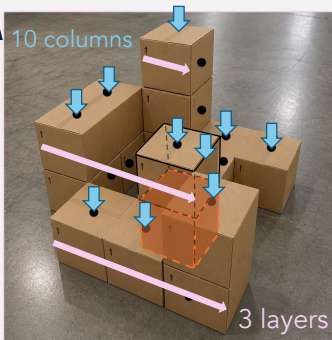


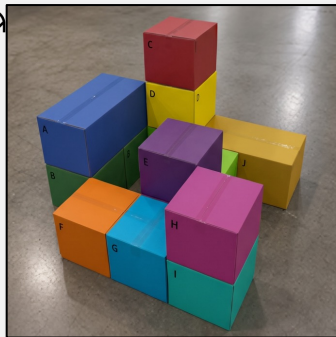
Q: How many boxes are in this warehouse?

 Human: Hierarchical



15 visible + 1 hidden  
= 16 total ✓

 Multimodal LLMs: ?



12 visible? + 0 hidden?  
= 13 total ✗

Finetuning on Spatial-IQ Data

GT Labels  
+



Hierarchical  
CoT Reasoning



15 + 1 = 16 ✓