1) Decoder only transformers are more scalable mainly because they use a simpler and more uniform structure. There is only one stack of layers, instead of separate encoder & decoder stacks. This makes training easier to parallelize & more efficient on large hardwares like GPUs & TPUs. Decoder-only models also reuse the same self-attention mechanism at every stage, avoiding the extra cross-attention step used in encoder-decoder models. This reduces memory usage & computation during both training & inference. Another advantage is that decoder-only models naturally fit the next-token prediction objective used for large scale pretraining. Overall, fewer moving parts & better computational reuse makes decoder-only transformers scale better for large language models.

2) Next-token prediction works because language itself contains patterns that reflect meaning, logic, structure. When a model learns to predict the next word, it is indirectly learning grammar, facts & relationship b/w ideas. Tasks like translation & summarization are already present in text data, where one sequence naturally follows another. By seeing many such examples, the model learns how one form of text maps to another. Reasoning also appears in text as step-by-step explanations, arguments & cause-effect relations over large datasets, this single objective captures many skills without needing separate training tasks.