Interpreting Pretrained Language Models via Concept Bottlenecks (Extended Abstract)*

Zhen Tan[1], **Lu Cheng**[2], **Song Wang**[3], **Bo Yuan**[4], **Jundong Li**[3], **Huan Liu**[1]
[1]Arizona State University, Arizona, USA
[2]University of Illinois Chicago, Illinois, USA
[3]University of Virginia, Virginia, USA
[4]Independent Researcher
{ztan36, huanliu}@asu.edu, lucheng@uic.edu, {sw3wv, jundong}@virginia.edu

Abstract

Pretrained language models (PLMs) achieve state-of-the-art results but often function as "black boxes", hindering interpretability and responsible deployment. While methods like attention analysis exist, they often lack clarity and intuitiveness. We propose interpreting PLMs through high-level, human-understandable concepts using Concept Bottleneck Models (CBMs). This extended abstract introduces C^3M (ChatGPT-guided Concept augmentation with Concept-level Mixup), a novel framework for training Concept-Bottleneck-Enabled PLMs (CBE-PLMs). C^3M leverages Large Language Models (LLMs) like ChatGPT to augment concept sets and generate noisy concept labels, combined with a concept-level MixUp mechanism to enhance robustness and effectively learn from both human-annotated and machine-generated concepts. Empirical results show our approach provides intuitive explanations, aids model diagnosis via test-time intervention, and improves the interpretability-utility trade-off, even with limited or noisy concept annotations. Code and data are released at https://github.com/Zhen-Tan-dmml/CBM_NLP.git

1 Introduction

Although Pretrained Language Models (PLMs) like BERT [Devlin *et al.*, 2018] have achieved remarkable success in various NLP tasks [Zhu *et al.*, 2020], they are frequently regarded as black boxes, posing significant obstacles to their responsible deployment in real-world scenarios, particularly in critical domains such as healthcare [Koh *et al.*, 2020]. To date, many existing works [Madsen *et al.*, 2022] leverage attention weights extracted from the self-attention layers to provide token-level or phrase-level importance. These low-level explanations are found unfaithful [Yin and Neubig, 2022] and lack readability and intuitiveness [Losch *et al.*, 2019], leading to unstable or even unreasonable explanations. To address these limitations, we seek to explain via human-comprehensible *concepts* that use more abstract features (e.g., general notions) as opposed to raw input features at the token level [Zarlenga *et al.*, 2022]. The foundation of this work is the Concept Bottleneck Models (CBMs) [Koh *et al.*, 2020] that interpret deep models (e.g., ResNet [He *et al.*, 2016]) for image classification tasks using high-level concepts (e.g., shape). For NLP tasks such as sentiment analysis, concepts can be Food, Ambiance, and Service as shown in Figure 1, where each concept corresponds to a neuron in the concept bottleneck layer. The final decision layer is then a linear function of these concepts. Using concepts greatly improves the readability and intuitiveness of the explanations compared to low-level features such as "lobster".

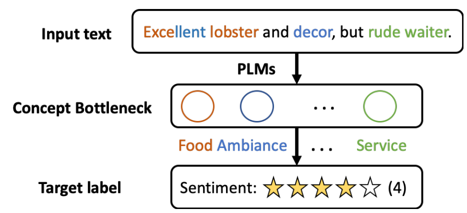


Figure 1: The illustration of CBE-PLMs. Through black-box PLMs, the input text x is mapped into an intermediate layer consisting of a set of human-comprehensible concepts c, which are then used to predict the target label y.

We propose to study *Concept-Bottleneck-Enabled Pretrained Language Models* (CBE-PLMs). There are two key challenges: ❶ First, existing CBMs [Koh *et al.*, 2020; Zarlenga *et al.*, 2022] require human-annotated concepts. This can be challenging for natural language since the annotator may need to read through the entire text to understand the context and label one concept [Németh *et al.*, 2020]. This limits the practical usage and scalability of CBE-PLMs. ❷ Second, many studies have identified the trade-off between interpretability and task accuracy using CBMs since the predetermined concepts may leave out important information for target task prediction [Zarlenga *et al.*, 2022]. Therefore, it is crucial to improve both interpretability and task performance to achieve optimal interpretability-utility trade-off.

❶ To tackle the first challenge, we propose leveraging Large Language Models (LLMs) trained on extensive human-

*This is an concise version of [Tan *et al.*, 2024b], recipient of the Best Paper Award at PAKDD 2024.

generated corpora and feedbacks, such as ChatGPT [OpenAI, 2023], to identify novel concepts in text and generate pseudo-labels (via prompting) for unlabeled concepts. Recent studies [OpenAI, 2023] exhibit that these LLMs encapsulate significant amounts of human common sense knowledge. By augmenting the small set of human-specified concepts with machine-generated concepts, we increase concept diversity and useful information for prediction. In addition, generated pseudo-labels offer us a large set of instances with noisy concept labels, complementing the smaller set of instances with clean labels. ❷ To further improve interpretability-utility trade-off (second challenge), we propose to learn from noisy concept labels and incorporate a concept-level MixUp mechanism [Zhang *et al.*, 2017] that allows CBE-PLMs to cooperatively learn from both noisy and clean concept sets. We name our framework for training CBE-PLMs as *ChatGPT-guided Concept augmentation with Concept-level Mixup* (C^3M). In summary, our contributions include:

- We provide the first investigation of utilizing CBMs for interpreting PLMs.
- We propose C^3M, which leverages LLMs and MixUp to help PLMs learn from human-annotated and machine-generated concepts. C^3M liberates CBMs from predefined concepts and enhances the interpretability-utility trade-off.
- We demonstrate the effectiveness and robustness of test-time concept intervention for the learned CBE-PLMs for common text classification tasks.

2 Related Work

2.1 Interpreting Pretrained Language Models

PLMs such as Word2Vec [Mikolov *et al.*, 2013], BERT [Devlin *et al.*, 2018], and the more recent GPT series [OpenAI, 2023] have demonstrated impressive performance in various NLP tasks. However, their opaque nature poses a challenge in comprehending how PLMs work internally [Diao *et al.*, 2022]. In order to improve the interpretability and transparency of PLMs, researchers have explored different approaches, such as visualizing attention weights [Galassi *et al.*, 2020], probing feature representations [Bills *et al.*, 2023], and using counterfactuals [Ross *et al.*, 2021], among others, to provide explanations at the local token-level, instance-level, or neuron-level. However, these methods often lack faithfulness and intuitiveness, and are of poor readability, which undermines their trustworthiness [Madsen *et al.*, 2022].

Recently, researchers have turned to global concept-level explanations that are naturally understandable to humans. Although this level of interpretability has been less explored in NLP compared to computer vision [Kim *et al.*, 2018], it has gained attention. For instance, a study [Vig *et al.*, 2020] investigates gender classification bias by examining the association of occupation words such as "nurse" with gender. In addition, the CBMs [Koh *et al.*, 2020] have emerged as novel frameworks for achieving concept-level interpretability in lightweight image classification systems. CBMs typically involve a layer preceding the final fully connected classifier, where each neuron corresponds to a concept that can be interpreted by humans. CBMs also show advantages in improving accuracy through human intervention during testing. Yet, the application of CBMs to larger-scale PLMs interpretation is under-explored. Implementing CBMs necessitates human involvement in defining the concept set and annotating the concept labels. Such requirements are challenging for natural language as humans may need to read through the entire text to understand the context and label one concept [Németh *et al.*, 2020].

2.2 Learning from Noisy Labels

Addressing inaccurately labeled or misclassified data in real-world scenarios is the goal of learning from noisy labels, with techniques including noise transition matrix estimation [Liu *et al.*, 2022], robust risk minimization [Englesson and Azizpour, 2021], and more. Recently, the resilience of semi-supervised learning methods like MixMatch [Berthelot *et al.*, 2019] and FixMatch [Sohn *et al.*, 2020] to label noise has been discovered by using pseudo-labels for unlabeled data. Inspired by them, we porpose to utilize an LLM (ChatGPT) as a fixed-label guesser, generating noisy intermediate concept labels to potentially predict task labels.

Notably, CBMs specialize in the interpretation and interactability of deep models for general classification tasks. While *Multi-Aspect Sentiment Analysis* [Zhang *et al.*, 2022] (MASA) shares similar goals when using aspects as concepts, it differs as concepts are not confined to fine-grained aspectual features and can be abstract ideas or broader notions throughout entire contexts. Aspect labels in MASA, primarily used for prediction accuracy, are not always mandatory. To summarize, this study pioneers the comprehensive exploration of utilizing concepts for interpreting large-scale PLMs, and provids a robust framework for harnessing the noisy signals from LLMs to achieve interpretable outcomes from lighter-weight PLMs, which can be easily understood by users.

3 Concept-Bottleneck-Enabled PLMs (CBE-PLMs) with C^3M

3.1 CBE-PLM Architecture

We adapt CBMs for PLMs by introducing a projector layer p_ψ after the PLM encoder f_θ. This layer maps the PLM's latent representation $z = f_\theta(x)$ to a concept activation vector $\hat{c} = p_\psi(z)$, where each dimension corresponds to a concept. A final predictor g_ϕ then maps these concept activations to the task label $\hat{y} = g_\phi(\hat{c})$. The model structure is $x \to z \to \hat{c} \to \hat{y}$. Concepts can be multi-class (e.g., positive/negative/unknown).

3.2 The C^3M Framework

C^3M enables training CBE-PLMs effectively even with limited human-annotated concepts ($\mathcal{D}_s$) and abundant unlabeled data ($\mathcal{D}_u$). It involves two main stages (illustrated conceptually in Figure 2 of the original paper [Tan *et al.*, 2024b]):

1. ChatGPT-guided Concept Augmentation:

- *Concept Set Augmentation:* We prompt ChatGPT, using human-specified concepts ($\mathcal{C}_s$) as examples (in-context

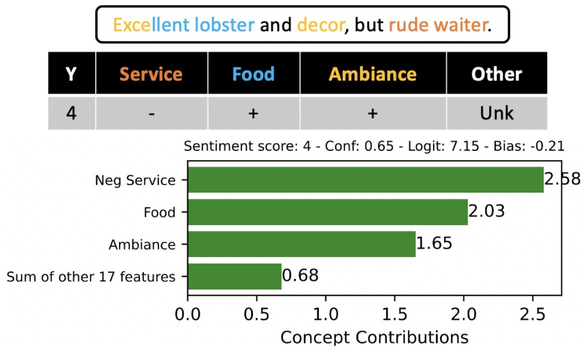


Figure 2: Illustration of the explainable prediction for a toy example in restaurant review sentiment analysis.

learning), to generate additional relevant concepts ($\mathcal{C}_a$). This expands the concept space.

- *Noisy Concept Label Annotation:* We use ChatGPT with few-shot prompting to generate pseudo-labels ($\tilde{c}_{sa}$ or $\tilde{c}_u$) for all concepts across both $\mathcal{D}_s$ and $\mathcal{D}_u$. This creates an augmented dataset $\tilde{\mathcal{D}}$ with noisy but comprehensive concept labels.

2. Concept-level MixUp (CM): Directly training on $\tilde{\mathcal{D}}$ treats clean and noisy labels equally, potentially harming performance. CM addresses this by encouraging linear behavior between examples. It interpolates latent representations, concept labels, and task labels between pairs of instances sampled from $\tilde{\mathcal{D}}_{sa}$ (containing original human labels c_s) and the shuffled full dataset $\mathcal{W} = \text{Shuffle}(\tilde{\mathcal{D}})$. The interpolated values $(\hat{z}^{(i,j)}, \hat{c}^{(i,j)}, \hat{y}^{(i,j)})$ are calculated using a mixing coefficient $\hat{\lambda} = \max(\lambda, 1-\lambda)$ where $\lambda \sim \text{Beta}(\alpha, \alpha)$. This generates mixed instances for training. The final loss $L_{\text{jointMixUp}}$ combines the standard joint CBM loss applied to these mixed instances, weighted by a factor τ. This allows the model to learn robustly from the noisy signals provided by the LLM while leveraging high-quality human annotations.

4 Experimental Highlights

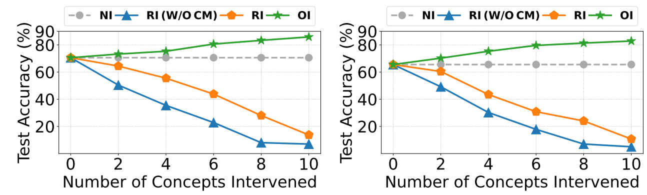
We evaluated CBE-PLMs trained with C^3M on sentiment classification tasks using the CEBAB dataset and a curated IMDB-C dataset (based on IMDB), using PLM backbones like BERT, RoBERTa, and GPT2. We compared against standard PLMs and baseline CBE-PLMs trained without CM.

Key findings (conceptualized in Table 1 based on [Tan *et al.*, 2024b]):

- **Interpretability with High Utility:** CBE-PLMs provide concept-level interpretability with competitive task performance compared to standard PLMs. Smaller models like LSTM even showed improved task accuracy, suggesting the trade-off is not necessary.
- **Effectiveness of C^3M:** Our framework (CBE-PLM-CM) consistently achieved the best concept prediction accuracy (interpretability). It significantly boosted performance, especially on the small IMDB-C dataset, by leveraging noisy labels effectively. C^3M maintained or improved task accuracy compared to standard PLMs, demonstrating an excellent interpretability-utility trade-off. Concept-level MixUp (CM) proved essential for robustness against noisy labels, preventing performance degradation seen when naively using augmented data.
- **Explainable Predictions:** CBE-PLMs allow visualizing concept contributions to the final prediction, offering intuitive insights as shown in Figure 2).
- **Test-time Intervention:** Users can correct mispredicted concept activations at test time to potentially improve task accuracy. C^3M significantly enhanced the effectiveness and robustness of this intervention, mitigating negative impacts from potential incorrect human corrections (see Figure 3).

Table 1: Representative Results Summary (Conceptual - Adapted from Table 1 in [Tan *et al.*, 2024b]). Comparing standard PLM, baseline CBE-PLM, and our CBE-PLM-CM (C^3M). Metrics: Task Acc/F1, Concept Acc/F1 (higher is better). C^3M improves concept accuracy (interpretability) significantly while maintaining or improving task accuracy.

Dataset	Model Type	CEBAB ($\tilde{\mathcal{D}}$) Concept Acc/F1 ↑	CEBAB ($\tilde{\mathcal{D}}$) Task Acc/F1 ↑	IMDB-C ($\tilde{\mathcal{D}}$) Concept Acc/F1 ↑	IMDB-C ($\tilde{\mathcal{D}}$) Task Acc/F1 ↑
CEBAB	PLM (BERT)	- / -	80.5 / 78.4	- / -	98.9 / 98.7
	CBE-PLM (BERT)	68.2 / 78.1	77.4 / 74.6	67.3 / 79.2	97.6 / 97.6
	CBE-PLM-CM (BERT)	**70.6 / 80.1**	**94.4 / 93.3**	**70.1 / 79.9**	**98.2 / 98.1**
IMDB-C	PLM (RoBERTa)	- / -	84.1 / 82.5	- / -	99.1 / 99.1
	CBE-PLM (RoBERTa)	69.9 / 79.3	82.3 / 80.1	71.0 / 79.9	98.5 / 98.1
	CBE-PLM-CM (RoBERTa)	**72.9 / 81.9**	**96.3 / 98.5**	**72.9 / 81.9**	**99.7 / 99.7**



(a) BERT (b) GPT2

Figure 3: The results of Test-time Intervention. "NI" denotes "no intervention", "RI (W/O CM)" denotes "random intervention on CBE-PLMs without the concept-level MixUp", "RI" denotes "random intervention on CBE-PLMs", and "OI" denotes "oracle intervention".

5 Conclusion

This work introduce Concept-Bottleneck-Enabled PLMs (CBE-PLMs) as a way to bring concept-level interpretability to complex language models. We propose the C^3M framework to effectively train these models by leveraging LLMs for concept augmentation and pseudo-labeling, combined with a concept-level MixUp strategy for noise robustness. Our approach yields models that are not only more interpretable through concept activations and visualizations but also maintain high task performance and benefit from test-time intervention. Our follow-up works include discussions on providing both local and global explanations [Tan *et al.*, 2024a], enabling autonomous test-time interventions [Tan *et al.*, 2025a], the faithfulness of post-hoc explanations [Tan *et al.*, 2025b], and the intrinsic barriers to explanations [Tan and Liu, 2025]. We hope our methods offer a practical path towards building more transparent, trustworthy, and interactive PLMs by effectively utilizing both limited human knowledge and large-scale AI capabilities.

Acknowledgements

This work is supported by the National Science Foundation (NSF) under grants IIS-2229461.

References

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019.

Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Shizhe Diao, Zhichao Huang, Ruijia Xu, Xuechun Li, Yong Lin, Xiao Zhou, and Tong Zhang. Black-box prompt learning for pre-trained language models. *arXiv preprint arXiv:2201.08531*, 2022.

Erik Englesson and Hossein Azizpour. Generalized jensen-shannon divergence loss for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:30284–30297, 2021.

Andrea Galassi, Marco Lippi, and Paolo Torroni. Attention in natural language processing. *IEEE transactions on neural networks and learning systems*, 32(10):4291–4308, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

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

Yang Liu, Hao Cheng, and Kun Zhang. Identifiability of label noise transition matrix. *arXiv preprint arXiv:2202.02016*, 2022.

Max Losch, Mario Fritz, and Bernt Schiele. Interpretability beyond classification output: Semantic bottleneck networks. *arXiv preprint arXiv:1907.10882*, 2019.

Andreas Madsen, Siva Reddy, and Sarath Chandar. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55(8):1–42, 2022.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

Renáta Németh, Domonkos Sik, and Fanni Máté. Machine learning of concepts hard even for humans: The case of online depression forums. *International Journal of Qualitative Methods*, 19:1609406920949338, 2020.

OpenAI. Gpt-4 technical report, 2023.

Alexis Ross, Ana Marasović, and Matthew E Peters. Explaining nlp models via minimal contrastive editing (mice). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3840–3852, 2021.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Zhen Tan and Huan Liu. Intrinsic barriers to explaining deep foundation models, 2025.

Zhen Tan, Tianlong Chen, Zhenyu Zhang, and Huan Liu. 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, 2024.

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

Zhen Tan, Jie Peng, Song Wang, Lijie Hu, Tianlong Chen, and Huan Liu. Tuning-free accountable intervention for llm deployment–a metacognitive approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25237–25245, 2025.

Zhen Tan, Song Wang, Yifan Li, Yu Kong, Jundong Li, Tianlong Chen, and Huan Liu. Are we merely justifying results ex post facto? quantifying explanatory inversion in post-hoc model explanations. *arXiv preprint arXiv:2504.08919*, 2025.

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

Kayo Yin and Graham Neubig. Interpreting language models with contrastive explanations. *arXiv preprint arXiv:2202.10419*, 2022.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lio, et al. Concept embedding models. In *NeurIPS 2022-36th Conference on Neural Information Processing Systems*, 2022.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. A survey on aspect-based sentiment analysis: tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. Incorporating bert into neural machine translation. In *International Conference on Learning Representations*, 2020.