanatory impact, especially for highly correlational tasks and datasets.

### 5.2 Quantitative Results on Text Classification

The experiment results on text classification datasets are presented in Table 2. Overall, HI-Concept not only achieves the best performance

| Dataset | Model | Method | Faithfulness | | | | | Causality | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Acc | Precision | Recall | F1 | Completeness | CACE | ΔAcc |
| IMDB | Transformer | $\beta$-TCVAE (Chen et al., 2018) | 43.53% | 50.23 | 50.03 | 33.08 | 27.36 | 0.037 | 1.50% |
| | | K-means (Likas et al., 2003) | 83.64% | 84.74 | 85.05 | 83.63 | 61.87 | 0.047 | 2.59% |
| | | PCA (F.R.S., 1901) | 85.18% | 85.56 | 86.20 | 85.15 | **62.36** | 0.001 | 0.01% |
| | | ConceptSHAP (Yeh et al., 2020) | 84.36% | 85.04 | 85.56 | 84.34 | 62.05 | 0.031 | 1.30% |
| | | *HI-concept* | **88.78%** | **90.07** | **87.50** | **88.24** | 58.10 | **0.150** | **11.06%** |
| | BERT | $\beta$-TCVAE (Chen et al., 2018) | 93.86% | 94.31 | 93.43 | 93.68 | 10.71 | 0.057 | 4.05% |
| | | K-means (Likas et al., 2003) | **98.69%** | 96.16 | 96.23 | 96.19 | 15.69 | 0.037 | 0.97% |
| | | PCA (F.R.S., 1901) | 96.68% | **96.65** | **96.68** | **96.67** | 15.33 | 0.002 | 0.02% |
| | | ConceptSHAP (Yeh et al., 2020) | 95.84% | 95.78 | 95.96 | 95.83 | 17.16 | 0.050 | 0.06% |
| | | *HI-concept* | 92.97% | 93.25 | 93.34 | 92.97 | **21.04** | **0.099** | **8.99%** |
| | Llama | $\beta$-TCVAE (Chen et al., 2018) | 20.56% | 33.41 | 33.36 | 13.30 | -14.29 | 0.001 | 0.15% |
| | | K-means (Likas et al., 2003) | 15.31% | 5.10 | 33.33 | 8.85 | -21.82 | 0.019 | 0.00% |
| | | PCA (F.R.S., 1901) | **95.15%** | **67.97** | **77.66** | **69.80** | **64.19** | 0.001 | 0.03% |
| | | ConceptSHAP (Yeh et al., 2020) | 18.83% | 42.83 | 34.95 | 14.88 | -1.78 | 0.005 | 1.60% |
| | | *HI-concept* | 87.87% | 53.27 | 68.60 | 55.29 | 59.83 | **0.042** | **28.69%** |
| AG-News | Transformer | $\beta$-TCVAE (Chen et al., 2018) | 98.91% | 98.94 | 98.94 | 98.93 | **66.73** | 0.049 | 6.62% |
| | | K-means (Likas et al., 2003) | 98.16% | 98.32 | 98.11 | 98.18 | 65.99 | 0.044 | 0.07% |
| | | PCA (F.R.S., 1901) | **99.99%** | **99.99** | **99.99** | **99.99** | 66.66 | 0.000 | 0.03% |
| | | ConceptSHAP (Yeh et al., 2020) | 73.01% | 59.36 | 74.34 | 64.88 | 47.07 | 0.000 | 0.00% |
| | | *HI-concept* | 99.50% | 99.50 | 99.51 | 99.50 | 66.70 | **0.046** | **7.12%** |
| | BERT | $\beta$-TCVAE (Chen et al., 2018) | 92.30% | 94.93 | 91.89 | 92.91 | 57.25 | 0.044 | 5.32% |
| | | K-means (Likas et al., 2003) | 86.83% | 92.74 | 85.42 | 87.53 | 52.62 | 0.028 | 7.15% |
| | | PCA (F.R.S., 1901) | 99.79% | 99.82 | 99.77 | 99.79 | 61.04 | 0.001 | 0.01% |
| | | ConceptSHAP (Yeh et al., 2020) | 93.46% | 93.70 | 94.62 | 93.66 | **62.69** | 0.025 | 4.44% |
| | | *HI-concept* | **99.90%** | **99.89** | **99.90** | **99.89** | 61.12 | **0.058** | **10.54%** |
| | Llama | $\beta$-TCVAE (Chen et al., 2018) | 1.27% | 0.25 | 20.00 | 0.50 | -23.89 | 0.000 | 0.01% |
| | | K-means (Likas et al., 2003) | 37.00% | 7.40 | 20.00 | 10.80 | 1.09 | 0.007 | 0.02% |
| | | PCA (F.R.S., 1901) | **85.41%** | **65.78** | **67.98** | **66.73** | **51.46** | 0.000 | 0.03% |
| | | ConceptSHAP (Yeh et al., 2020) | 17.01% | 35.37 | 35.20 | 15.87 | -7.73 | 0.002 | 2.83% |
| | | *HI-concept* | 81.52% | 48.59 | 55.99 | 51.53 | 43.07 | **0.039** | **27.79%** |

Table 2: Faithfulness (Acc, Precision, Recall, F1, Completeness) and causality (CACE, ΔAcc) evaluation of different text classification methods. The best result is bolded, and the second-best result is underlined.

| Method | CACE | Keywords |
|---|---|---|
| CS | 0.134 | apple, NASA, Microsoft, new, sun, red, super, game |
| CS | 0.000 | one, two, gt, new, cl, lt, first, world, mo, last |
| HI-C | 0.130 | us, bush, u, eu, new, peoples, china, high, gt, world |
| HI-C | 0.003 | us, update, new, mo, two, first, knicks, last, one, hen |

Table 3: Generated concepts with Average Impact (CACE) from AG-News dataset, BERT model. CS is ConceptSHAP, HI-C is *HI-concept*. Each line is one concept, represented by keywords, which are ordered by descending importance.

in causality, but improves on faithfulness as well. For faithfulness metrics (Acc, Precision, Recall, F1, and Completeness), *HI-concept* achieves the best or second-best results for almost all datasets and models. Notably, for the cases achieving second-best performance, the best model for faithfulness is PCA. PCA, however, as a completely different group of methods, is often faced with the issue of low causal impact (shows CACE close to 0 in Table 2). While considering causality metrics (CACE and ΔAcc), our *HI-concept* exhibits a significantly greater superiority. Causality metrics for baseline methods are mostly minimal, which implies that most explanatory features discovered are correlational and unreliable. In comparison, concepts discovered by *HI-concept* show significant improvements in both causality and faithfulness, especially for pretrained models such as BERT, Llama, and T5, whose results are shown in appendix E. This validates the hypothesis that HI-Concept can result in more improvements for larger pre-trained models with more complex architectures. With more parameters and pretraining, these models could encode more correlational information and contain more spurious correlations. As shown with the toy example in §5.1, HI-Concept's causality awareness would be more beneficial in highly correlational scenarios.

### 5.3 Qualitative Analysis of Text Classification

We take a closer look at BERT for AG-News to qualitatively examine the discovered concepts in terms of *causality* and *explainability*.

**Causality.** Table 3 visualizes the most and least causal concepts obtained from both baseline ConceptSHAP and our *HI-concept*. The words are orga-

| Method | Visualization |
|---|---|
| ConceptSHAP | dream team leads spain 44 - 42 at halftime athens, greece - as expected, the u.s. men's basketball team had its hands full in a quarterfinal game against spain on thursday... |
| *HI-concept* | dream team leads spain 44 - 42 at halftime athens, greece - as expected, the u.s. men's basketball team had its hands full in a quarterfinal game against spain on thursday ... |

Figure 4: Qualitative comparison from AG-News: "World" news misclassified as "Sports" by BERT.

| | Accuracy | Confidence | Time Spent |
|---|---|---|---|
| Plain | 72.5% | 3.2 | 10.7 |
| ConceptSHAP | 68.5% | 2.7 | 10.6 |
| Polyjuice | 73.5% | 2.6 | 7.6 |
| *HI-concept* | **80.5%** | **3.5** | **9.3** |

Table 4: Human study for explainability evaluation.

nized by descending concept importance scores (described in §3.3). For the most causal concept (*i.e.*, larger CACE), the one by ConceptSHAP implies technological news, but has some confounding keywords from the sports category (*e.g.*, "red", "super", "game"). The one by *HI-concept* clearly points to political news, without confounding words that belong to other categories. While for the least causal concept, the ConceptSHAP only consists of purely correlational and non-semantically meaningful words. Instead, *HI-concept* still contains class-specific words (*e.g.*, "us", "knicks"), which result in non-zero CACE. Overall, *HI-concept* results in a set of more task-relevant and semantically meaningful concepts.

**Explainability.** Fig. 4 shows the failure case ("World" news misclassified as "Sports") highlighted with the top concept discovered. ConceptSHAP discovers a top concept related to the keywords "leads", "as expected", or "on thursday", which are not informative as to why the model classified this input as "Sports". On the contrary, *HI-concept* could precisely point out why: BERT is looking at keywords such as "dream team", "game", and country names. Such examples show the potential of *HI-concept* being used in understanding the model's failure processes, which we further investigate in §5.5 with a carefully designed human study.

### 5.4 Generalization to Concept Insertion

As mentioned in §3.4, we study the causal impact of concepts by generalizing to a novel *insertion* operation. With the insertion of the found concepts one by one, we expect to observe *gradual improvement* of the recovering accuracy of the concept model. For example, we first evaluate the concept model (with 10 concepts) with only the most important concept, while masking all other concepts. Then, we evaluate the concept model with the two most important concepts, while masking all other concepts. The process goes on until we mask 0 concepts. As we unmask more and more concepts, the model performance is expected to gradually improve in order for each concept to have some causal importance. Formally, at the step $m \in 1, \ldots, n$, the concept model reconstruction becomes $g_\theta(p_c(x_j)|c_{i\in C\setminus C_m} = 0)$, where $C_m$ is the set of most important $m$ concepts.

Fig. 5 shows the trend results on the AG-News dataset. The concept is inserted in the order of descending importance. Obviously, our *HI-concept*, plotted as the red line, is the only method that shows gradual improvement consistently for all base models. While for other comparison methods, a single concept can already result in maximum accuracy, *e.g.*, all baselines on T5 and Llama, indicating less-causal sets of concepts overall.

### 5.5 Human Study

To systematically test whether derived features are explainable to humans, we design a human study to test the degree to which "a user can correctly and efficiently predict the method's results", which is the explainability definition by Kim et al. (2016). Inspired by the forward simulation design from Hase and Bansal (2020), we carefully conduct the following human study: We first show 100 randomly selected examples from AG's test set to users and ask them to predict the model's news topic classification. Then, we show the same examples again but with assistive information from *HI-concept*, including textual highlights and topic keywords, and ask users to predict the model's decision again. As a comparison, we show examples augmented by ConceptSHAP instead. For each question, we let users rate their confidence and record the time spent in seconds. Moreover, to test against local counterfactuals, which is a popular group of explainability methods, we also include Polyjuice (Wu et al., 2021) as another baseline. Polyjuice is a generator method that utilizes a finetuned GPT-2 model for producing diverse local counterfactuals to a sentence. Thus, it enables an automated approach to derive token explanations with Shapley values. Ideally, good explanations could help

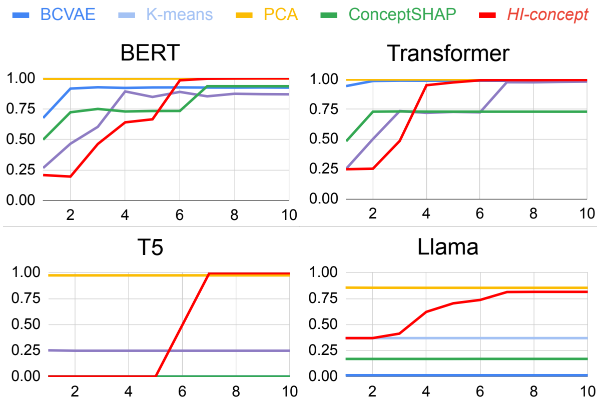


Figure 5: Effects of concept insertion on accuracy on AG-News dataset. Each figure represents a different model where the number of inserted concepts (x-axis) is plotted against accuracy (y-axis).

| Method | Acc | CACE | ΔAcc |
|---|---|---|---|
| Without Auto-Encoding Loss | 93.46% | 0.028 | 6.11% |
| Without Prediction Loss | 68.00% | 0.035 | **17.41%** |
| Without Regularizer Loss | 95.76% | 0.041 | 6.23% |
| Without Causality Loss | **99.92%** | 0.029 | 2.95% |
| *HI-concept* | 99.90% | **0.058** | 10.54% |

Table 5: Ablation on BERT for IMDB with faithfulness (Acc) and impact (CACE, ΔAcc) evaluation.

users better predict the model outcomes, thus increasing usability by resulting in higher accuracy and higher confidence. More details on the design can be found in Appendix F.

As shown in Table 4, when the users are given assistive information provided by *HI-concept*, their accuracy of predicting the model's decisions improved from 72.5% to 80.5%. On average, users also report higher confidence in their predictions and spend less time on the questions. When given correlational explanations by ConceptSHAP, however, both prediction accuracy and confidence decrease. Polyjuice, as a local counterfactual baseline, results in a human prediction accuracy of 73.5%. It surpasses the conceptSHAP baseline (68.5%) but still lags behind HI-Concept (80.5%). Moreover, HI-Concept also maintains the highest confidence score over all the baselines, outperforming Polyjuice by 1.1 (on a scale of 1-5). We note that users with Polyjuice spend less time (7.6s v.s 9.3s of HI-Concept) for the decision. It could be because Polyjuice tends to assign high importance to a selected few words, while giving minimal importance to others. This leads to quicker decision-making by users but is also accompanied by low accuracy and confidence. Overall, our study achieves the Cohen's Kappa agreement of 0.74, which is considered substantial agreement (Landis and Koch, 1977).

### 5.6 Ablation Study

To further investigate the effect of different loss objectives and various hyperparameters, we conduct multiple ablation studies.

**Loss objectives.** To ensure that the designated 4 objectives behave as expected, we conduct ablation studies for BERT on AG-News and report the results in Table 5. As observed, each designed loss plays its own role. Specifically, eliminating prediction loss leads to a large decrease in Acc, resulting in an unfaithful model. Therefore, even though its model explanations are more causal (large ΔAcc), the results cannot be trusted. Meanwhile, the auto-encoding and regularizer loss contribute to both faithfulness and causality, while causality loss mostly helps to ensure the causal metric. The full *HI-concept* method discovers a set of concepts with both good causality and faithfulness.

**Layer to Interpret.** We experiment on the 3rd, 6th, 9th, and 12th BERT layer respectively, all with 10 concepts. Overall, as shown in Fig. 6, the later layers tend to discover more class-coherent concepts. The beginning layers, however, could discover more abstract features and also lexical word clusters, such as concepts with only nouns or adjectives. This finding is confirmed by topic coherence metrics shown in Appendix G.1 and findings from Dalvi et al. (2021), where they observe that BERT finds more lexical information in the earlier layers. The detailed results are presented in Appendix G.1.

**Number of Concepts.** We experiment with 3, 5, 10, 50, and 100 concepts on the penultimate layer. The detailed results are presented in Appendix G.2. We find that a concept number close to the number of output classes usually gives higher prediction changes, while increasing the number results in higher recovering accuracy. When the number of concepts becomes larger, concepts usually become more coherent. However, with too large a number of concepts, the performance will decrease, as more noise is introduced into the training process.

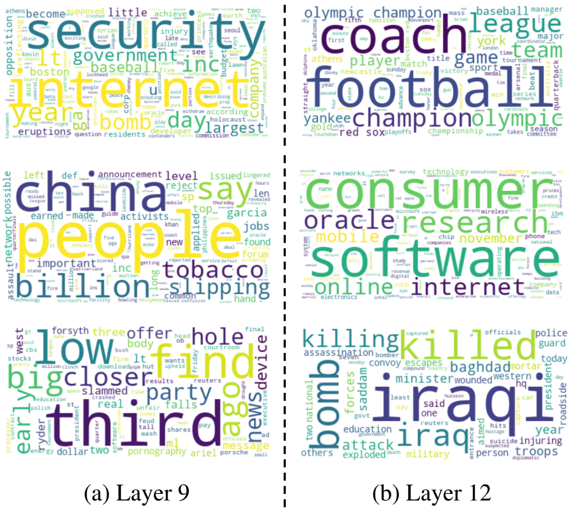
## 6 Related Work

**Concept-based Explanations** have been a explainability method that derive user-friendly, high-level concepts. Kim et al. (2018) first proposes TCAV, which derives concept vectors by training a linear classifier between a concept's examples and random counterexamples. Koh et al. (2020) provides a complete survey on concept bottleneck models and their interventions. Yeh et al. (2020) proposes an adapted Shapley value metric to evaluate completeness of explanations. However, as existing methods do not differentiate between correlational and causal information, their performances on NLP tasks are problematic, especially on LLMs with pretraining. Thus, some works measure their causal impacts by hidden space interventions (Harradon et al., 2018), counterfactuals (Feder et al., 2021b; Wu et al., 2023), or constructing relevant datasets (Abraham et al., 2022). However, they do not explicitly *optimize* for higher causal effects.

**Causality-aware Explanations** have two common methods. *Probing* methods (Conneau et al., 2018; Belinkov et al., 2020; Elazar et al., 2021) train an external model - a *probe* - to predict properties from the latent representations. However, it suffers from inherent flaws (Barrett et al., 2019; Belinkov, 2022), such as poor generalization. *Causal Mediation Analysis (CMA)* (Pearl, 2022; Vig et al., 2020) measures output change following a counterfactual intervention in an intermediate variable. Both methods can be viewed as supervised concept discovery algorithms. However, they could be limited as they rely on human-constructed features, requiring expertise. Thus, it may be beneficial to develop unsupervised explanation features. Specifically, in NLP, causality shows a promising path forward (Feder et al., 2021a), as it can offer insights into the model's inner workings. Most current methods attempt to causally explain LMs by generating *counterfactual* inputs (Alvarez-Melis and Jaakkola, 2017; Veitch et al., 2021; Wu et al., 2021).

## 7 Conclusions

We propose *HI-concept* to derive impactful concepts to explain the black-box language model's decisions. Our framework not only derives high-impact concepts that mitigate the confounding issue with the proposed causal objective, but also advances previous evaluations via both quantitative global accuracy change and qualitative insertion study. Extensive experiments, visualizations, figures, and human studies prove that our *HI-concept* can produce semantically coherent and user-friendly concept explanations.



(a) Layer 9

(b) Layer 12

Figure 6: Wordclouds of concepts generated on the 9th (left) and 12th (right) layer. The 9th layer includes a government concept, a China concept, and an Adjective (mostly) concept. The 12th layer includes a sports concept, a technology concept, and a political concept.

## Limitations

Regarding potential concerns, *HI-concept* only encourages high impact in post-hoc model explanations and should serve as an assistive tool instead of being accepted as ground-truth.

As a future venue to our work, we believe that *HI-Concept* sets a good foundation for future research on causal NLP explainability, especially for deriving human-friendly explanations. To improve it further, a similar causal objective could be used to address spurious correlations during training. It also has the potential of being carried over to other domains, such as vision or tabular tasks. The high-level attributes in the hidden space can also be used in downstream applications to provide better controllability for the users.

## Ethics Statement

*HI-concept* demonstrates the potential to play an important role in practical scenarios such as debugging and transparency. As AI ethics have become a major concern in real-life applications, such explanations can help users better identify bias and promote fairness.

## Acknowledgement

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG-PhD/2021-01-001).

## 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 *Advances in Neural Information Processing Systems*, volume 35, pages 17582–17596. Curran Associates, Inc.

David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 412–421, Copenhagen, Denmark. Association for Computational Linguistics.

Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial removal of demographic attributes revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6330–6335.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2020. On the linguistic representational power of neural machine translation models. *Computational Linguistics*, 46(1):1–52.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. 2018. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Fahim Dalvi, Abdul Rafae Khan, Firoj Alam, Nadir Durrani, Jia Xu, and Hassan Sajjad. 2021. Discovering latent concepts learned in bert. In *International Conference on Learning Representations*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *stat*, 1050:2.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E. Roberts, Brandon M. Stewart, Victor Veitch, and Diyi Yang. 2021a. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *CoRR*, abs/2109.00725.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021b. CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386.

Karl Pearson F.R.S. 1901. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.

Cyril Goutte and Eric Gaussier. 2005. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *European conference on information retrieval*, pages 345–359. Springer.

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

Michael Harradon, Jeff Druce, and Brian Ruttenberg. 2018. Causal learning and explanation of deep neural networks via autoencoded activations. *arXiv preprint arXiv:1802.00541*.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552, Online. Association for Computational Linguistics.

Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, 29.

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.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Aristidis Likas, Nikos Vlassis, and Jakob J. Verbeek. 2003. The global k-means clustering algorithm. *Pattern Recognition*, 36(2):451–461. Biometrics.

Andrew L. 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: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Sherin Mary Mathews. 2019. Explainable artificial intelligence applications in nlp, biomedical, and malware classification: a literature review. In *Intelligent computing-proceedings of the computing conference*, pages 1269–1292. Springer.

Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. 2020. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1):18–33.

Judea Pearl. 2009. *Causality: Models, Reasoning and Inference*, 2nd edition. Cambridge University Press.

Judea Pearl. 2022. Direct and indirect effects. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 373–392.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?". In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

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.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Victor Veitch, Alexander D'Amour, Steve Yadlowsky, and Jacob Eisenstein. 2021. Counterfactual invariance to spurious correlations in text classification. *Advances in Neural Information Processing Systems*, 34.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in Neural Information Processing Systems*, 33:12388–12401.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.

Zhengxuan Wu, Karel D'Oosterlinck, Atticus Geiger, Amir Zur, and Christopher Potts. 2023. Causal proxy models for concept-based model explanations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 37313–37334. PMLR.

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. *Advances in Neural Information Processing Systems*, 33:20554–20565.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*.

| Dataset | Train | Test | Label dim. | Avg. size |
|---|---|---|---|---|
| Toy (image) | 48k | 12k | 15 | (240, 240) |
| IMDB (text) | 37.5k | 2.5k | 2 | 215 |
| AG (text) | 120k | 7.6k | 4 | 43 |

Table 6: A summary of the datasets.

# Appendix for "Explaining Language Models' Predictions with High-Impact Concepts"

## A Training details

In practice, we only turn on the causal loss after a certain number of epochs (usually half of the overall number of epochs) to make sure that the surrogate model first learns to faithfully reconstruct from the set of concepts before optimizing for the impactful ones. This is because learning the two conflicting objectives at once will usually result in low accuracy. We also note that some contextual information is still needed to maximize the accurate reconstruction of hidden activations $\phi(\mathbf{x})$. Thus, the causality loss is enforced on all concepts except the last one $\mathbf{c}_n$, which is used as a 'context concept'. During model inference, the last (non-impactful) concept is unused.

After training, we post-process discovered concepts to filter out unused ones. While the number of concepts $n$ is user-selected, as in many topic models, it is an inherent flaw as it requires a certain level of domain expertise. For example, in a movie review dataset with only 2 output classes, if an unfamiliar user sets $n$ to 200, the model will naturally discover many noisy concepts and only a few useful ones. To ensure that the noisy concepts are eliminated, we post-process the concepts and filter out the unused ones (with an impact $I_{\text{ind}}(\mathbf{c}_i)$ close to 0). Thus, a more desirable number of concepts is returned even if the user provides an overestimate of $n$. In our experiments, we see that, after filtering, the model always achieves a better or same prediction-reconstruction performance as before. However, even with this post-processing, specifying too large a number of concepts can still be dangerous as it harms the concept model's training process.

## B Hyperparameters used

For all concept experiments, the following parameters are universally applied as a selected default, which demonstrated better performances during experiments: For regularizer losses, $\lambda_1 = 0.1$ and $\lambda_2 = 0.5$. In $\text{TH}(\cdot, \beta)$ function, threshold is set to be $\beta = 0.1 = \frac{1}{n}$, where n is the number of concepts selected. For the top-$N$ neighborhood, $N = \frac{1}{4}\text{BS}$, where BS is the effective batch size, which we have set as 128 during the experiments. For the masking strategy, we always recommend masking random concepts with a probability of $0.2$ as the optimal strategy, as masking maximum concepts may lead to a highly uneven distribution of $\text{I}(\mathcal{C})$ among discovered concepts.

As all dataset class sizes are small (2 in IMDB/toy or 4 in AG-News), the number of concepts is chosen to be 10 for all experiments. When the number of classes is larger, we recommend choosing a larger number of concepts to ensure a faithful reconstruction of the original input.

For training the concept model, we always use an Adam optimizer with a learning rate of $3e-4$. All models are all trained using 100 epochs. In the *HI-concept* models, causal loss is always turned on at half of the overall number of epochs. After turning on causal loss, all parameters are set to untrainable except for the concept vectors, which ensures that the reconstruction ability is not forgotten.

The same hyperparameters are set for the conceptSHAP models, which are also found to generate the optimal performances. The threshold is set to be $\beta = 0.3$, as recommended by the original paper on NLP datasets.

For the causal loss regularizer, $\lambda_{\text{c}} = 1$ is set for all experiments, except for $\lambda_{\text{c}} = 3$ in the case of IMDB with BERT. A higher $\lambda_{\text{c}}$ will usually lead to a higher output change ($\text{I}(\mathcal{C})$ and $\Delta$Acc), accompanied by a decrease in faithfulness (RAcc).

To reproduce, all experiments were run with a random seed of 0.

A summary of the datasets is provided in 6. IMDB and AG-news are both licensed for non-commercial use.

## C Quantitative metrics

**Faithfulness:** To ensure that the surrogate model can accurately mimic the original model's prediction process, we evaluate whether the captured concept probabilities $p_{\mathcal{C}}(\mathbf{x})$ can recover the original model's predictions $\psi\big(\phi(\mathbf{x})\big)$ with the established metrics below:

(*i*) *Recovering Accuracy* (Acc): As defined in Yeh et al. (2020), for the set of concepts $\mathcal{C}$, RAcc measures the prediction reconstruction accuracy using concept scores:

$$\mathrm{RAcc}(\mathcal{C}) = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{\mathbf{x}_j \in \mathcal{D}_{\text{test}}} \mathbb{1}\Big(\psi(\phi(\mathbf{x}_j)) = \psi(g_\theta(p_{\mathcal{C}}(\mathbf{x}_j)))\Big)$$

(*ii*) *Precision, Recall, F1*: To provide a thorough study, we also include common metrics including precision, recall, and F1 (Goutte and Gaussier, 2005).
(*iii*) *Completeness*: As defined in Yeh et al. (2020), completeness measures whether $\mathcal{C}$ is sufficient in recovering predictions. Denoting $\sup_g \mathbb{P}_{x,y \in \mathcal{D}_{\text{test}}}[y = \arg\max_{y'} \psi_{y'}(g_\theta(p_{\mathcal{C}}(\mathbf{x}_j)))]$ as the best accuracy by predicting the label just given the concept scores, and $a_r$ as the accuracy of random prediction, completeness is formulated as:

$$\text{Completeness}(\mathcal{C}) = \frac{\sup_g \mathbb{P}_{x,y \in \mathcal{D}_{\text{test}}}[y = \arg\max_{y'} \psi_{y'}(g_\theta(p_{\mathcal{C}}(\mathbf{x}_j)))] - a_r}{\mathbb{P}_{x,y \in \mathcal{D}_{\text{test}}}[y = \arg\max_{y'} f_{y'}(x)] - a_r}$$

**Causality:** To systematically evaluate causality, we conduct synthetic experiments, derive qualitative examples, draw trend graphs, and conduct human studies. In quantitative experiments, we use the following quantitative metrics:
(*i*) *Causal Concept Effect (CACE)*: As defined in Goyal et al. (2019), CACE for a concept $c$ is the change in prediction after removing it. Then, we compute the average CACE to evaluate a set of concepts $\mathcal{C}$:

$$\mathrm{CACE}(\mathbf{c}_i) = \mathbb{E}\big[\psi(g_\theta(p_{\mathcal{C}}(\mathbf{x}_j))) - \psi(g_\theta(p_{\mathcal{C}\setminus\{i\}}(\mathbf{x}_j)))\big]$$

(*ii*) *Recovering Accuracy Change ($\Delta$Acc)*: Doshi-Velez and Kim (2017) state: "Causality implies that the predicted change in output due to a perturbation will occur in the real system". Therefore, if a concept $\mathbf{c}_i$ is a crucial factor used by the model to make predictions, omitting it should disrupt the ability to faithfully recover predictions:

$$\Delta\mathrm{Acc}(\mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{\mathbf{c}_i \in \mathcal{C}} |\mathrm{RAcc}(\mathcal{C}) - \mathrm{RAcc}(\mathcal{C} \setminus \{\mathbf{c}_i\})|$$

# D Toy example

We conduct experiments on a synthetic (toy) image dataset with ground truth concepts in order to test the validity of our method and confirm the claim that higher confounding effects within the dataset lead to more correlational explanations, thus calling for a more causal explainability approach. Specifically, We extend the toy dataset design of Yeh et al. (2020) to make it more realistic by inserting spurious correlations.

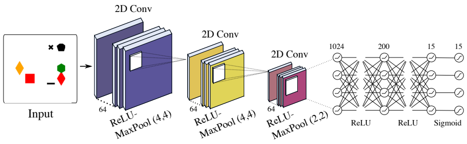


Figure 7: Convolutional Neural Network used for classifying the toy dataset.

## D.1 Data generation

As a synthetic setup, at most 15 shapes are randomly scattered on a blank canvas at random locations with random color selections (as noise). For each image sample $x_j$, $\boldsymbol{z}^{j}_{\{1:15\}}$ are binary variables of whether or not a shape is present in $x_j$ with each $\boldsymbol{z}^{j}_{s}$ sampling from a Bernoulli distribution with probability 0.5. Then, a 15-class target $\boldsymbol{y_j}$ is constructed with respect to whether the first 5 shapes ($\boldsymbol{z}^{j}_{\{1:5\}}$) are present or not with human-designed rules. For example, $\boldsymbol{y_1} = \sim (\boldsymbol{z_1} \cdot \boldsymbol{z_3}) + \boldsymbol{z_4}$. A total of $60,000$ examples are generated as the toy dataset using a seed of 0.

The setup mentioned above is, in fact, far away from realistic scenarios, as it does not consider possible confounding. Thus, to make it more realistic, we insert spurious correlations between the pairs $(\boldsymbol{z}^{j}_{\{1:5\}}, \boldsymbol{z}^{j}_{\{6:10\}}), (\boldsymbol{z}^{j}_{\{6:10\}}, \boldsymbol{z}^{j}_{\{11:15\}})$ with a correlation factor $p_{\text{cor}}$. For example, when $z_1 = 1$, $z_6 = \text{Bernoulli}(p_{\text{cor}})$; when $z_1 = 0$, $z_6 = \text{Bernoulli}(1 - p_{\text{cor}})$.

## D.2 CNN classification model used for the toy example

The CNN classification model used for the toy dataset is shown in Fig. 7. Specifically, 3 convolutional layers with a kernel size of 5 and 64 output channels were used, each followed by a ReLU activation and max pooling layer. Then, the result is flattened into a linear vector, followed by 2 linear layers and a sigmoid activation function. The output is a 15-dimensional binary classification probability. The model is trained for 100 epochs with an Adam optimizer with learning rate $3e-4$. For reproducibility purposes, the model is initialized and trained with a seed of 0.

## D.3 Visualizations

As an example visualization, in Fig. 8, two random images from the toy dataset are displayed on the left, while three example concepts discovered by *HI-concept* are plotted on the right. We could observe that *HI-concept* is able to derive meaningful

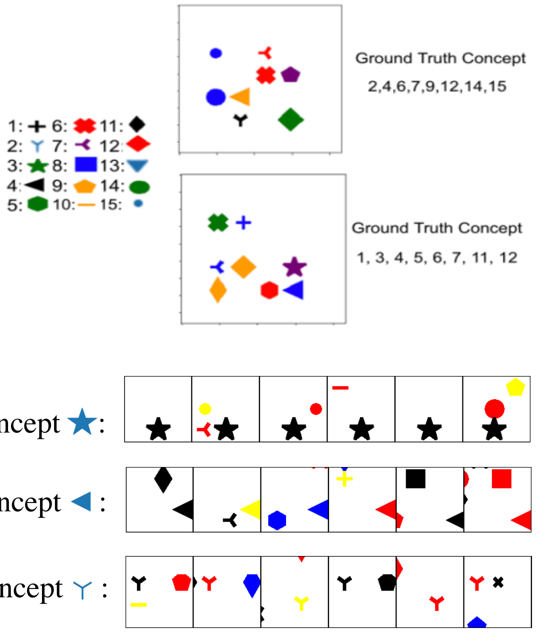


Figure 8: Examples from the toy dataset and concepts discovered.

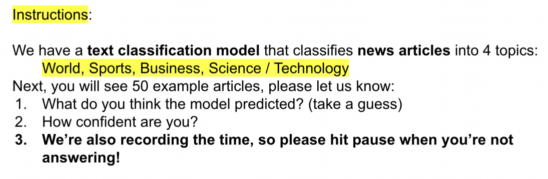
clusters as concepts, which provide a sanity check for usability of the latent concepts.

### D.4 Results on toy dataset

From the results shown in Table 1, we could observe that, as we increase $p_{\text{cor}}$ to mimic an increase in confounding levels in real life, our *HI-concept* consistently outperforms the baseline by a bigger margin. HI-Concept achieves higher impacts ($\mathrm{I}(\mathcal{C})$) and higher accuracy change ($\Delta$Acc), while maintaining the best RAcc, indicating faithfulness to the original predictions. Moreover, we note that the improvement is even stronger in real data experiments, as the added artificial confounding is more complicated in real-life scenarios.

## E Text classification results on T5

The results on pretrained and finetuned T5 model can be found in Table 7. Similar to Llama, as T5 is also a generative model instead of a classification model, the output space is much larger and harder to reconstruct. In this case, only the PCA method is able to accurately reconstruct the output classifications. All baseline methods generate features with minimal impact on outputs. Only *HI-concept* maintains both good reconstruction performance and high impact at the same time.

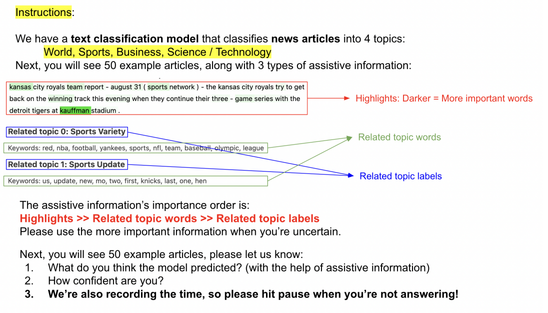


Instructions:

We have a **text classification model** that classifies **news articles** into 4 topics:
World, Sports, Business, Science / Technology
Next, you will see 50 example articles, please let us know:

1. What do you think the model predicted? (take a guess)
2. How confident are you?
3. **We're also recording the time, so please hit pause when you're not answering!**

Figure 9: Human study instructions for plain examples.



Instructions:

We have a **text classification model** that classifies **news articles** into 4 topics:
World, Sports, Business, Science / Technology
Next, you will see 50 example articles, along with 3 types of assistive information:

kansas city royals team report - august 31 ( sports network ) - the kansas city royals try to get back on the winning track this evening when they continue their three - game series with the detroit tigers at kauffman stadium .

**Related topic 0: Sports Variety**
Keywords: red, nba, football, yankees, sports, nfl, team, baseball, olympic, league
**Related topic 1: Sports Update**
Keywords: us, update, new, mo, two, first, knicks, last, one, hen

The assistive information's importance order is:
**Highlights >> Related topic words >> Related topic labels**
Please use the more important information when you're uncertain.

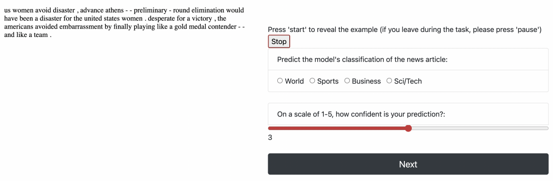
Next, you will see 50 example articles, please let us know:

1. What do you think the model predicted? (with the help of assistive information)
2. How confident are you?
3. **We're also recording the time, so please hit pause when you're not answering!**

Figure 10: Human study instructions for *HI-concept* augmented examples.

## F Human study setup

For the human study, 100 examples are randomly selected from the test set $\mathcal{D}_{\text{test}}$. The questionnaire takes the format of a self-constructed website. Firstly, we show the examples without any assistive information, where the instructions are shown in Fig. 9 and an example question looks like Fig. 11. Secondly, the same examples are shown with assistive information derived from Concept-SHAP. Lastly, the examples are shown with assistive information derived from *HI-Concept*. The instructions are shown in Fig. 10 and an example question looks like Fig. 12. 4 volunteers (Ph.D. students) each answered 50 plain examples and 50 augmented examples. The volunteers are all proficient in English. The volunteers report an average time of approximately 30 minutes for answering all 100 questions. As the volunteers are working also in AI-related areas and are briefed about the purpose and usage of survey data beforehand, they understand fully the data collection and usage. Thus, implicit consent is granted by participation.



us women avoid disaster , advance athens - - preliminary - round elimination would have been a disaster for the united states women . desperate for a victory , the americans avoided embarrassment by finally playing like a gold medal contender - - and like a team .

Press 'start' to reveal the example (if you leave during the task, please press 'pause')

Predict the model's classification of the news article:

○ World ○ Sports ○ Business ○ Sci/Tech

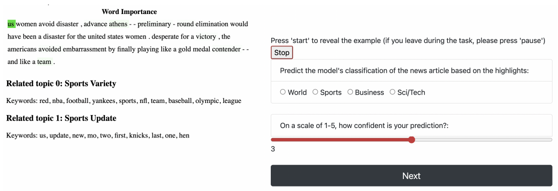
On a scale of 1-5, how confident is your prediction?:

3

Figure 11: Human study question and answer.

| Dataset | Model | Method | Acc | Precision | Recall | F1 | Completeness | CACE | ΔAcc |
|---|---|---|---|---|---|---|---|---|---|
| IMDB | T5 | $\beta$-TCVAE (Chen et al., 2018) | 0.00% | 0.00 | 0.00 | 0.00 | -23.70 | 0.000 | 0.00% |
| | | K-means (Likas et al., 2003) | 75.85% | 37.92 | 50.00 | 43.13 | 26.83 | 0.025 | 1.06 |
| | | PCA (F.R.S., 1901) | 98.86% | 99.04 | 97.85 | 98.43 | **48.42** | 0.000 | 0.02% |
| | | ConceptSHAP (Yeh et al., 2020) | 0.00% | 0.00 | 0.00 | 0.00 | -23.70 | 0.000 | 20.21% |
| | | *HI-concept* | **99.50%** | **99.65** | **98.98** | **99.31** | 48.87 | **0.153** | **62.47%** |
| AG-News | T5 | $\beta$-TCVAE (Chen et al., 2018) | 0.00% | 0.00 | 0.00 | 0.00 | -20.60 | 0.000 | 0.00% |
| | | K-means (Likas et al., 2003) | 24.87% | 6.22 | 25.00 | 9.96 | 4.40 | 0.011 | 1.49% |
| | | PCA (F.R.S., 1901) | 97.38% | 97.40 | 97.37 | 97.38 | 73.12 | 0.000 | 0.01% |
| | | ConceptSHAP (Yeh et al., 2020) | 0.00% | 0.00 | 0.00 | 0.00 | -20.60 | 0.000 | 0.01% |
| | | *HI-concept* | **99.46%** | **99.46** | **99.46** | **99.46** | **73.70** | **0.075** | **72.37%** |

Table 7: Faithfulness (Acc, Precision, Recall, F1, Completeness) and causality (CACE, ΔAcc) evaluation of pretrained and finetuned T5.



**Word Importance**

us women avoid disaster , advance athens - - preliminary - round elimination would have been a disaster for the united states women . desperate for a victory , the americans avoided embarrassment by finally playing like a gold medal contender - - and like a team .

**Related topic 0: Sports Variety**

Keywords: red, nba, football, yankees, sports, nfl, team, baseball, olympic, league

**Related topic 1: Sports Update**

Keywords: us, update, new, mo, two, first, knicks, last, one, hen

Press 'start' to reveal the example (if you leave during the task, please press 'pause')

Stop

Predict the model's classification of the news article based on the highlights:

○ World ○ Sports ○ Business ○ Sci/Tech

On a scale of 1-5, how confident is your prediction?:

3

Next

Figure 12: Human study question and answer.

As one resulting concept is "a group of words that are meaningful" (Dalvi et al., 2021), which could take some time for humans to read, we also employ an LLM (GPT-3.5) to summarize the words into an assistive label. The resulting labels allow humans to quickly grasp the gist of an abstract concept. Specifically, we used the GPT-3.5-turbo model with the following prompt:

"You're an expert in topic labeling. Please come up with a short word or phrase that summarizes the topic with the keywords below:

[set of keywords]"

## G Hyperparameter comparisons

The proposed method of *HI-concept* includes many tunable hyperparameters, including the top-N neighborhood, threshold, etc. While these parameters are set at the default mentioned in Appendix B, there are two hyperparameters that users can customize the most: the layer to interpret at and number of concepts . To better understand how these two parameters may affect the generated concepts, we conduct comparisons on both. We evaluate in terms of impact and topic quality. For impact, we have reported the number of effective concepts left after post-processing, the recovering accuracy (RAcc), the Average Impact ($\mathrm{I}(\mathcal{C})$), and the induced change in accuracy (ΔAcc). For topic quality, we have reported coherence scores, including averaged Pointwise Mutual Information (PMI) (c_uci score), normalized PMI (c_npmi score), c_v score which measures how often the topic words appear together in the corpus, and word2vec similarity (Röder et al., 2015).

The following comparisons are all conducted on the AG-news dataset with BERT, where the other hyperparameters mentioned in Appendix B stay the same.

### G.1 Layer-wise comparison

To compare what each layer discovered, as BERT has 12 layers, we experimented on the 3rd, 6th, 9th, and penultimate layer respectively, all with 10 concepts.

Quantitatively, we plotted out the effective number of concepts, recovering accuracy, impact and accuracy change in Fig. 13. All layers demonstrate similar performances in recovering accuracy, which is close to 100%. The intermediate layers, especially the 6th layer, produce a higher average impact and recovering accuracy. This is because the intermediate layers discover concepts on the token-level, while the penultimate layer concepts are sentence-level (on the [CLS] token). Thus, the token-level concepts will have more fine-grained control.

Qualitatively, we plotted some wordclouds of the keywords in discovered concepts in Fig. 6. We could see that, in the penultimate layer, concepts are more concentrated on each class. For example, the first concept would correspond to the class "Sports", the second to "Sci/Tech", and the third to "World" news. The emphasis on events is also clearer, such as the third one talking about the Iraq War. However, When we move to earlier layers, the concepts' class labels are more mixed together. In the 9th layer, the first concept concerns government, which includes terms such as "government", "internet", "security", "bomb", "baseball", etc. It could, however, correspond to many class labels, such as "Sci/Tech", "World", or even "Sports".

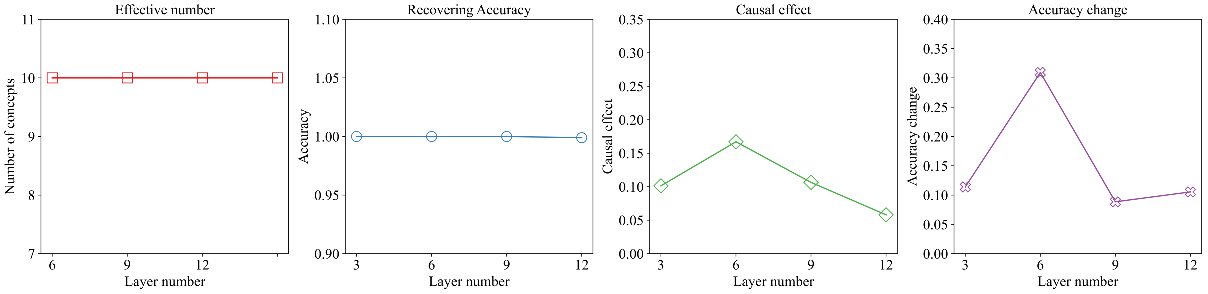


Figure 13: Layer-wise effective number of concepts, RAcc ↑, I($\mathcal{C}$) ↑, and $\Delta$ Acc ↑.

Similarity, the second concept talks about China, including “china”, “billion”, “people”, “activitists”, “announcement”, etc. The third concept is interesting as it covers mostly adjective words which do not seem to correlate too much in semantic meanings, such as “low”, “big”, “closer”, and “third”. Similar observations are also confirmed in papers such as (Dalvi et al., 2021), which derives concepts using agglomerative hierarchical clustering combined with human annotations in BERT latent representations. They observe that BERT finds more lexical information in the earlier layers.

In terms of topic quality, we evaluated the concept keywords using coherence metrics. As shown in Fig. 14, all coherence scores showed a general trend of concepts becoming more coherent as the layer number increases. The conclusion is consistent with the wordcloud visualizations.

Thus, in real-life debugging scenarios, we recommend using the penultimate layer, which will find more coherent topics. However, there could be continued work to discover information learned in the prior layers and to investigate how information flows through layers in a hierarchical way.

### G.2 Number of concepts

In the penultimate layer of BERT, we experiment with 3, 5, 10, 50, and 100 concepts.

From Fig. 15, we could see that the performance is very dependent on the number of concepts. The effective number of concepts, recovering accuracy, average impact, and accuracy change all appear to be elbow-shaped. In this case, 5 concepts provided the highest impact on output predictions, as it is close to the number of classes (4) in the AG-News dataset. Increasing the number of concepts to 10 would yield a better recovering accuracy. As the number of concepts increases to 50 and 100, we observe that the model fails to learn completely. In practice, we have often observed the best number to be positively correlated with the number of dataset classes. In other words, a dataset with more classes will require a higher number of concepts for faithful reconstruction. In terms of topic coherence, we could observe from Fig. 16 that the topic coherence scores usually oscillate, but mostly display a generally upward trend of becoming more coherent as the number of concepts increases.

## H Classification models used for text experiments

### H.1 Transformer classification model trained from scratch

The self-trained transformer model used during text experiments follows a simple structure: the input text is truncated to max length 512 and passed to an embedding layer of dimension 200. Then, the embeddings are passed through a positional encoding layer with dropout rate 0.2. Then, 6 transformer layers follow with a hidden dimension of 200 and 2 heads. Finally, we mean pool the transformed embeddings and pass through a linear classifier head. The linear outputs are activated with a Sigmoid function to produce class probabilities.

To train the transformer model, we use either the IMDB or AG-News dataset. We train for 10 epochs with a batch size of 128 and an Adam optimizer with learning rate $3e-4$. We also use a learning rate step scheduler with step size 1 and gamma of 0.95.

### H.2 Pretrained and finetuned BERT model

For AG-News, we take the finetuned version of bert-base-uncased model on huggingface: “fabriceyhc/bert-base-uncased-ag_news”. For IMDB, we finetuned by ourselves on the bert-base-uncased model. The hyperparameters used for both finetuning are reported in Appendix H.1, where LR stands for learning rate and BS stands for batch size.

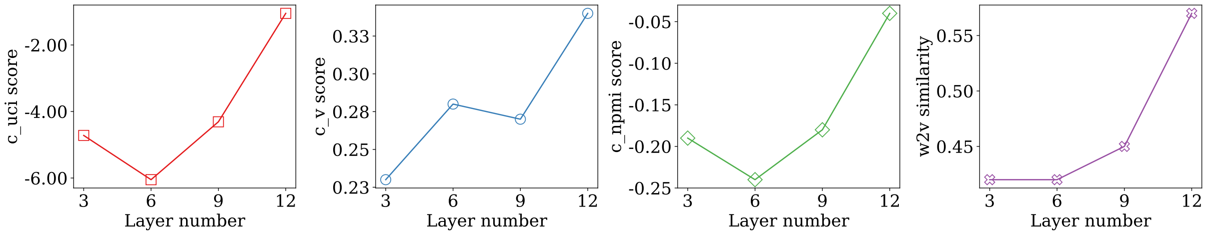


Figure 14: Layer-wise Topic Coherence Comparison.

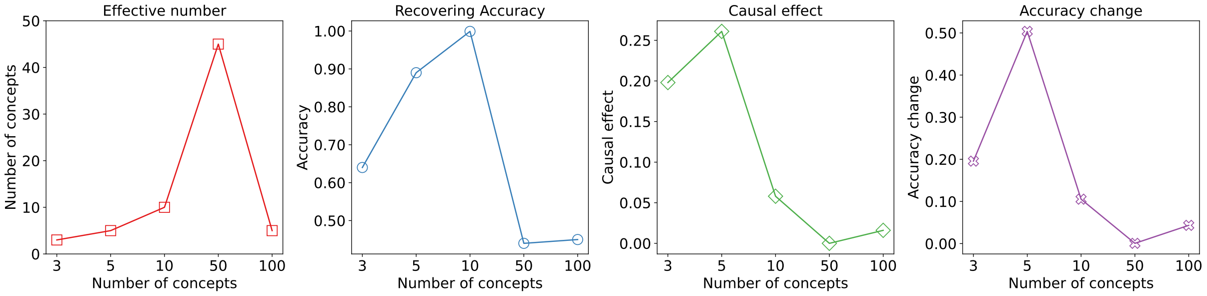


Figure 15: Concept-wise effective number of concepts, RAcc ↑, I($\mathcal{C}$) ↑, and $\Delta$ Acc ↑.

| Dataset | AG-News | IMDB |
|---|---|---|
| LR | $5e-5$ | $3e-4$ |
| train BS | 8 | 8 |
| eval. BS | 16 | 16 |
| seed | 42 | 42 |
| optimizer | Adam<br>betas $= (0.9, 0.999)$<br>epsilon $= 1e-8$ | Adam<br>betas $= (0.9, 0.999)$<br>epsilon $= 1e-8$ |
| LR scheduler | linear | linear |
| warmup steps | 7425 | 1546 |
| training steps | 74250 | 15468 |

Table 8: Hyperparameters for finetuning BERT model.

The huggingface code and models are all licensed under Apache 2.0, which allows for redistribution and modification. Similarly, the codebase used for replicating the visualization method (Chefer et al., 2021) and the baseline method (Chen et al., 2018) are licensed under the MIT license, which allows for redistribution of the code.

### H.3 T5 and Llama

As T5 and Llama are both generative models, when calculating impact, we simplify outputs by filtering to only the classification classes (e.g., words "Positive", "Negative" for IMDB) and summing all other vocab probabilities as "Other".

For T5, we finetune on IMDB and AG-News separately using the same hyperparameters: max seq length of 512, learning rate of $3e-4$, weight decay of 0.0, adam epsilon of $1e-8$, warmup steps of 0, train batch size of 10, eval batch size of 10, num train epochs of 2, and gradient accumulation steps of 8.

The T5 model is licensed under Apache 2.0, which allows for redistribution and modification.

For Llama, we use the 7B model licensed under GPL 3.0, which allows for redistribution and modification. Specifically, we use the following in-context learning prompt:

**IMDB** Given a movie review, classify its sentiment into positive or negative.

### Moview review: Sorry, gave it a 1, which is the rating I give to movies on which I walk out or fall asleep. In this case I fell asleep 10 minutes from the end, really, really bored and not caring at all about what happened next.

### Sentiment:

negative

### Movie review: Zentropa has much in common with The Third Man, another noir-like film set among the rubble of postwar Europe. Like TTM, there is much inventive camera work. There is an innocent American who gets emotionally involved with a woman he doesn't really understand, and whose naivety is all the more striking in contrast with the natives.<br /><br />But I'd have to say that The Third Man has a more well-crafted storyline. Zentropa is a bit disjointed in this respect. Perhaps this is intentional: it is presented as a dream/nightmare, and making it too coherent would spoil the effect. <br /><br />This movie is

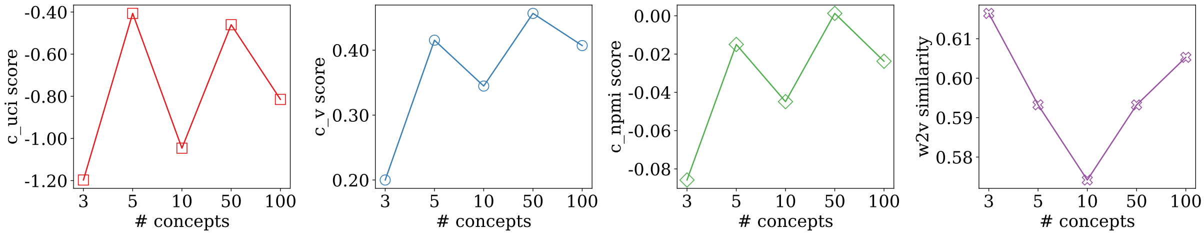


Figure 16: Concept-wise Topic Coherence Comparison.

| Dataset | $\beta$-TCVAE | kmeans | PCA | conceptSHAP | *HI-concept* |
|---|---|---|---|---|---|
| IMDB | 475.9 | 37.7 | 0.8 | 199.3 | 227.2 |
| AG | 1525.6 | 15.51 | 2.5 | 1749.65 | 2242.1 |

Table 9: A summary of runtime (in seconds) on datasets for BERT.

unrelentingly grim–"noir" in more than one sense; one never sees the sun shine. Grim, but intriguing, and frightening.
### Sentiment:
positive

### Moview review:
**INPUT**
### Sentiment:

**AG-News** Given a news article, classify its category into World, Sports, Business, or Tech.

### News article:
IBM to hire even more new workers By the end of the year, the computing giant plans to have its biggest headcount since 1991.
### Topic:
Tech

### News article: Fears for T N pension after talks Unions representing workers at Turner Newall say they are 'disappointed' after talks with stricken parent firm Federal Mogul.
### Topic:
Business

### News article:
**INPUT**
### Category:

# I Run-time

As our model optimizes for causality loss, the runtime is slightly longer than the baseline method ConceptSHAP (Yeh et al., 2020), but is still short. A summary of runtime is shown in Appendix I. All models shown are run on the GTX 1080Ti graphic card with 12 GB memory. Generally, as post-hoc explainability methods, the runtimes are very light and, therefore, a concern that is less important than the model quality. For example, on a dataset of size 50k such as IMDB, it only takes 227.2 seconds (3.8) minutes to train our *HI-concept* model.