We would like to thank the referee for their report on our manuscript; however, we disagree with their criticisms. We believe that the numerical results presented are new and interesting enough on their own to justify publication in PRD (though we did make several attempts at developing a heuristic understanding) and that the referee has several scientific misunderstandings about our manuscript in particular and the study of gravitational collapse in AdS in general. We enumerate our responses to the referee's points below with a short description of the revisions we made to our manuscript in each case.

1. The referee states that "the text is too long so it difficult to extract the main results/observations (which are not very new)." We are aware of many PRD articles that are longer than this manuscript, so we are not entirely sure of the referee's objection. However, we have added a small amount to the introduction section (penultimate full paragraph on page 2) emphasizing that a systematic study of stability features in a two-parameter phase space (mass and width) has never been done before. We also emphasize that our results include the first evidence for chaotic gravitational collapse in AdS for a system with only one length scale (when using AdS units), the massless scalar. The other chaotic systems — massive scalars, Einstein–Gauss-Bonnet gravity, two thin shells — all have an additional length scale. There are numerous other results/observations presented throughout the text; while unusual for a paper in formal theory, it is similar in style to a phenomenology paper.

2. The referee also states that our manuscript gives "the results of numerical calculations but it does not provide much of the explanation and insight into the problem of stability of AdS." This seems to miss our point, which is exploring stability behavior across parameter space, partly as a means of providing background for future analytical studies. Another main goal was to identify *nonperturbative* behaviors, which are intrinsically difficult to understand analytically. Nonetheless, we have attempted to find some heuristic understanding of the different phases in section IV.

3. The referee does not find our criteria for the distinction between the unstable and metastable phases to be clear. We have tried to give a rigorous definition; to clarify, we have added additional text to the first paragraph of section III.A. Both unstable and metastable initial data have t_H well fit by $a\epsilon^{-p} + b$ for evolutions with large enough t_H. The difference is that the best fit power p is statistically significantly different from 2 for metastable data. By this we mean that the best fit value is more than two standard errors from 2 when our fit is restricted either to $t_H > 60$ or $t_H > 80$ and more than one standard error from 2 when the fit is restricted to $t_H > 100$. The last criterion is relaxed because constraints on time and computational resources allow us to find only a small sample of such long evolutions for all the initial data we consider. We think that these criteria should now be clear.

4. It is not entirely clear from the report if the referee disagrees entirely with having a metastable phase as a classification at all. If the referee means that there is no sudden change in the order parameter p between unstable and metastable phases, we agree. It may be more reasonable to view the metastable phase instead as part of a second-order phase transition, and we have added a sentence to that effect in the last paragraph of section II.A.

5. The referee objects to the fact that we round $p \approx 5.6$ to 6 in the last paragraph of section III.A but do not round $p \approx 2.08$ to 2 in the previous paragraph. The reason we made the distinction is that the first instance is statistically equivalent to 6 (based on the quoted error), while the second instance is not statistically equivalent to 2. However, to be absolutely clear, we have changed the p values in the text to give the fit value to all significant digits as well as the errors. In addition, although quoting errors in parentheses is a fairly common notation, it may have contributed to this misunderstanding. We have therefore changed to a plus-minus notation for errors, eg, we have replaced 5.6(8) with 5.6 ± 0.8, throughout the entire manuscript. We have also added parenthetical comments indicating that values following ± are standard errors in the fit values at several points.

6. The referee seems to object to our classification of $\mu = 5, \sigma = 1.7$ initial data as metastable because the best fit value is very close to $p = 2$. While we agree that it seems to be a borderline case, we have to follow the mathematically precise definition that we adopted for the metastable phase vs the unstable phase. (It may be possible to adopt another definition that will change this sort of borderline classification, but we do not feel that is reason to reject the manuscript.) The referee suggests that we have mis-classified this initial data because of "a low resolution in amplitude." To check this hypothesis, we have added 5 new amplitudes with horizon formation times $60 < t_H < 300$. The fit restricting to $t_H > 60$ changed from $p \approx 2.08 \pm 0.02$ to $p \approx 2.07 \pm 0.02$, the fit restricting to $t_H > 80$ was unchanged, and the fit with $t_H > 100$ changes from $p \approx 2.10 \pm 0.04$ to $p \approx 2.11 \pm 0.03$. This is not a statistically significant difference, and the classification of the initial data does not change.

7. The referee also suggests that we have misclassified this initial data because t_H is only piecewise continuous as a function of ϵ; for massless scalars, there is a series of transition amplitudes below which a pulse must travel across AdS and back an additional time before collapsing. As the referee notes, the change in t_H is $\Delta t_H \sim \pi$, the crossing time for AdS, leading to "steps" of t_H as a function of ϵ. Δt_H is shorter for massive scalars, since the pulse cannot travel to the boundary; an initial pulse can also spread into a wider pulse, leading to even smaller values of Δt_H. However, in the perturbative regime (say for $t_H > 60$), it is not practical to measure the width of these steps numerically, and even slightly different amplitude values are typically separated by several steps in t_H. It is

therefore reasonable to fit to the perturbative scaling. We have added a paragraph to section II.A (third paragraph in the revision). Specifically considering the $\mu = 5, \sigma = 1.7$ initial data, the $\epsilon = 0.1$ and $\epsilon = 0.095$ points have $\Delta t_H \sim 12.5$, or 4 steps. That's an average step width of $\Delta\epsilon \sim 0.00125$, or $\sim 1.25\%$. If we naively take this to represent a systematic error of the same relative level that conspires to reduce the fit value of p, this initial data is almost precisely marginal for classification as unstable or metastable. However, we argue that, because the amplitudes we evolve are effectively located randomly in the steps of t_H, this effect is folded into to the statistical uncertainties of the fit. (One way to interpret the referee's comments is to agree with this.)

8. The referee does not like some of our figures. One comment is that figure 13a is not clear; its purpose is simply to show that this initial data is nonmonotonic and should do that at 100% scale in print (and can of course be magnified by electronic readers). Similar comments apply to the former figure 15a (now 16a). The referee also thinks figure 7 "do[es] not say much." We have replaced both subfigures of figure 7 with improved figures showing the evolutions at three times each to illustrate the movement of secondary pulses on top of the main pulse. Hopefully the referee finds more information content (and notes that we mention that we have studied the time evolution in more detail).

9. The referee states, "Given the plots showing results of convergence tests (in particular Fig. 16) I have my doubts about the numerics." We would like to point out that the referee is focusing on one convergence test with poor results while ignoring many successful convergence tests presented throughout the appendix, which validate all our key results on irregular initial data (the most difficult evolutions from a computational point of view). In particular, we show that the evolutions used to find approximately 2σ evidence for a non-zero Lyapunov coefficient in the massless scalar are already convergent with only a quarter as many grid points as we used, and we validate both non-monotonic and chaotic behavior for other scalar masses and initial pulse widths. As further validation of chaotic behavior, we have added convergence tests for the $\mu = 5, \sigma = 0.34$ amplitudes used to find a non-zero Lyapunov exponent at the 3σ level in table II. These convergence tests are discussed in the appendix and shown in the new figure 15. We also quote convergence testing from reference [34] (by two of us) demonstrating convergence for another evolution for $\mu = 20$ scalars. And even the evolutions shown in figure 16 is convergent for a significant fraction of their evolutions, so some of the results from those evolutions, such as the characteristic behavior of the spectral evolution shown in the early times of figure 10d, have also been validated by convergence tests. So the referee (or reader) may want to discount the particular evolution shown in figure 16, but they *must* acknowledge that we have validated our key results by convergence testing, the gold standard for numerical evolution. We have added some text throughout the

appendix to emphasize this point. We also modified the last paragraph of section III.B slightly.

10. In addition, while convergence testing is the gold standard for validating our evolutions, it is not the only reliability test. We have made two other checks on the evolutions presented in figure 16 and found no problem. Therefore, we believe those evolutions are reliable. We already discussed this point in our previous submission.

11. The referee suggests "In order to demonstrate that the (non-)monotonic behavior of t_H is real one should compare plots of t_H vs ϵ with different spatial/temporal resolutions. No such test was presented. Also, checking behavior of $t_H(\epsilon)$ for different threshold values for $A(x_H, t_H)$ would be valuable." This is unnecessary since we have validated nonmonotonicity explicitly in several of our convergence tests. In fact, convergence testing does check t_H vs ϵ with different resolutions, and it is not practical to re-run many amplitudes at higher resolutions. In addition, while the threshold value $A(x_H, t_H)$ is resolution-dependent, we did use slightly different formulae in this work and in reference [34] by two of us. We find a difference in t_H only at the 10^{-5} level for a sample of initial data (μ, σ, ϵ) that we evolved separately for the two works. But we emphasize that this is not necessary to demonstrate nonmonotonicity.

12. While making the revisions, we noticed that we had inadvertently been using an outdated version of figure 5b, which we have updated (the only change is to add data points corresponding to amplitudes $\epsilon = 3.52, 3.51$, which are discussed later in that subsection).

In short, we believe our work is correct and sufficiently novel and interesting to merit publication in PRD. We have made a number of revisions to our manuscript, including some improvements suggested by the referee. However, in aggregate, we disagree with the referee for the reasons laid out in detail above.