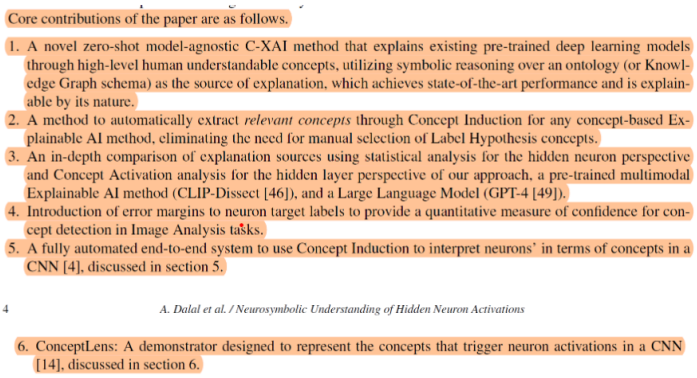
Reviewer 1:

State of the art is in places a list of approaches without a clear indication of what is missing in those approaches or the advantage of the proposed framework for understanding neuron activations via concept induction. Some of the approaches listed in Sec 2 are improving on existing comparative baseline such as CBM, but it is not clear why concept induction would be a better way of identifying relevant concepts than SAM or Q-SENN or Label-free CBM. This should be clear from the beginniing while at the moment it is only reported in a scattered manner after the evaluation experiments. A table illustrating a more quantitative comparative analysis with the desirable features and how they are missing (or not) in recent approaches to concept attribution would clearly help.	We appreciate the reviewer's feedback and agree that our manuscript should more explicitly highlight why our method provides a superior alternative. In the revised manuscript, we have added a paragraph at the end of Section 2 that clearly outlines these points. This revision directly addresses the reviewer's concern by clarifying the specific gaps and highlighting our neurosymbolic framework's distinct advantages. Regarding the suggestion to include a quantitative comparative table, we considered this option carefully. We found that the narrative format in Section 2 allows us to capture nuanced qualitative differences (e.g., in transparency, scalability, and the dynamic nature of learned concepts) that a fixed table might miss. Nonetheless, we remain open to adding a supplementary summary table if the reviewer believes it would enhance clarity. We believe this addition clearly and succinctly demonstrates why concept induction is a better method for identifying relevant concepts in our study. Recent trends highlight the potential of large language models (LLMs) to bridge the gap between model complexity and human-understandable explanations. LLMs like GPT-3 and GPT-4 have been used in few-shot learning contexts to generate concepts with minimal human intervention [47], providing a scalable solution to automated concept discovery. However, these approaches still require post-processing to filter and refine generated concepts for practical use [12, 70]. While LLMs show promise in automating concept generation, challenges remain in aligning explanations with human common sense and ensuring that they cater to diverse user needs, whether system developers or end-users. Our approach distinguishes itself by leveraging symbolic deductive reasoning over a comprehensive background knowledge base derived from Wikipedia, comprising approximately 2 million interconnected classes to generate explanations. Unlike methods that depend on manual selection or post hoc filtering of candidate concepts, our framework systematically extracts human-understandable labels directly from the underlying data, reducing potential biases and ensuring scalability. Moreover, by operating as a white-box system, Concept Induction provides inherent transparency; each explanation can be traced back to logical reasoning steps, which contrasts with black-box methods as discussed above, while effective, do not reveal the underlying rationale behind their output. In this way, our approach not only offers improved interpretability but also facilitates a more scalable and systematic framework for understanding and comparing neural activations.
It would be good to more clearly indicate what is the exact addition of this paper to all previous conference contributions, as opposed to just saying it provides a joint perspective, discussion and a demonstrator that was only pre-print: what experiments are new/not previously reported? What exact components are used to bridge these separate conference contributions in this paper?	In our manuscript, we have clearly stated the motivation and contributions in the Introduction. In particular, we provide a bullet-point list (page 3 & 4) of our core contributions, which include:

	 Core contributions of the paper are as follows. 1. A novel zero-shot model-agnostic C-XAI method that explains existing pre-trained deep learning models through high-level human understandable concepts, utilizing symbolic reasoning over an ontology (or Knowledge Graph schema) as the source of explanation, which achieves state-of-the-art performance and is explainable by its nature. 2. A method to automatically extract *relevant concepts* through Concept Induction for any concept-based Explainable AI method, eliminating the need for manual selection of Label Hypothesis concepts. 3. An in-depth comparison of explanation sources using statistical analysis for the hidden neuron perspective and Concept Activation analysis for the hidden layer perspective of our approach, a pre-trained multimodal Explainable AI method (CLIP-Dissect [46]), and a Large Language Model (GPT-4 [49]). 4. Introduction of error margins to neuron target labels to provide a quantitative measure of confidence for concept detection in Image Analysis tasks. 5. A fully automated end-to-end system to use Concept Induction to interpret neurons' in terms of concepts in a CNN [4], discussed in section 5. 4 A. Dalal et al. / Neurosymbolic Understanding of Hidden Neuron Activations 6. ConceptLens: A demonstrator designed to represent the concepts that trigger neuron activations in a CNN [14], discussed in section 6. As stated in our manuscript, in this extended version, we have bridged our previous contributions by integrating these distinct lines of work into a unified framework. In addition to reiterating our core contributions, the current paper introduces several new elements: ● A joint perspective that unifies the central narrative and finer-grained analyses from our earlier work. ● An expanded literature review and discussion that contextualize our approach within the broader XAI landscape. ● New experimental evaluations and a demonstrator system (ConceptLens) showcase the practical applicability of our method. This paper is an extended merger of several conference contributions: [16] is the central one for the overall narrative; [15] is an extension with a finer-grained analysis; [7] goes in detail on using LLMs as an alternative to concept induction; [4] reports on our automation of the analysis process (see Section 5); This paper extends these by providing a joint perspective, additional literature review, more discussion, and a demonstrator system (see Section 6) previously only reported as a pre-print [14]. We believe that these clarifications and the extended merger of our previous work not only emphasize the novelty of the current submission but also provide a more comprehensive understanding of our approach. We are, however, open to further clarifying these aspects if required.
Overall this line of research is an interesting direction for investigation although it would still be difficult to transfer it to specific domains such as health and diagnostics, without considerable effort, since the generality of concepts would not fit the specific domain. What would be required for such domain adaptation? Also in terms of evaluation and benchmark, how could this represent a benchmark for other domains?	We thank the reviewer for raising an important point regarding the domain adaptation of our approach. In our manuscript (page 27 line 30), we explicitly state that although our current experiments are conducted on a CNN architecture with ADE20K Image data using a Wikipedia-derived concept hierarchy, the model-agnostic nature of our framework lends itself to adaptation across various domains. For example, in specialized domains such as health or diagnostics, a domain-specific Knowledge Graph—curated with relevant medical terminologies and relationships—would replace the generic Wikipedia hierarchy. This substitution would likely

	require additional efforts in knowledge curation and fine-tuning to capture the nuances of the domain. Furthermore, the evaluation framework we propose (including statistical analyses and concept activation benchmarks) can be directly adapted to assess the quality and relevance of explanations in these new contexts, thereby serving as a valuable benchmark for interpretability across diverse application areas. We believe that the model-agnostic nature of our framework serves as a promising starting point, and we plan to explore such domain adaptations in future work through collaborations with domain experts. Our focus has been primarily on assessing the comparative effectiveness of Concept Induction within the confines of Convolutional Neural Network architecture using ADE20K Image data. Nevertheless, it is imperative to investigate its suitability across different architectures and with diverse datasets. Given the model-agnostic nature of our approach, our results suggest its potential applicability across a range of neural network architectures, datasets, and modalities. While we utilized a Wikipedia Concept Hierarchy comprising 2 million concepts, it would be intriguing to observe the outcomes of our approach when powered by a domain-specific Knowledge Graph in specialized domains such as Medical Diagnosis.
Selection of thresholds seems to be experimentally determined (e.g. error margin threshold). Can authors suggest a specific rationale for the choice of values?	In our experiments, we selected threshold values—such as the 80% cutoff for positive activations and the 20% cutoff for negatives—to robustly separate images that strongly activate a neuron from those with minimal activation. This selection ensures that the positive set predominantly comprises images in which the target concept is clearly expressed, while the negative set includes images with little to no activation, thereby reducing false positives. Furthermore, employing multiple thresholds (e.g., >20%, >40%, and >60%) in our error margin analysis allows us to systematically assess how different activation intensities influence the reliability of concept detection. These choices represent our best guess for balancing sensitivity and specificity, and although heuristic, they strike an appropriate balance between recall and precision, as supported by our statistical evaluations. Our approach to selecting target labels has already been stated in the manuscript. We have now clarified this rationale in the revised manuscript. *A. Dalal et al. / Neurosymbolic Understanding of Hidden Neuron Activations* 13 We define a target label for a neuron to be *confirmed* if it activates for $\geqslant 80\%$ of its target images regardless of how much or how often it activates for non-target images. The cut-offs for neuron activation and label hypothesis confirmation are chosen to ensure strong association and responsiveness to images retrieved under the target label, but 80% is somewhat arbitrary and could be chosen differently. For each neuron, we calculate the maximum activation value across all images. We then take the positive example set P to consist of all images that activate the neuron with *at least* 80% of the maximum activation value, and the negative example set N to consist of all images that activate the neuron with *at most* 20% of the maximum activation value (or do not activate it at all). We selected these thresholds as our best guess (further refinement may be possible in future) based on experimental observations to ensure that the positive set is predominantly comprised of images in which the target concept is clearly expressed, while the negative set is limited to images with minimal or no activation, thereby reducing overlap and enhancing the reliability of the subsequent concept extraction. The *Computation of Non-TLA* Concept Induction generates a number of concept labels for each neuron unit, ranked by some accuracy measure. Herein, we consider the Top-3 labels (ranked by coverage score) for each of the 64 neurons in the dense layer. Using the Target-Label image dataset (each image falls under the target label), the TLA is calculated, and, using a Non-target Label image dataset (none of the images contain the target label), the Non-TLA is calculated. To obtain a nuanced understanding of how activation levels affect the reliability of the neuron–concept association, we calculate TLA and Non-TLA for each neuron at specified activation value thresholds, namely > 0, $> 20\%$, $> 40\%$, and $> 60\%$ of the max activation value that was recorded for that neuron. These thresholds are our best guess for balancing sensitivity and specificity, and we acknowledge they are heuristic and may be refined in future work. For example, (see Table 10), neuron 43 activates at $> 40\%$ of its max activation value in about 19.7%

A brief comparison with disentangled representation approaches is missing, for example those where hidden units are associated with concepts via Network Dissection and work building on this approach.	Our revised manuscript now includes a concise paragraph in Section 2 that contrasts our approach with methods like Network Dissection. While Network Dissection maps hidden units to semantic concepts using manually curated labels, it does not capture the full hierarchical and dynamic nature of learned representations nor incorporate an explicit reasoning process. In contrast, our method leverages symbolic deductive reasoning over a large-scale Wikipedia-derived knowledge base (≈2 million classes) to automatically extract human-understandable labels, eliminating the need for manual candidate selection and yielding transparent, white-box explanations that can be quantitatively evaluated. We believe this addition clearly addresses the reviewer's concern. The application of background knowledge, including the use of large ontologies, has been explored to generate more automated and systematic explanations. Semantic Web technologies [11, 17] and methods like Concept Induction [51, 56] have demonstrated the utility of formal logic and structured data to explain deep learning models, though these approaches often focus on input-output relationships rather than internal model activations. While methods such as Network Dissection (e.g., [75]) provide valuable insights by mapping hidden units with semantic concepts by comparing neuron activations against a pre-defined set of labels (typically derived from human-annotated datasets), they do not capture the full hierarchical and dynamic nature of learned concepts, nor do they incorporate an explicit reasoning process. Notably, CLIP-Dissect [46] employs zero-shot learning to associate im-
Typos/notes: Is be (page 3 line 11) CCN training (should be CNN training) (page 5 line 38) Responses to be returns (should be returnED) (page 10 line 36) The conceptLens page link (footnote 7 page 37) seems to have problems loading so none of the operations showed in the demo video can be performed.	Corrected.
For the sake of readability and focus, I think Section 4 represents a different investigation/analysis than the one presented as a core in the paper, and should therefore be a separate submission. The paper is way too long as a 42 pages and there is a neat split between the first 28 pages and the direction in which LLM is replacing ECII concept induction. The evaluation is different (humans are used here) and the goal is different (producing and evaluating explanations for humans). Unlike Section 5 and 6, where a tool is presented that follows the steps and methods provided in the first 28 pages, Section 4 is a clear diversion so I suggest removing it.	We appreciate the reviewer's feedback on readability and focus. While Section 4 presents a distinct analysis, it provides valuable insights into the trade-offs between white-box and black-box explanation approaches, particularly from a human-centered perspective. Given its relevance to concept-based explainability, we believe it complements the main study.

Reviewer 2:

The introduction is good, although it reads quite long. If keeping it so, it may be good to clearly state the problem, research question and goal of the paper as a very first thing (eg a first paragraph of what is coming), and then deepdive into the rest. A working example is often a good idea. Also, considering the length of the paper, it may be good to have a coincise description of what goes where and how can the reader go through.	We appreciate the reviewer's insightful comments regarding the introduction's structure. While the introduction is relatively long, it is necessary to establish the foundation for our interdisciplinary work, which integrates Explainable AI, Neurosymbolic AI, and Concept Induction. ● The problem and motivation are introduced gradually (page 2, lines 26 and 36) to provide context before presenting our Concept Induction-based method. Stating the research question too early may make it harder for readers to grasp the underlying challenges. ● A concrete example is useful but introducing it before key concepts may reduce clarity. Instead, we progressively build toward examples in Sections 3 and 4, where they align naturally with our method and evaluation. ● A roadmap is already included (page 4, line 7) to guide the reader through the structure of the paper. While we acknowledge the comment on length, we believe the current structure is well-justified given the topic's complexity. However, we are open to further refinements if needed.
The related work section is complete to my (up to a certain point expert) knowledge, but at the end of the section I am a bit left unclear on how the paper stands out. I would suggest to clearly state, for each body of work, how the presented paper improves upon.	We appreciate the reviewer's feedback. We have revised the Related Work section to explicitly highlight how our approach improves upon existing methods. Specifically, we now: Clearly state how our method differs from other approaches, Emphasize our use of symbolic deductive reasoning, and Contrast our white-box approach with black-box models ensuring greater interpretability and transparency. These refinements (page 5, line 37) clarify our contributions, and we welcome any further suggestions.
Minors: - When referring to specific sections >> "Section" with capital 's' (Section 4, Section 5, etc) - Section 2 : called it 'related work' ? (literature review may be misleading, as one would think of a systematic review of the field) - Section 3.1 : "Preliminaries" ? - For opening quotes : use this character 2 times in latex : ` >> eg ``kitchen'' (to close quotes : 2 times this char : ')	Corrected

Reviewer 3:

The literature review would benefit from being presented in a table/graph form, comparing the main axes (such as the neural/symbolic nature, the degree of supervision required, etc.) and possibly proposing a proper taxonomy of the works cited.	We thank the reviewer for the valuable suggestion to present the literature review in a table/graph form. We considered this option; however, we opted for a narrative format in order to provide detailed qualitative insights into the strengths and limitations of each method. Our approach allows us to discuss nuances—such as the degree of supervision required, the dynamic versus static nature of the concept pools, and the neural versus symbolic components—in a cohesive, contextual manner that a table might not capture. That said, we are open to providing a supplementary summary table if the reviewer believes it is needed to further enhance clarity.
Also, the relationship between the related works cited and the present work could be discussed further, e.g., by discussing under which aspect your proposal improves the weaknesses of each method.	We thank the reviewer for this insightful suggestion. In response, we have expanded Section 2 to further discuss the relationship between the cited works and our own approach. In the revised text, we now clearly explain how our proposal improves upon the weaknesses of each method. For example, we highlight that while many approaches rely on manually curated or static concept pools, our method leverages symbolic deductive reasoning over a background knowledge base, which offers both scalability and inherent transparency. We also explain that our white-box approach provides explicit reasoning steps, in contrast to black-box methods that lack interpretability. We believe these revisions provide a more comprehensive understanding of how our work advances the state of the art by directly addressing the limitations observed in related works. Recent trends highlight the potential of large language models (LLMs) to bridge the gap between model complexity and human-understandable explanations. LLMs like GPT-3 and GPT-4 have been used in few-shot learning contexts to generate concepts with minimal human intervention [47], providing a scalable solution to automated concept discovery. However, these approaches still require post-processing to filter and refine generated concepts for practical use [12, 70]. While LLMs show promise in automating concept generation, challenges remain in aligning explanations with human common sense and ensuring that they cater to diverse user needs, whether system developers or end-users. Our approach distinguishes itself by leveraging symbolic deductive reasoning over a comprehensive background knowledge base derived from Wikipedia, comprising approximately 2 million interconnected classes to generate explanations. Unlike methods that depend on manual selection or post hoc filtering of candidate concepts, our framework systematically extracts human-understandable labels directly from the underlying data, reducing potential biases and ensuring scalability. Moreover, by operating as a white-box system, Concept Induction provides inherent transparency; each explanation can be traced back to logical reasoning steps, which contrasts with black-box methods as discussed above, while effective, do not reveal the underlying rationale behind their output. In this way, our approach not only offers improved interpretability but also facilitates a more scalable and systematic framework for understanding and comparing neural activations.
Regarding the discussion of “explaining a neural network through concepts” (cfr. p3,r11), reporting some works related to “having a neuron active for many concepts at once” could be beneficial. To this extent, the literature on disentangled representations (Bengio et al., 2013; Locatello et al., 2019) could be useful. Other useful keywords are “polysemantic neurons” (i.e., neurons that fire under multiple	In our revised Section 2, we already briefly discuss Network Dissection in the context of disentangled representations and explain that our approach is distinct because it uses symbolic deductive reasoning over a large-scale knowledge base to automatically extract human-understandable labels. Additionally, our analysis in Section 3.3.3 on neuron ensembles implicitly captures the polysemantic nature of neurons. While we acknowledge the existence of literature on disentangled representations and polysemantic neurons, our

stimuli).	primary focus is on generating transparent, reasoning-based explanations—a goal that is not directly addressed by those methods.
While I understand the utility of having the notions related to each section structured to give the background needed at the beginning of each section, some common preliminary notions could be moved to a background section before entering Section 3. This section could also help provide a visual example to help understand all the inputs/outputs involved in the system. In my opinion, this would help to make the paper less of a collection of existing published papers and more of a comprehensive work on the topic.	Figure 1 on page 6 provides a comprehensive overview of our entire pipeline, clearly illustrating all inputs and outputs involved in the system. We introduce the necessary background concepts in context, ensuring a logical flow throughout the manuscript. This approach allows readers to engage with key notions at the point of use, enhancing clarity and comprehension. However, we are open to incorporating a brief background subsection if the reviewer believes it is necessary to further improve readability and can be done without disrupting the narrative flow.
(cfr. p26,r36) It is quite strange that only the Resnet50V2 achieved high validation accuracy scores, while other architectures show a big gap with the training accuracy, especially when using early stopping. Do other metrics highlight this issue (e.g., top-k accuracy) as well? Could you compare the confusion matrices? Also, is patience=3 / learning rate=0.001 sufficient/necessary to fine-tune this task? Usually, you could get better results in fine-tuning with lower learning rates and/or providing more epochs. While I understand the argument of the low need for high accuracy, the explanations should be made on a sufficiently reliable/performant model, and I can't see how Resnet50v2 has such a wide margin compared to the classic Resnet50.	We thank the reviewer for the valuable feedback regarding model performance and hyperparameter choices. In our extensive experiments, we evaluated several architectures (including VGG16, InceptionV3, Resnet50, Resnet50V2, Resnet101, and Resnet152V2) and tested various hyperparameter configurations (different learning rates, patience values, and number of epochs). Ultimately, Resnet50V2 achieved the best overall performance, with consistent training and validation accuracy levels. Although our primary focus is on generating and interpreting explanations rather than maximizing classification accuracy, the 87% validation accuracy achieved by Resnet50V2 is robust enough for our purposes. The choices of a patience of 3 and a learning rate of 0.001 were derived from extensive preliminary tuning; while further fine-tuning (e.g., lower learning rates or more epochs) might yield incremental improvements, such modifications were not necessary given that our task prioritizes explanation fidelity. Overall, our current approach strikes a sufficient balance between model performance and the reliability of the generated explanations.
Regarding the statistical testing: in p13,r23 you state the usage of the Mann-Whitney U test that does not require normal distributions. It is unclear to me whether this test should be corrected or not (due to the multiple analyses performed) and why. Also, you mention there is no reason to assume that activation values follow a normal distribution; can you show an example?	We thank the reviewer for raising these important points regarding our statistical testing. To address them: ● Multiple Comparisons: We employed the Mann-Whitney U test because it does not assume normality of the data—a key advantage given the nature of neuron activations. Although multiple hypothesis tests can increase the risk of Type I errors, our observed p-values are extremely small (often <0.00001). Even if a conservative correction (such as Bonferroni) were applied, the corrected p-values would remain highly significant. We therefore believe that a formal correction would not alter our conclusions. ● Non-Normality of Activation Values: Activation values in deep neural networks, particularly those

	derived from non-linear functions (e.g., ReLU), often exhibit skewed distributions. For example, as reported in Table 5, neuron 1 shows a mean activation of 4.17 for target images but a median of 4.13, while for non-target images, the mean is 0.67 and the median is 0.00. This pronounced difference between mean and median is indicative of a skewed (non-normal) distribution, thereby justifying the use of a non-parametric test such as the Mann-Whitney U test. We hope this explanation clarifies our rationale and demonstrates that our statistical analyses are both appropriate and robust.
I am not sure of the usefulness of Table 6-7-8. In particular, they show the raw performance in both training and test settings. Wouldn't a chart be more informative, especially while comparing the results of GPT/CLIP/Concept Induction? Those tables could be moved to an Appendix if possible. Also, I am unsure of the utility of having the training accuracy reported as well, if not discussed in the paper.	Thank you for the suggestion. We considered adding charts, however we were unable to come up with a good and meaningful way to visualize the data in the tables without adding unnecessary redundancy and length. We would be happy to receive concrete suggestions.
Regarding the "Further discussion" subsection, there are a couple of claims that could be discussed better: 9a. P27,r3: "it is unclear how to craft the pool of candidate concepts"; can you expand on this topic? 9b. P27,r5: "tailored to the application scenario"; can you provide an example? 9c. P27,r9: "it is equally vital to thoughtfully design this pool"; could you better explain what are the risks of a poorly designed pool? 10. How would this extend to other datasets? Can you make an example?	We thank the reviewer for the detailed suggestions regarding the "Further Discussion" subsection. In response, we have expanded this section to clarify certain points: ● P27, r3: creating an effective candidate concept pool is challenging because it must be both broad enough to capture a wide range of relevant concepts and sufficiently structured to filter out irrelevant or overly generic terms. ● P27, r5: For instance, in a medical diagnostic application, a tailored candidate pool would include clinical terminology (e.g., "cardiomegaly," "pleural effusion") and relationships specific to medical imaging, thereby capturing the nuances necessary for accurate interpretation. ● P27, r9: Manuscript says - it is equally vital to thoughtfully design this pool. Neglecting this aspect could result in overlooking crucial concepts essential for gaining insights into hidden layer computations. Our approach offers a way to integrate rich background knowledge and extract meaningful concepts from it. [illegible] within a given candidate pool, it is equally, if not more, vital to thoughtfully design this pool. Neglecting this aspect could result in overlooking crucial concepts essential for gaining insights into hidden layer computations. Our approach offers a way to integrate rich background knowledge and extract meaningful concepts from it.

	If an application does not require comprehensive concept-based explanations, CLIP-Dissect/GPT-4 may serve as a useful solution, especially when time is limited. However, for detailed concept-based analysis, preparing background knowledge and leveraging Concept Induction is crucial. For CLIP-Dissect/GPT-4, it is unclear how to meticulously craft the pool of candidate concepts since it is difficult to manually curate a static set that is broad enough to capture all pertinent concepts while remaining specific enough to avoid noisy or ambiguous labels. By employing a background knowledge base, it is possible to define a large pool of potential explanations, tailored to the application scenario, with additional relationships among concepts. For example, in a medical diagnostic application, an ideal candidate pool would include specialized clinical terminologies (e.g., "cardiomegaly" or "pleural effusion") that are essential for accurate interpretation—an adjustment that is hard to achieve with a generic vocabulary. Concept Our focus has been primarily on assessing the comparative effectiveness of Concept Induction within the confines of Convolutional Neural Network architecture using ADE20K Image data. Nevertheless, it is imperative to investigate its suitability across different architectures and with diverse datasets. Given the model-agnostic nature of our approach, our results suggest its potential applicability across a range of neural network architectures, datasets, and modalities. While we utilized a Wikipedia Concept Hierarchy comprising 2 million concepts, it would be intriguing to observe the outcomes of our approach when powered by a domain-specific Knowledge Graph in specialized domains such as Medical Diagnosis.
The limitations of the work could be summed up in a specific section at the end of the paper (e.g.: activation patterns involving more than one neuron, requirement of labeled data, single dataset analysis, concept formation across multiple layers). Mitigations and/or suggestions for implementing these improvements could be reported as well.	In our manuscript, the "Further Discussion" section already addresses many of these points. For instance, we acknowledge (1) the model-agnostic nature of our approach, (2) the possibility of extending our method to other datasets, and (3) the potential use of domain-specific knowledge graphs for specialized applications such as medical diagnosis. As for the question of activation patterns involving multiple neurons, we do not regard this as a limitation of our approach; rather, it reflects the intrinsic nature of neural networks, where concepts can be distributed across multiple units. In fact, our work explicitly examines neuron ensembles in Section 3.3.3 to account for these distributed activations. If the reviewer deems it necessary, we can add a short "Limitations and Future Work" section to summarize these points more explicitly. However, we believe our existing "Further Discussion" already captures the essence of these limitations (e.g., focusing on the dense layer, the need for labeled data, single dataset use) and outlines how we plan to address them in future research. Our focus on dense layer activations, while providing valuable insights, represents only a part of what the deep representation encodes. The dense layer likely relates to clear-cut concepts that separate output classes, aligning well with our goal of identifying high-level, interpretable concepts. However, these concepts are influenced by combinations of features from previous layers. This limitation underscores the complex nature of deep neural networks, where concepts identified at the dense layer result from hierarchical feature compositions throughout the network. While our method offers meaningful insights into these high-level concepts, it may not fully capture the nuanced feature interactions in earlier layers. Nonetheless, focusing on the dense layer allows us to extract concepts more directly relevant to the network's final decision-making process, balancing interpretability with the complexity of internal representations. Future work could explore extending our method to analyze concept formation across multiple layers, potentially revealing a more comprehensive picture of the network's decision-making process. One drawback of utilizing Concept Induction (and GPT-4) is its dependency on object annotations, which serve as data points in the background knowledge. In contrast, CLIP-Dissect operates without the need for labels and can function with any provided set of images. Our focus has been primarily on assessing the comparative effectiveness of Concept Induction within the confines of Convolutional Neural Network architecture using ADE20K Image data. Nevertheless, it is imperative to investigate its suitability across different architectures and with diverse datasets. Given the model-agnostic nature of our approach, our results suggest its potential applicability across a range of neural network architectures, datasets, and modalities. While we utilized a Wikipedia Concept Hierarchy comprising 2 million concepts, it would be intriguing to observe the outcomes of our approach when powered by a domain-specific Knowledge Graph in specialized domains such as Medical Diagnosis.
Minors:	We have now included a brief definition of the Levenshtein string

1. The Levenshtein string similarity metric is undefined (p29,r41)	similarity metric in the revised version of the paper.
Grammar and general layout: 1. P3,r11: “Neural Network through concepts is be a two-step process” -> “[...] is a two-step process” 2. p5,r43: should have a brief discussion before creating the subsubsection 3.1.1, to avoid the empty subsection. 3. P20,r28: k-fold cross validation vs p22,r37, K-fold cross validation; keep a consistent notation 4. P37,r44: necessitate -> necessitates 5. P29,r3: beforew -> before	P3, r11: corrected P3, r11: added P20, r28: corrected P37, r44:corrected P29, r3:corrected