

Output Probability



Feed Forward



Norm

Cross-Attention & Gating Network

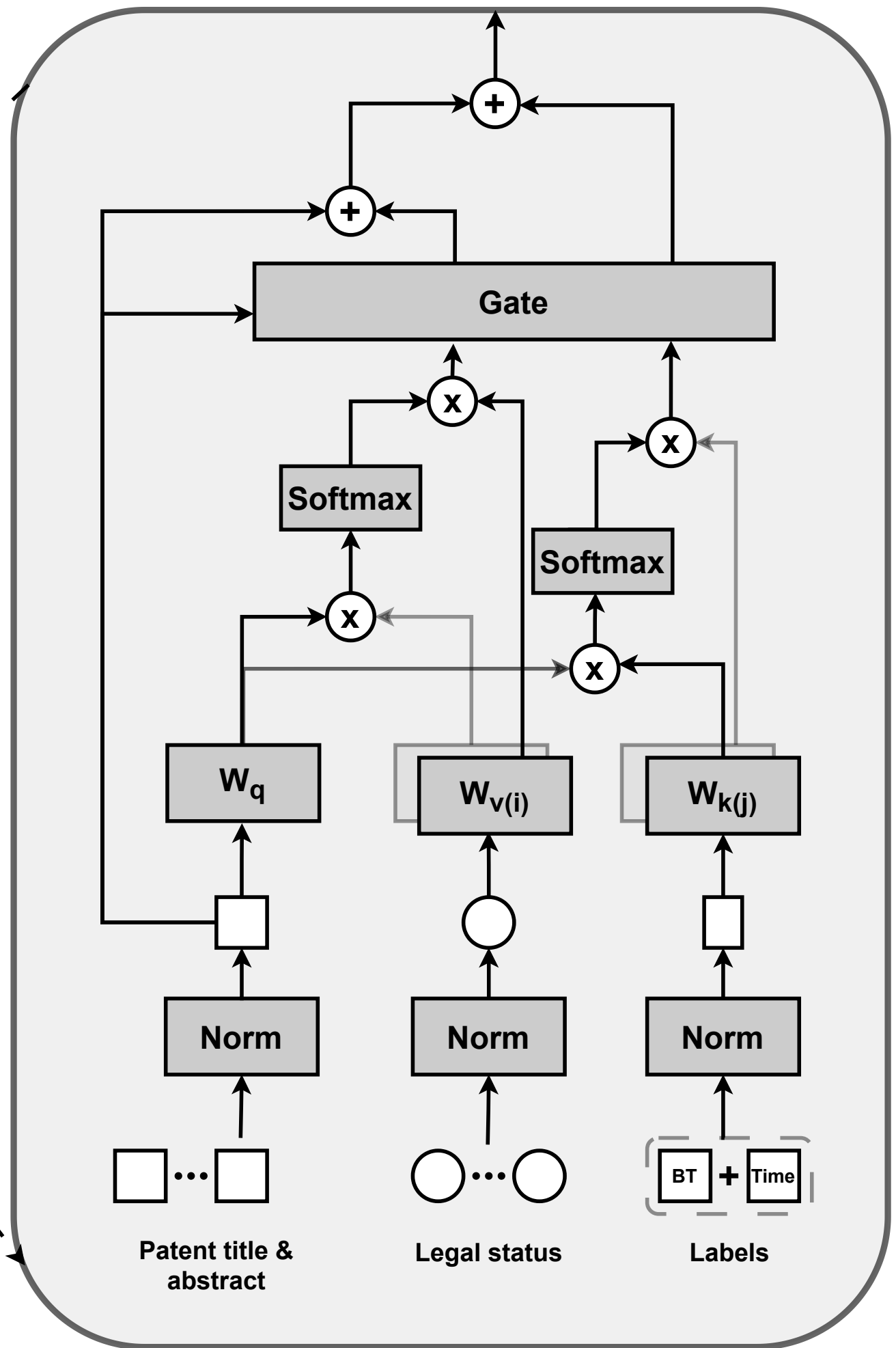


Multi-Head Self-Attention

Norm

Input Tokens

Transformer Block



Cross-Attention & Gating Network