Rebuttal

Thank you to all the reviewers for your helpful comments. We appreciate that the reviewers found our data-centric approach to capturing uncertainty by rethinking ground truth annotations to be "interesting" (R2, R3), "qualitatively better" (R1), and "useful in real-world applications" (R1). Many of the responses focused on the more technical, computer-vision aspects. While this is undoubtedly important, our paper crucially brings attention to the importance of human factors in responsible and robust AI (especially medical), which need appropriate evaluative methods. Our data-centric approach provides, we feel, an important supplement to the model-centric research which dominates the field.

Results and Analysis. *User task load increase appears significant* (R2). Direct comparison of the task load factors across annotation methods can be misleading because a single CC annotation collects more absolute information than a single standard annotation. The phrasing that CCs do not "require significantly higher annotator effort" is admittedly unclear. In the paper, we will clarify that the overall increase in task load is, as R1 writes, "an acceptable extra workload for the annotators" given the additional amount of information acquired. We believe this is supported by the paper's analysis of Table 3 (§5.2¶1). • *Max contours may additionally identify lesions standard annotations do not* (R1). We agree with R1 that validating whether max contours did lead to higher recall of TP lesions would strengthen the case for a CC annotations. However, we had limited ability to directly validate this as: (1) we did not have access to domain experts that could provide high-confidence validation of the identification of lesions nor final diagnosis data; and (2) there is disagreement in the criteria for what is considered a TP lesion segment—i.e., whether a lesion segmentation is *visually* sound or ultimately *diagnostically* significant. §5.1 and §6¶1 relatedly show that max contours provide positive signals for more pixels *likely* to be lesions than standard annotations, though. • *Comparison with SOTA models is "mainly visual"* (R3). Our focus is proposing CCs as a novel annotation method; the modeling results validate its applicability in deployment settings. As CCs present an alternate representation of the ground truth, we must make assumptions when comparing against existing metrics that are based on inherently differently represented data. In this work, we make a best effort to do an approximately apples-to-apples comparison between our work and existing systems; §5.1 rigorously investigates the properties of these two ground truth representations, which affect downstream models. However, we agree with R3 that this evaluation can be limited. To address this and reproducibility concerns raised by R2, we will release the results of our annotation as a dataset so future work can conduct additional comparisons on the modeling front.

Contributions. *Images used already exist* (R4). The focus of our work is not on generating a novel set of images. Rather, we contribute a new annotation method and compare it with singular annotations on LIDC. • *No modeling method contributions* (R4). Our goal was not to make a novel modeling contribution. Rather, we tested and showed CC's compatibility with a wide array of already existing general segmentation models. We see this as a twofold strength of CCs. First, CCs demonstrate that effective uncertainty representations can be achieved not only with complex model-centric approaches but also with general models and relatively simple data-centric modifications. Second, because CCs are 'model-agnostic', existing segmentation explainability methods, training strategies, metrics, etc. also apply to CC-trained uncertainty-aware models, whereas their direct applicability to the aforementioned specialist models is much less clear.

Novelty. *There already exists similar work in uncertainty modeling and annotation* (R1, R2). Our work differs substantively from existing work in three ways: First, CCs do not derive uncertainty representations from singular boundaries (e.g., morphological dilation in Yeung et al. cited by R2 or probabilistic sampling in Phiseg). These approaches make certain assumptions about how uncertainty is distributed relative to the singular boundary, which can reflect errors in how the specific boundary is formed. By using two contours, CCs explicitly collect information on uncertainty *structural to the image* which do not adhere to the regularity of the previous assumptions (see Figure 5, CC prediction row for examples). Second, CCs focus on the human interpretation of the results and retain hard boundaries of each contour. Whereas other approaches use 'softening' procedures which yield specialized models whose uncertainty maps are unclear to interpret (see §2.1¶2, 3), modeling CCs is a straightforward and conventional segmentation problem, unlike many model-centric approaches. Third, CCs are more information-efficient than the standard annotation method, in that they enable single annotators to provide confidence bounds which would otherwise depend on multiple annotators' singular boundaries (§5.1), which opens up new opportunities for more fully utilizing specialists' expertise. In sum, other works in medical uncertainty-aware segmentation have overlooked the interpretation bottlenecks in representing the ground truth in a singular manner. In constrast, our work can begin inquiry into uncertainty-aware segmentation systems which both give annotators more control and end-users more transparency by critically examining how the data is annotated to begin with. As R1 states, the key novelty and significance of our work is in the "introduction of uncertainty directly into the input of the system" with CCs. We will make this clearer by clarifying our novel contributions in the related work section.

User Study Details. *IRB approval and reproducibility* (R2). We received IRB approval for our study and will include the IRB approval number in the paper (after deanonymization). We will also provide all relevant information (software, annotation prompts, recruitment) as supplemental material. Given that we will release the anonymized annotation dataset as well, we don't expect significant effort to be needed to reproduce the study. • *Justifying recruitment of undergraduate students* (R2). All undergraduate students recruited as participants for our study are pursuing studies in medical or biological science fields. As explained in our paper (§4.2¶1), we successfully adjusted the difficulty of the task without compromising the experimental focus to match the expertise of our participant group, with no observable difference in our annotations compared to the original annotations in the LIDC dataset.