When stacking self-attention layers with positional encoding in a deep architecture, one common issue is that positional information can start to fade as it moves through the layers. The model may increasingly focus on content-based relationships and underutilize positional cues, which weakens its ability to understand the order of the sequence, something that's often pretty important. Deep transformer models also tend to have high capacity, so they're prone to overfitting on smaller datasets. On top of that, without careful initialization, normalization, and residual connections, training can become unstable, gradients might vanish or explode, making convergence harder. And of course, the deeper the stack, the more expensive it becomes in terms of compute and memory, which can be a real bottleneck unless you apply optimizations like sparse attention or hierarchical structures.