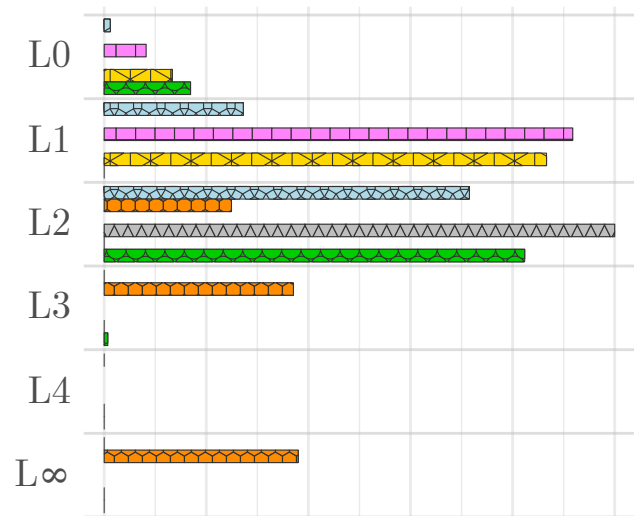
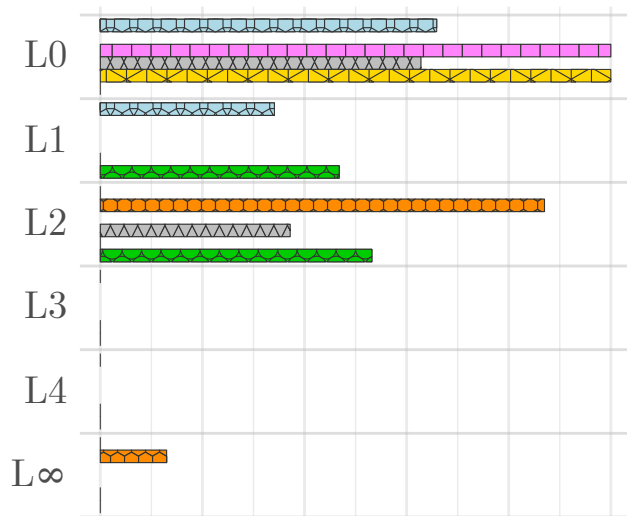


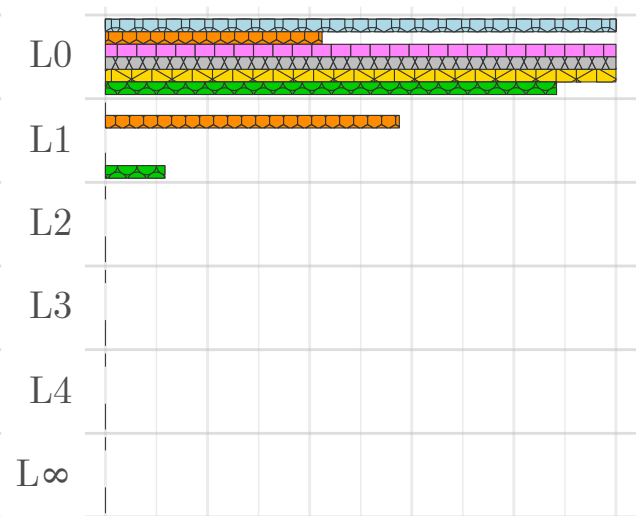
$p=2/3$ (baseline)



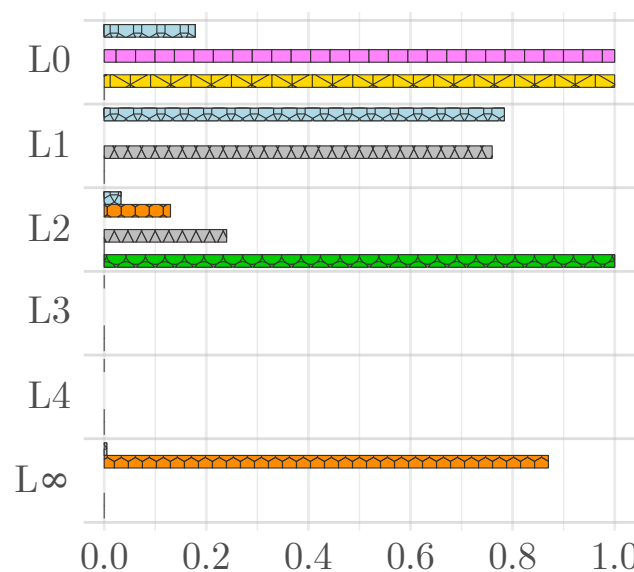
$p=1/2$



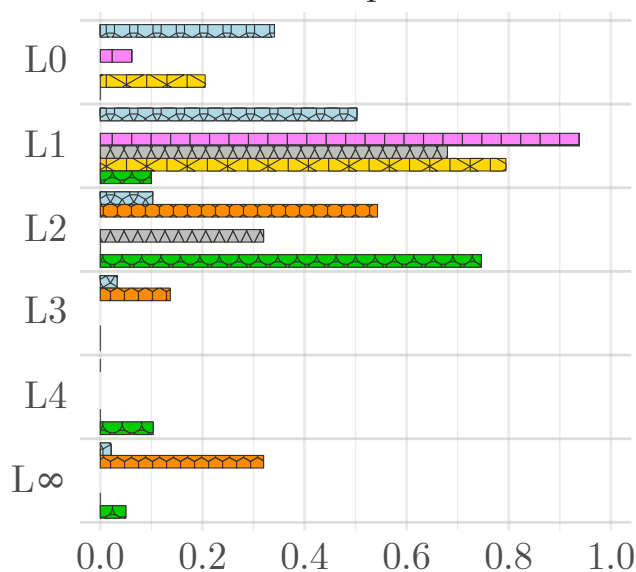
$p=4/3$



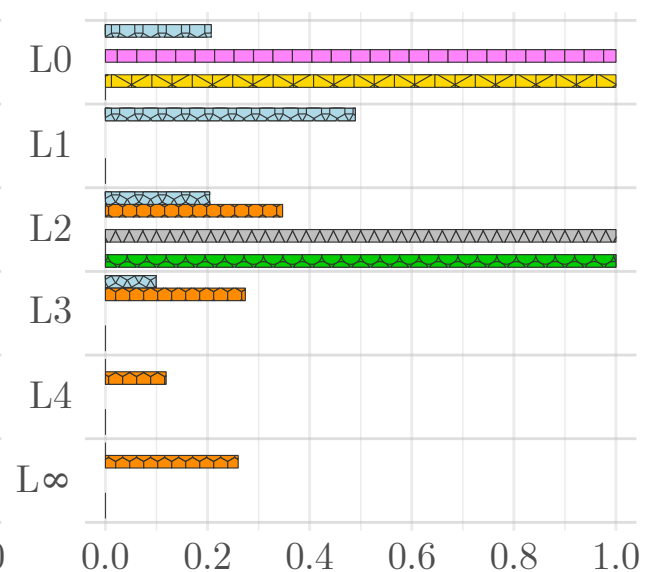
$n=2$



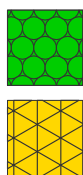
$n = \text{Unspecified}$



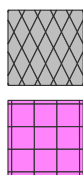
Median



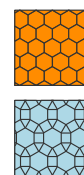
Average Proportions



Gemini-2.0 Flash (Thinking)



Claude-3.5 (Sonnet)



ChatGPT-o1



Gemini-1.5 Flash



Claude-2.1



ChatGPT-4