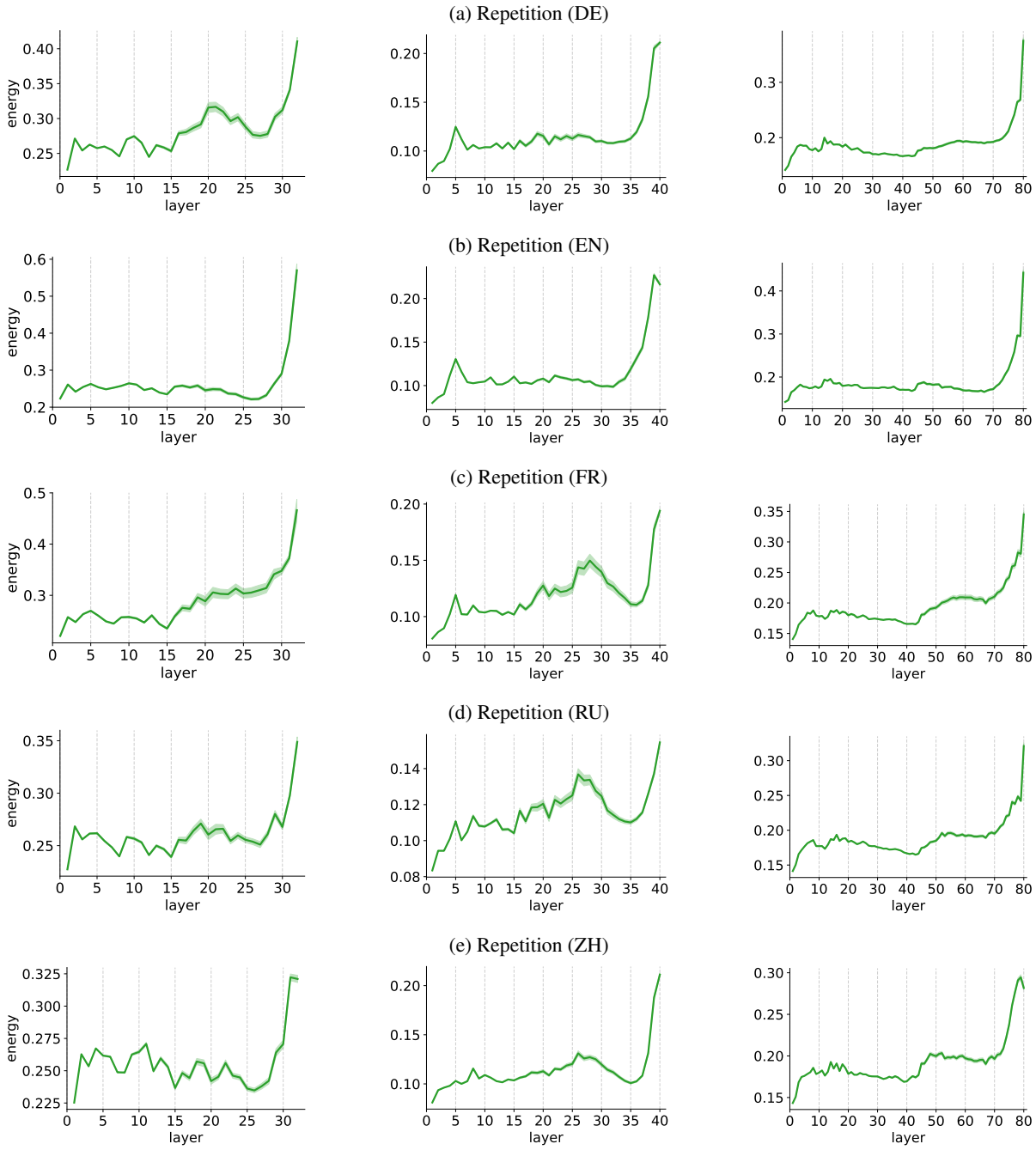

Figure 8: Figures illustrate the energy plots for the repetition task where Llama-2 7B, 13B, and 70B are tasked with copying a non-English word. There is one column per model size. The x-axis shows the layer number of the model, and the y-axis the energy. Means and 95% Gaussian confidence intervals have been computed over the input examples, numbers in Appendix A.

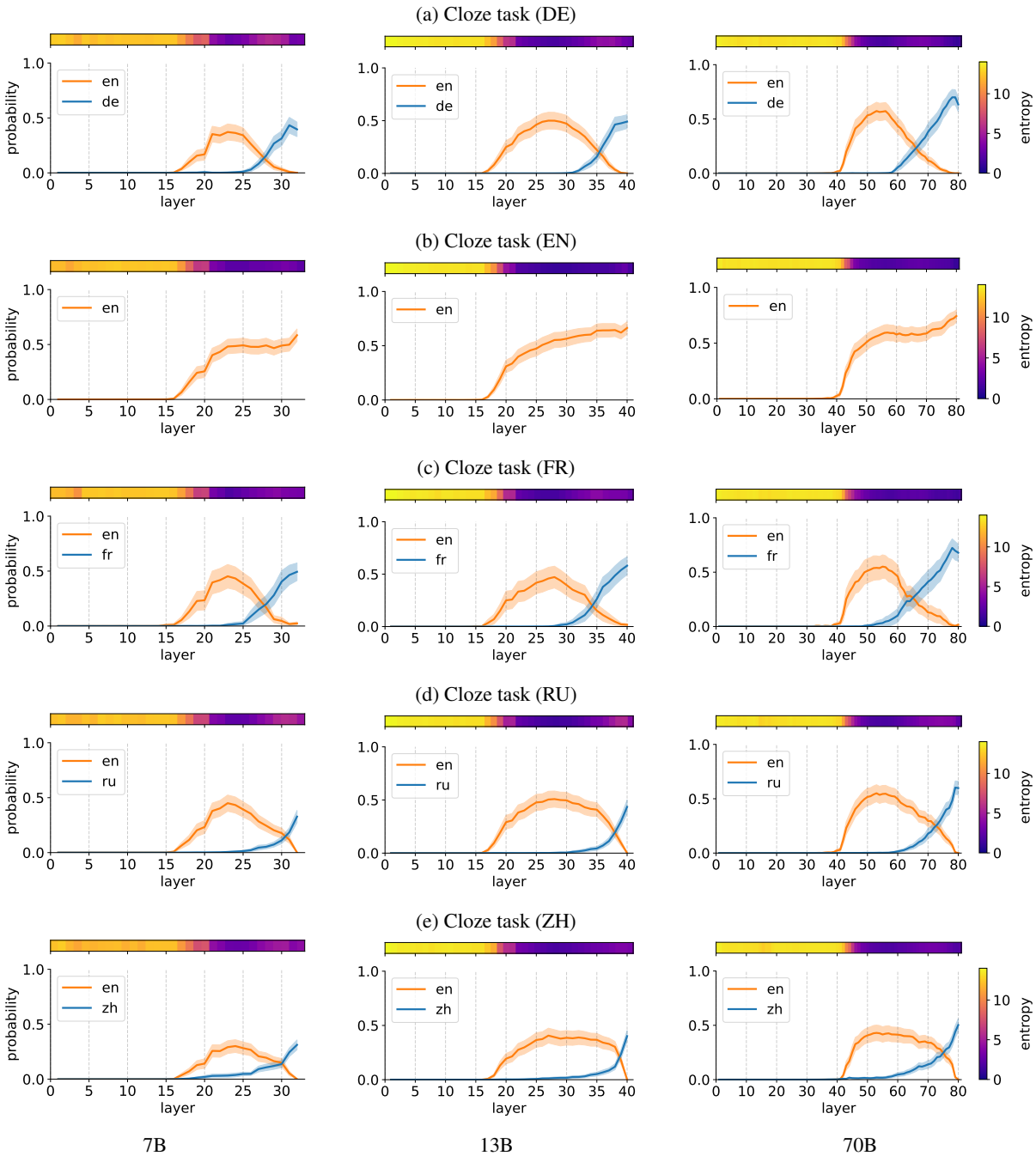

Figure 9: Figures show the same plots only for the cloze task where the correct token is defined in a fill-in-the-blank setting. In the plots, we illustrate the results for German. Means and 95% Gaussian confidence intervals have been computed over the input examples, numbers in Appendix A.

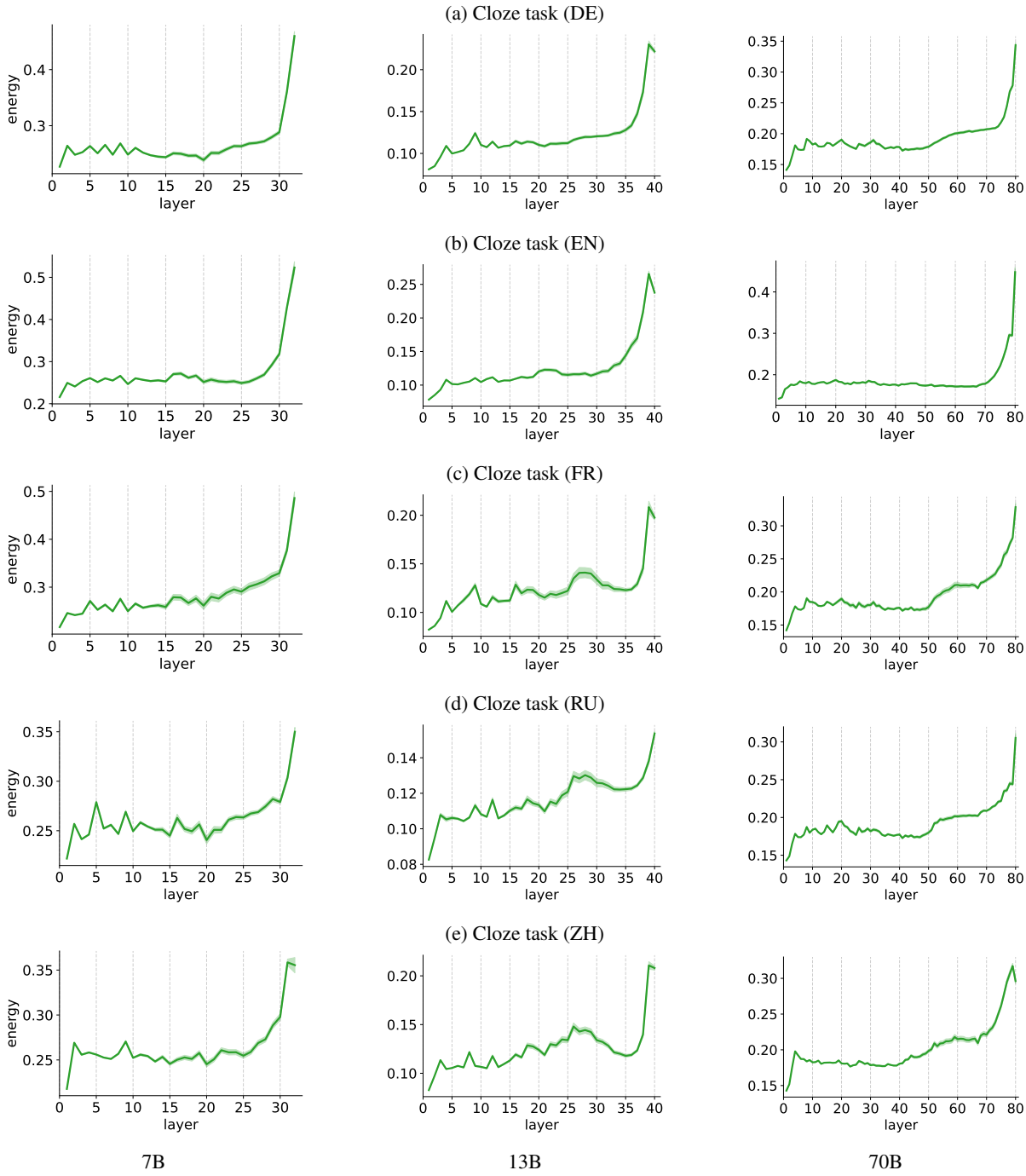

Figure 10: Figures show the same plots only for the cloze task where the correct token is defined in a fill-in-the-blank setting. In the plots, we illustrate the results for German. Means and 95% Gaussian confidence intervals have been computed over the input examples, numbers in Appendix A.