Fig. 1

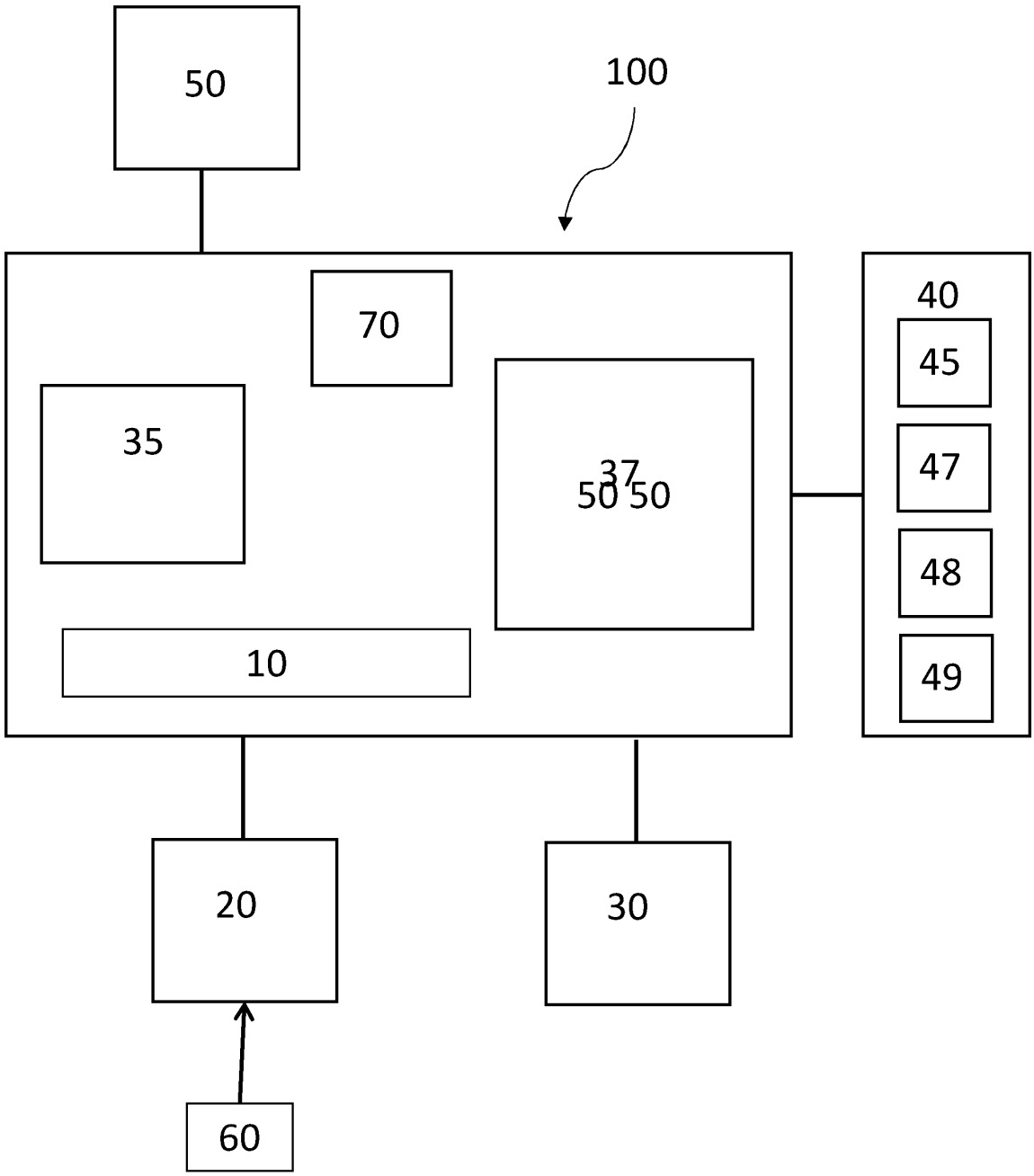


Fig. 2

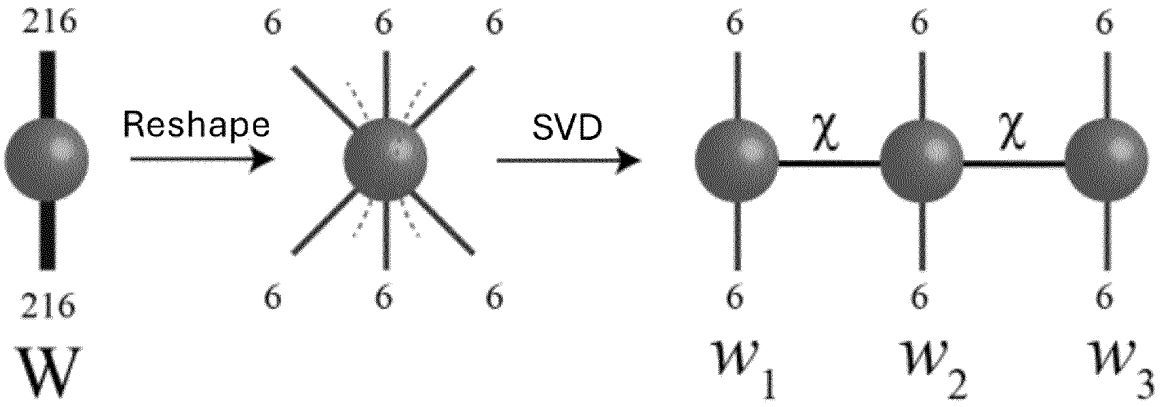


Fig. 3

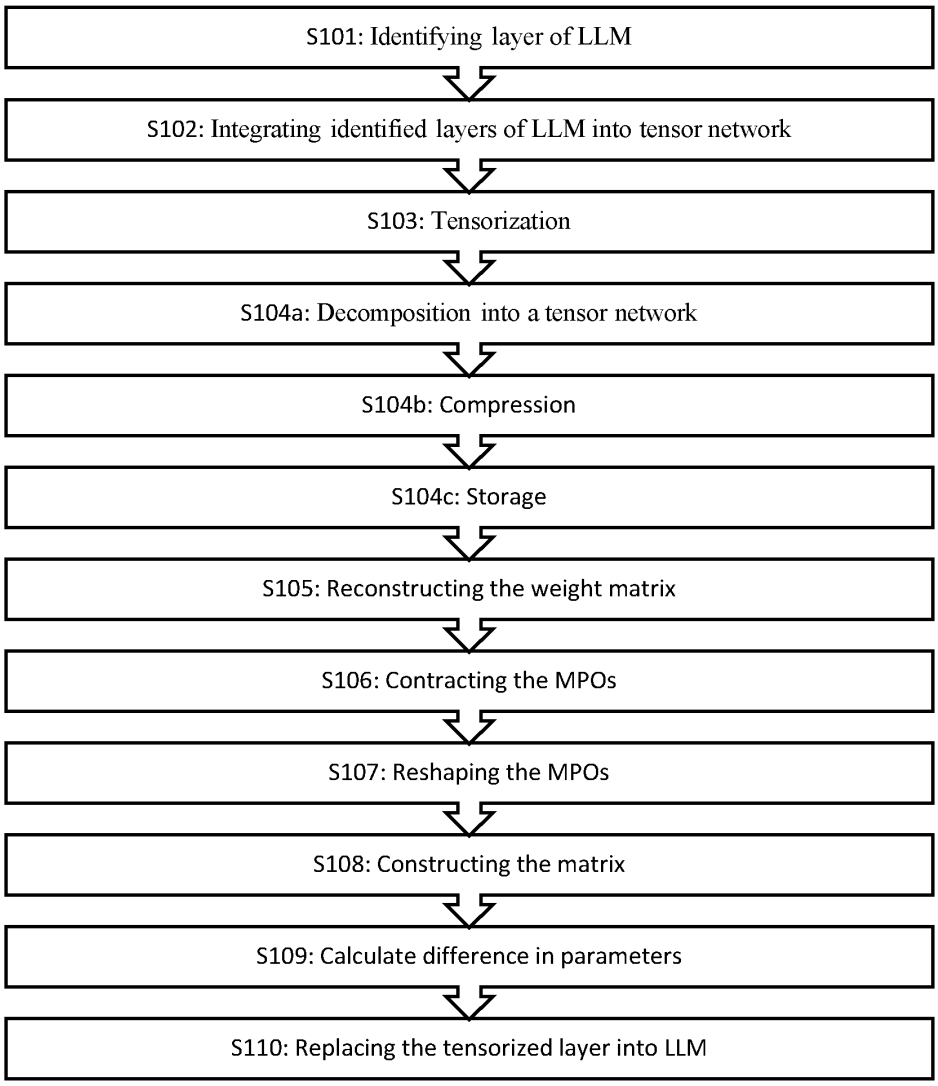

S101: Identifying layer of LLM
S102: Integrating identified layers of LLM into tensor network
S103: Tensorization
S104a: Decomposition into a tensor network
S104b: Compression
S104c: Storage
S105: Reconstructing the weight matrix
S106: Contracting the MPOs
S107: Reshaping the MPOs
S108: Constructing the matrix
S109: Calculate difference in parameters
S110: Replacing the tensorized layer into LLM

Fig. 4

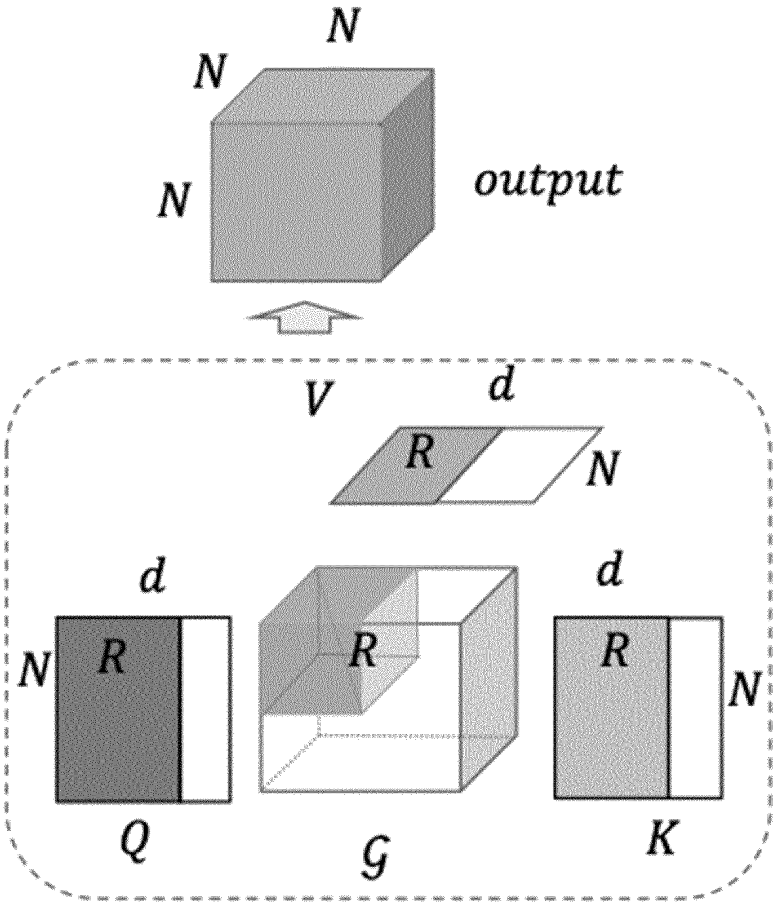


Fig. 5

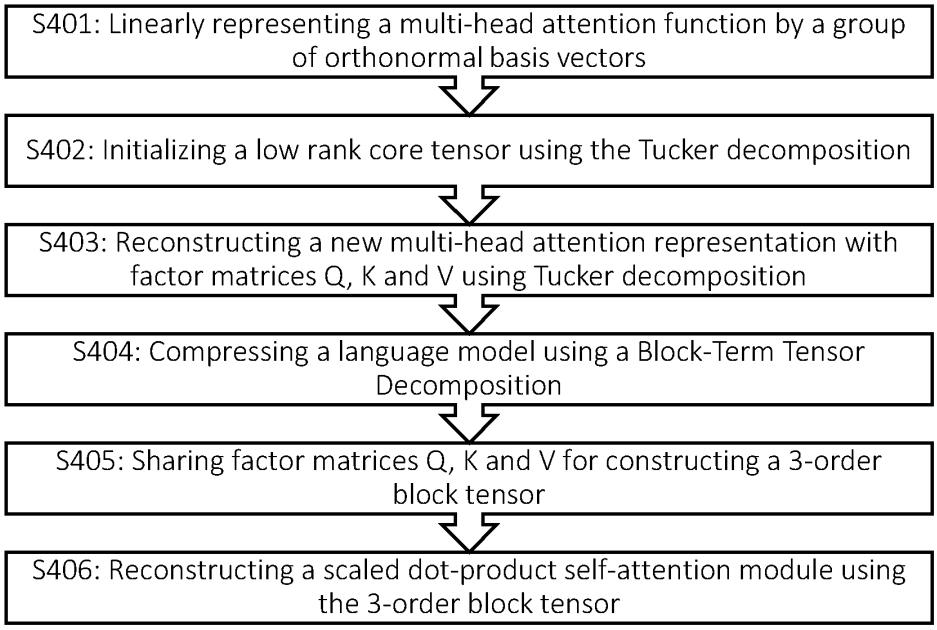

S401: Linearly representing a multi-head attention function by a group of orthonormal basis vectors
S402: Initializing a low rank core tensor using the Tucker decomposition
S403: Reconstructing a new multi-head attention representation with factor matrices Q, K and V using Tucker decomposition
S404: Compressing a language model using a Block-Term Tensor Decomposition
S405: Sharing factor matrices Q, K and V for constructing a 3-order block tensor
S406: Reconstructing a scaled dot-product self-attention module using the 3-order block tensor

Fig. 6

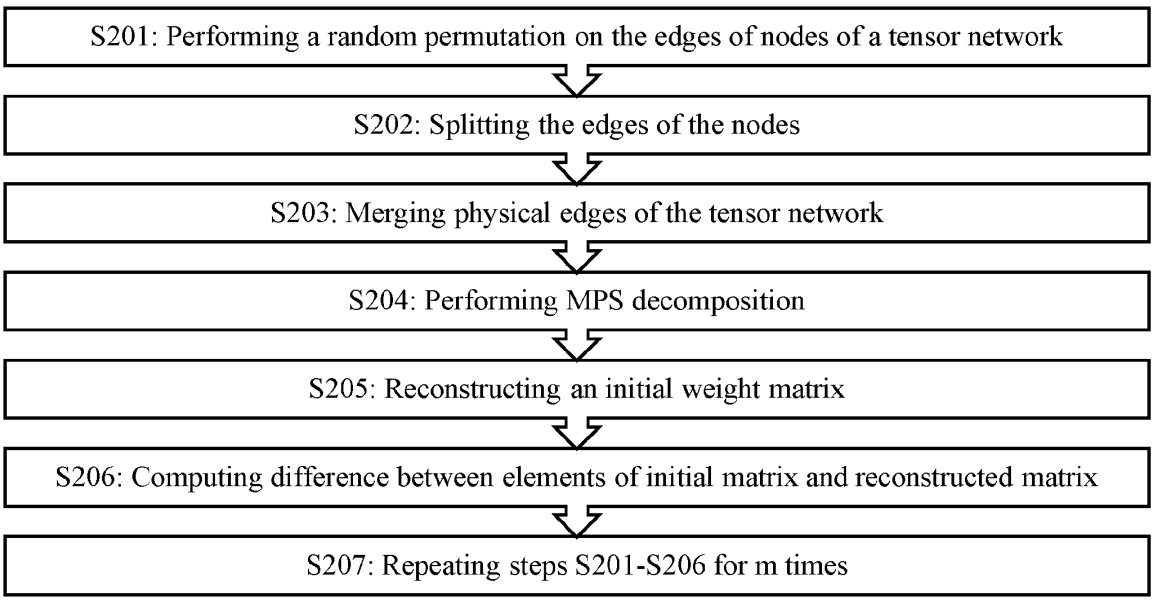

S201: Performing a random permutation on the edges of nodes of a tensor network
S202: Splitting the edges of the nodes
S203: Merging physical edges of the tensor network
S204: Performing MPS decomposition
S205: Reconstructing an initial weight matrix
S206: Computing difference between elements of initial matrix and reconstructed matrix
S207: Repeating steps S201-S206 for m times

Fig. 7

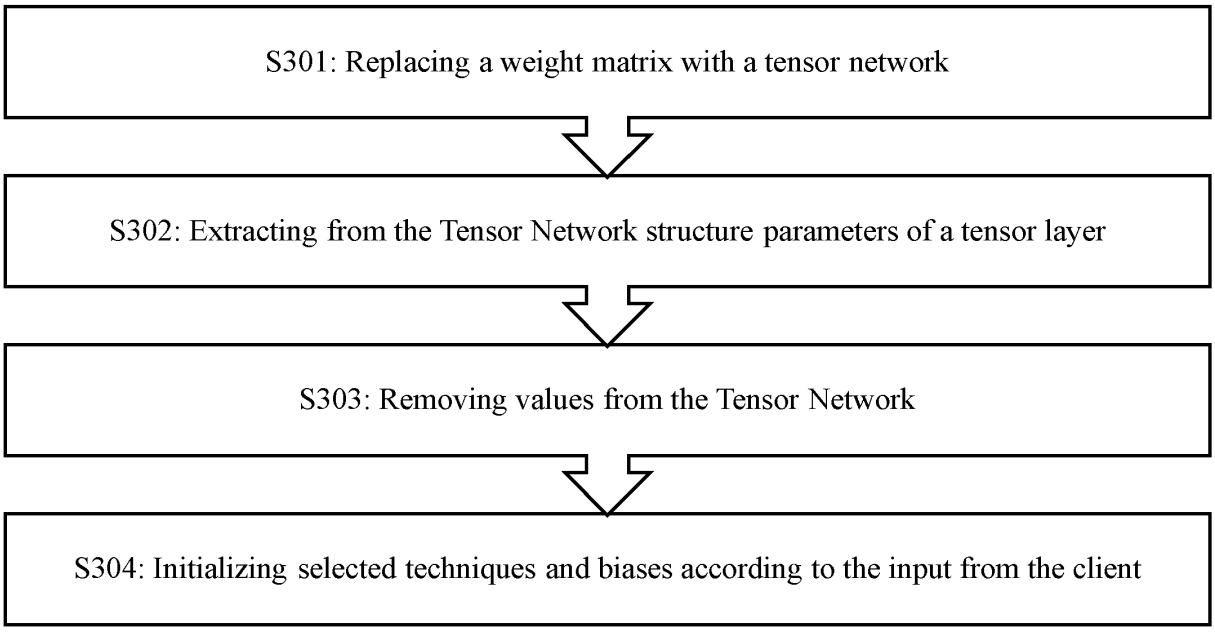

S301: Replacing a weight matrix with a tensor network
S302: Extracting from the Tensor Network structure parameters of a tensor layer
S303: Removing values from the Tensor Network
S304: Initializing selected techniques and biases according to the input from the client

Fig. 8

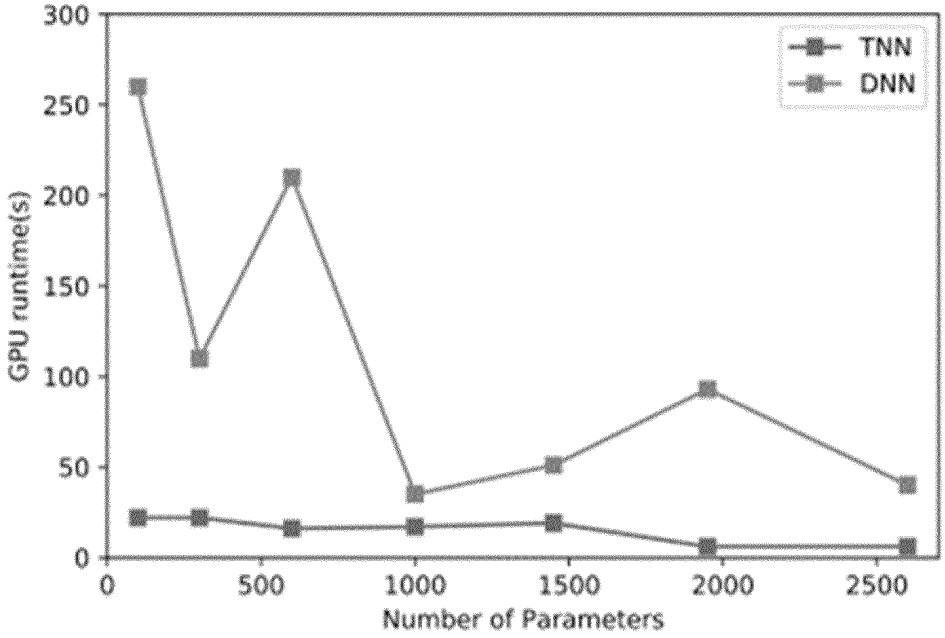

300
250
200
150
100
50
0
GPU runtime(s)
0
500
1000
1500
2000
2500
Number of Parameters
TNN
DNN

Fig. 9

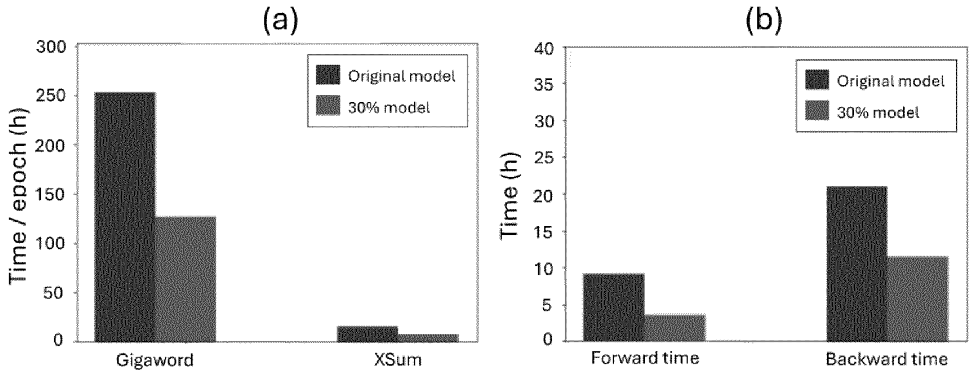
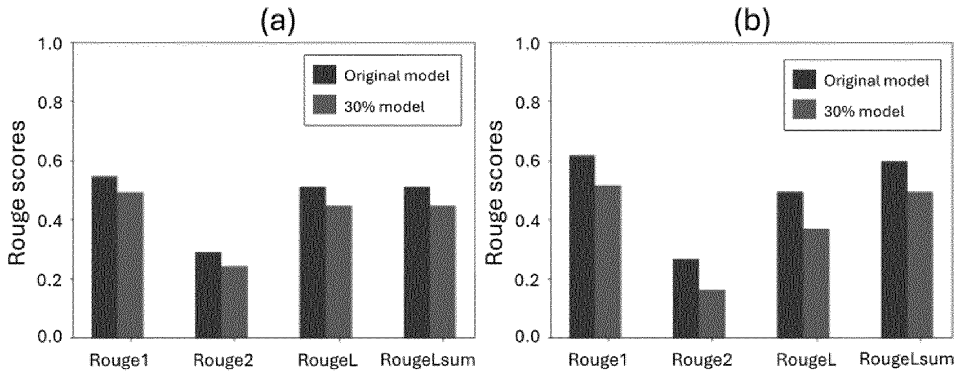

(a)
Time / epoch (h)
300
250
200
150
100
50
0
Original model
30% model
Gigaword
XSum
(b)
Time (h)
40
35
30
25
20
15
10
5
0
Original model
30% model
Forward time
Backward time

Fig. 10


(a)
Rouge scores
1.0
0.8
0.6
0.4
0.2
0.0
Original model
30% model
Rouge1
Rouge2
RougeL
RougeLsum
(b)
Rouge scores
1.0
0.8
0.6
0.4
0.2
0.0
Original model
30% model
Rouge1
Rouge2
RougeL
RougeLsum