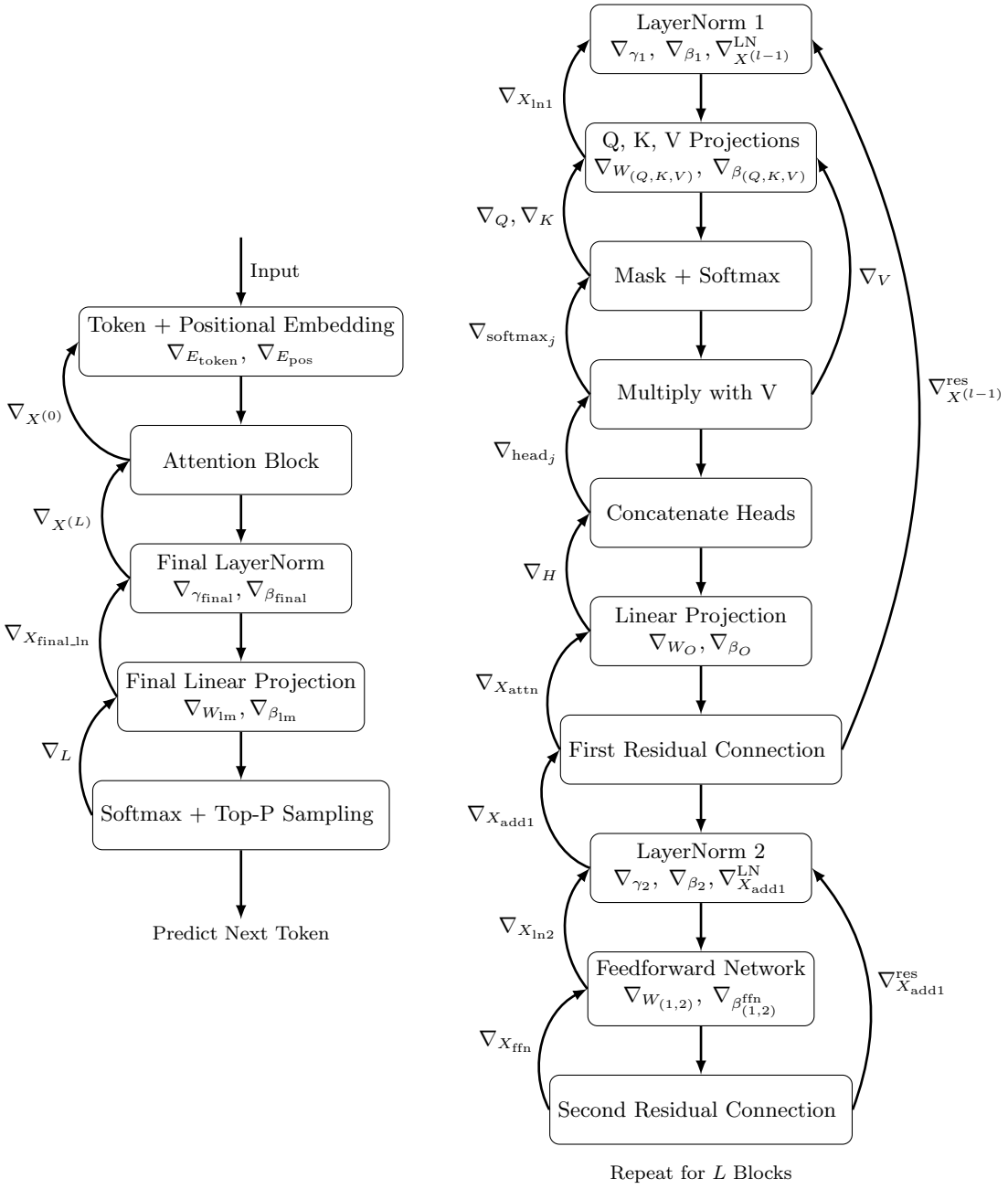

Figure 1: Overview of the GPT transformer architecture. (Left) High-level structure of the autoregressive transformer; (Right) internal breakdown of the attention mechanism, this fits into the attention block on the left. The arrows pointing downward signify the forward pass while the arrows pointing upward signify the backward propagation of gradients. The gradients listed in each block represent the gradients that are calculated for that block but aren't directly sent to previous blocks.