Algorithm 1 The pipeline of PTQ1.61

```
Require: Pretrained LLM $M$, input activation $X$
Ensure: Quantized LLM $M_q$
$M_P = \mathrm{P}(M)$                                  ▷ Preprocessing
for each block $b$ in $M_P$ do
    $m_s \leftarrow \mathrm{Top}_{20\%}(\mathrm{mean}(|X[:, c, \ldots]|))$
    for each channel $c$ in $W$ do
        if $m_s[c] = 1$ then
            Quantize $(W_c)$                           ▷ Salient weights
        else
            Binarize $(W_c)$                           ▷ Unsalient weights
        end if
    end for
    $Y_{\mathrm{full}} \leftarrow b(X)$
    $Y_q \leftarrow b_q(X)$
    $Y_{fq} \leftarrow b(Y_{q-1})$
    $\mathrm{L}_{norm} = \mathrm{MSE}(Y_q, Y_{full}) + \mathrm{NLC}(Y_q, Y_{full})$
    $\mathrm{L}_{homo} = \mathrm{MSE}(Y_q, Y_{fq}) + \mathrm{NLC}(Y_q, Y_{fq})$
    $\mathrm{Loss} = \mathrm{L}_{norm} + \mathrm{L}_{homo}$
    Loss.backward()                                    ▷ Optimize scaling factors of quantized weights
end for
return Quantized model $M_q$
```