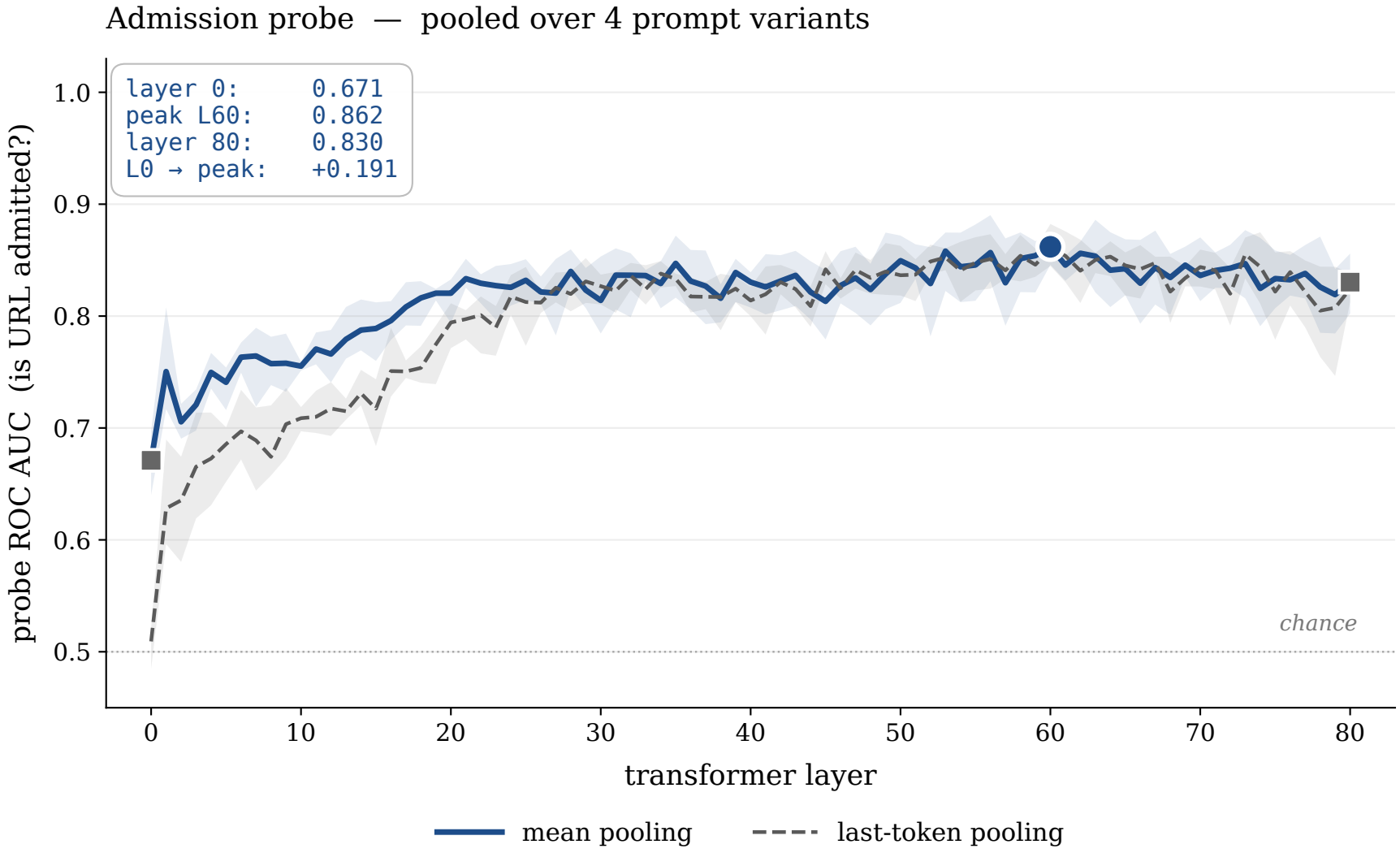

For each URL span in the rerank prompt, the label is 1 if the (model, variant) admitted that URL, else 0.
Linear probe (logistic regression on frozen hidden states), per (layer, pooling). Llama-3.3-70B + Qwen-2.5-72B. Shaded band: min–max envelope across the 4 prompt variants.