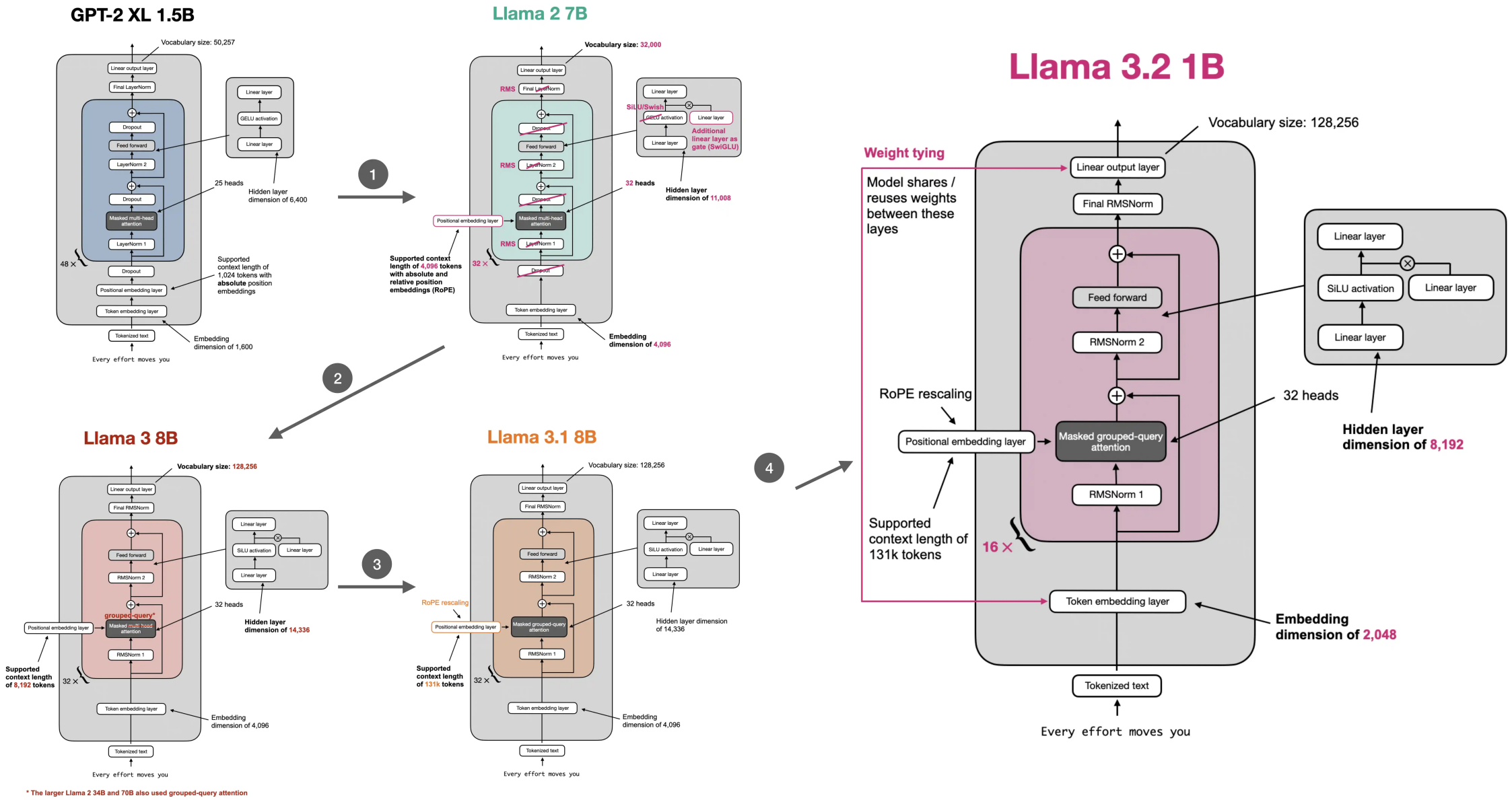
GPT-2 XL 1.5B
Vocabulary size: 50,257
Linear output layer
Final LayerNorm
Dropout
Feed forward
LayerNorm 2
Dropout
Masked multi-head attention
LayerNorm 1
48 ×
Dropout
Positional embedding layer
Token embedding layer
Tokenized text
Every effort moves you
Linear layer
GELU activation
Linear layer
25 heads
Hidden layer dimension of 6,400
Supported context length of 1,024 tokens with absolute position embeddings
Embedding dimension of 1,600
1
Llama 2 7B
Vocabulary size: 32,000
RMS Final LayerNorm
RMS LayerNorm 2
RMS LayerNorm 1
32 ×
Positional embedding layer
Masked multi-head attention
Dropout
Feed forward
Token embedding layer
Tokenized text
Every effort moves you
SiLU/Swish
Linear layer
Additional linear layer as gate (SwiGLU)
32 heads
Hidden layer dimension of 11,008
Supported context length of 4,096 tokens with absolute and relative position embeddings (RoPE)
Embedding dimension of 4,096
2
Llama 3 8B
Vocabulary size: 128,256
Final RMSNorm
RMSNorm 2
RMSNorm 1
32 ×
grouped-query*
Positional embedding layer
Token embedding layer
Tokenized text
Every effort moves you
Feed forward
SiLU activation
Linear layer
32 heads
Hidden layer dimension of 14,336
Supported context length of 8,192 tokens
Embedding dimension of 4,096
3
Llama 3.1 8B
Vocabulary size: 128,256
RoPE rescaling
Masked grouped-query attention
Supported context length of 131k tokens
4
Llama 3.2 1B
Weight tying
Model shares / reuses weights between these layes
Vocabulary size: 128,256
Linear output layer
Final RMSNorm
Feed forward
RMSNorm 2
RMSNorm 1
16 ×
Token embedding layer
Tokenized text
Every effort moves you
Linear layer
SiLU activation
Linear layer
Linear layer
32 heads
Hidden layer dimension of 8,192
RoPE rescaling
Positional embedding layer
Masked grouped-query attention
Supported context length of 131k tokens
Embedding dimension of 2,048
* The larger Llama 2 34B and 70B also used grouped-query attention