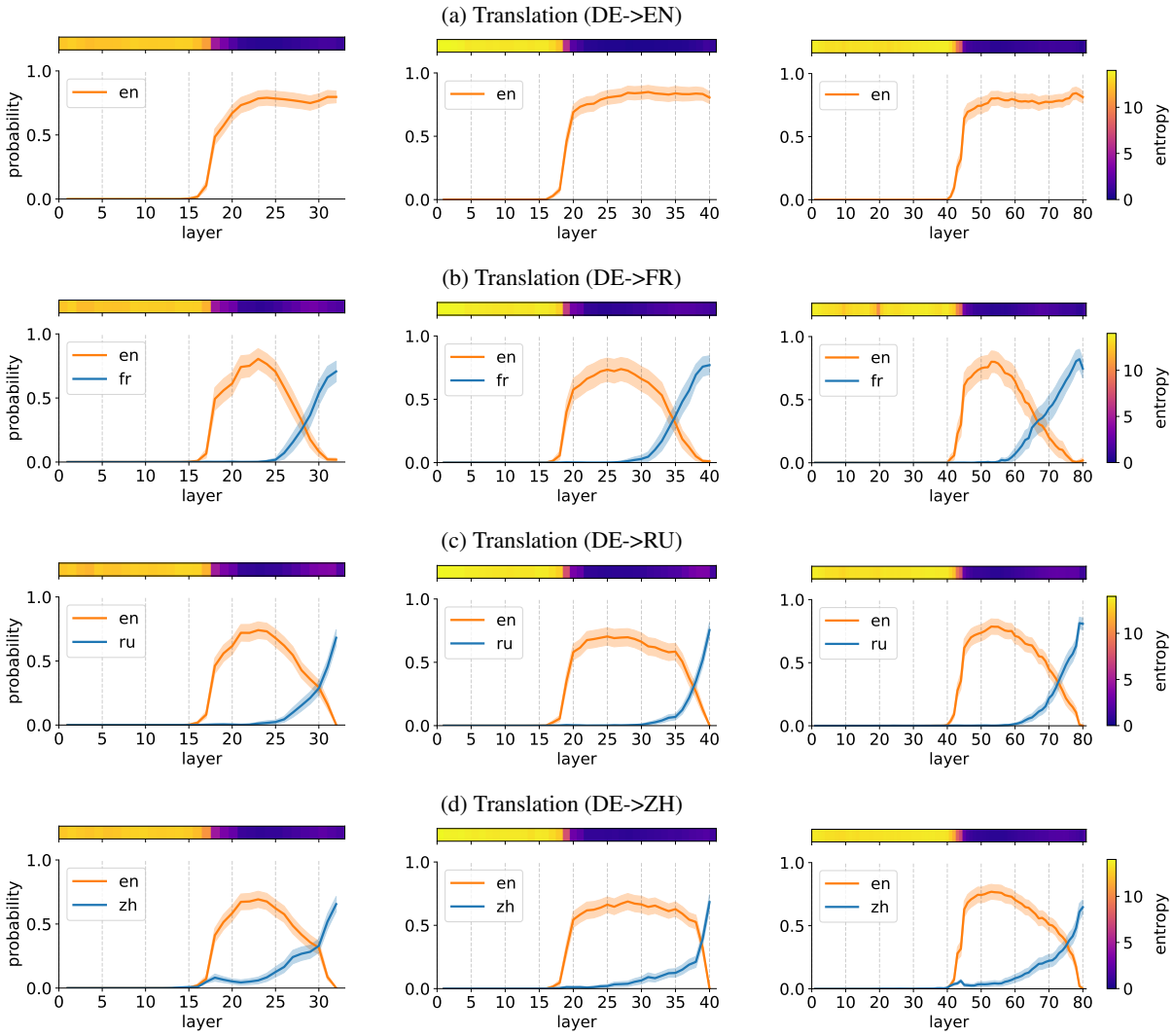

Figure 11: Figures illustrate the translation task where Llama-2 7B, 13B, and 70B are tasked with translating a word from non-English input language to output language. There is one column per model size. The x-axis shows the layer number of the model, and the y-axis the total probability mass falling on the correct token across languages. The orange line illustrates the probability of the correct target word in English and the blue line shows it for the non-English output language. We do not include the probability the input language since it is zero throughout. Means and 95% Gaussian confidence intervals have been computed over the input examples, numbers in Appendix A.

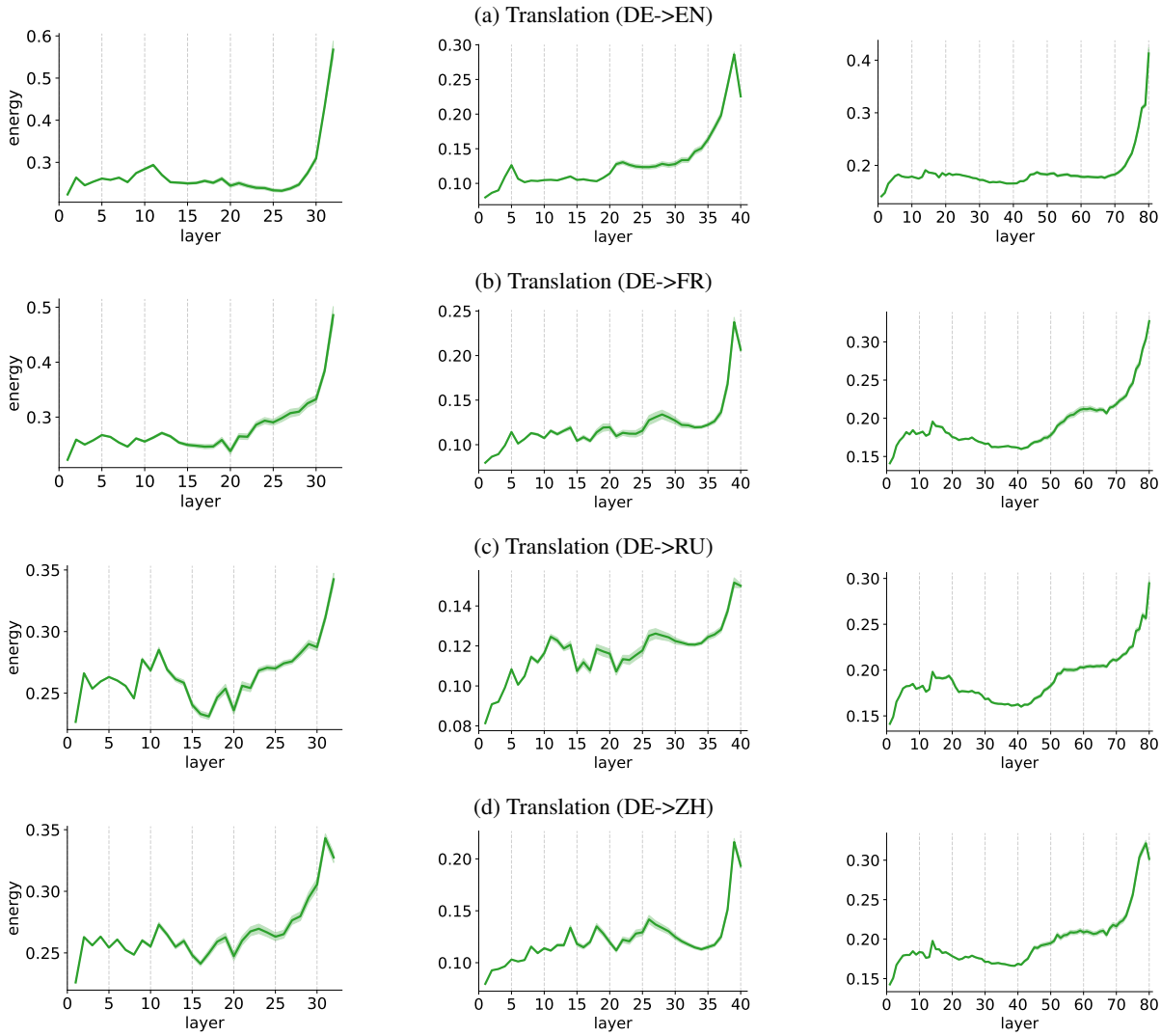

Figure 12: Figures illustrate the translation task where Llama-2 7B, 13B, and 70B are tasked with translating a word from non-English input language to output language. There is one column per model size. The x-axis shows the layer number of the model, and the y-axis the energy. Means and 95% Gaussian confidence intervals have been computed over the input examples, numbers in Appendix A.

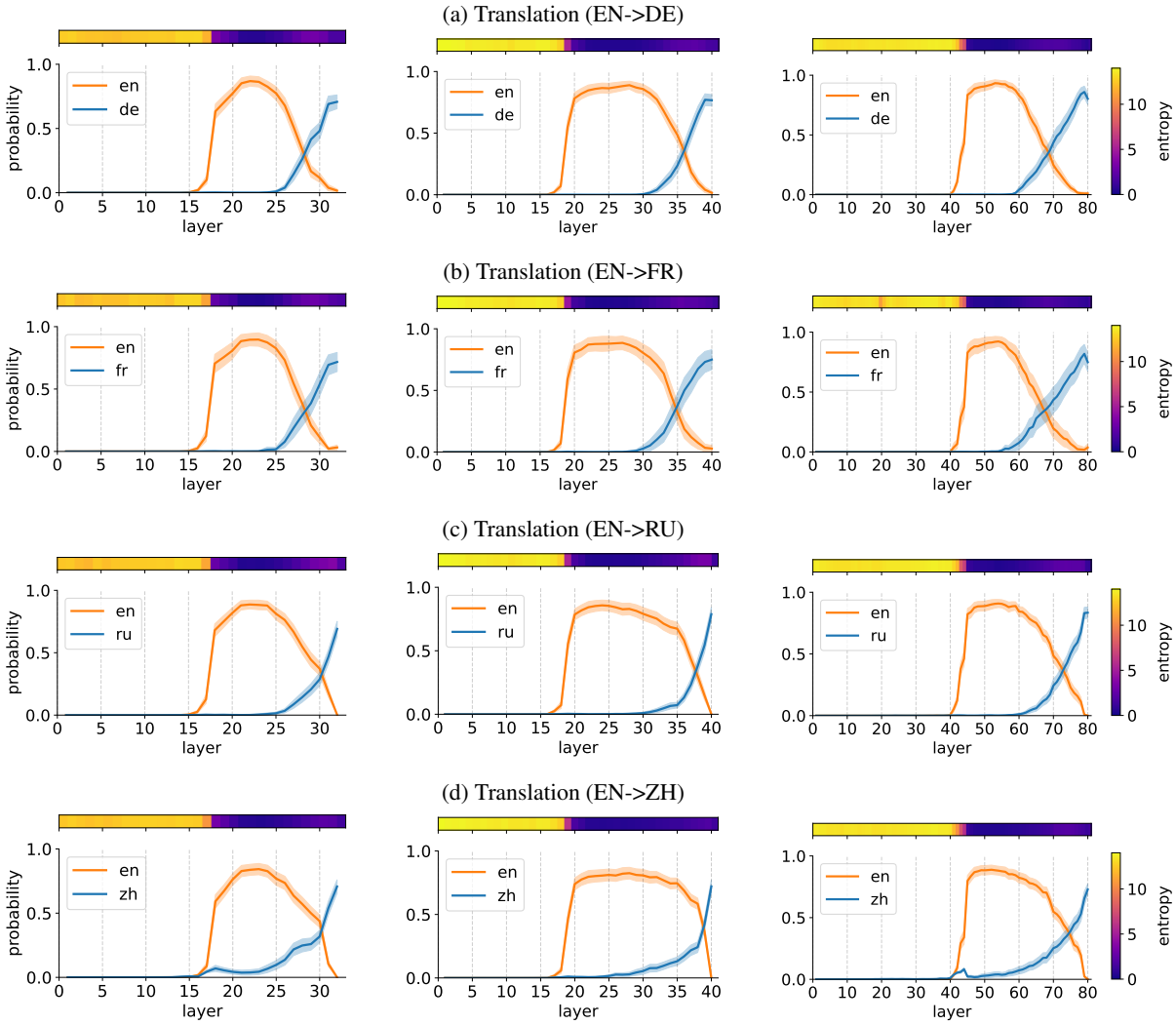

Figure 13: Figures illustrate the translation task where Llama-2 7B, 13B, and 70B are tasked with translating a word from English input language to output language. There is one column per model size. The x-axis shows the layer number of the model, and the y-axis the total probability mass falling on the correct token across languages. The orange line illustrates the probability of the correct target word in English and the blue line shows it for the non-English output language. We do not include the probability the input language since it is zero throughout. Means and 95% Gaussian confidence intervals have been computed over the input examples, numbers in Appendix A.