ChatGLM3?????

```
## data flow
```

```
                    query     ->  "□□"
                      |
                  tokenizer   -> input_ids  [6]
                      |
```

rotary_pos_emb embedding -> [1, 6, 4096] \ / GLMBlock x 28 -> [6, 1, 4096] <---| RMSNorm -> [6, 1, 4096] | final_layernorm [-1:] -> [1, 1, 4096] | Linear -> [1, 1, 65024] | output_layer 4096->65024 softmax -> [1, 65024] | multinomial -> [1] | cat([input_ids, next_tokens]) ---| ↓ tokenizer.decode( )

GLMBlock

```
   input
```

/
/ RMSNorm hidden_states -> [6, 1, 4096] | | /
		pow(2) -> [6, 1, 4096]							mean -> [6, 1, 1]			↓			rsqrt(+ eps) -> [6, 1, 1]		\ /		mul -> [6, 1, 4096]		\ weight -> [4096]		\ /	RMSNorm mul -> [6, 1, 4096]
SelfAttention x -> [6, 1, 4096]					Linear -> [6, 1, 4608] 4096->4608		/
	q k v [6, 1, 32, 128] [6, 1, 2, 128] [6, 1, 2, 128]		/
	pos_emb pos_emb \ -> cat(x0*y0-x1*y1, x1*y0-x0*y1, x, y)							expand expand -> [6, 1, 32, 128] [6, 1, 32, 128]		permute permute permute -> [1, 32, 6, 128] [1, 32, 6, 128] [1, 32, 6, 128]		\ /			|---- matmul	-> [1, 32, 6, 128] [1, 32, 128, 6] -> [1, 32, 6, 6]			add(mask) / -> [1, 32, 6, 6]		attention| softmax / -> [1, 32, 6, 6] dim:-1			\ /		|---- matmul -> [1, 32, 6, 6] [1, 32, 6, 128] -> [1, 32, 6, 128] -> [6, 1, 4096]	SelfAttention Linear -> [6, 1, 4096] 4096->4096	/	dropout \ / Add
/
RMSNorm hidden_states -> [6, 1, 4096]		/
		pow(2) -> [6, 1, 4096]							mean -> [6, 1, 1]			↓			rsqrt(+ eps) -> [6, 1, 1]		\ /		mul -> [6, 1, 4096]		\ weight -> [4096]		\ /	RMSNorm mul -> [6, 1, 4096]	/	mlp /		Linear -> [6, 1, 27392] 4096->27392		/
	chunk1 chunk0 -> [6, 1, 13696]			
			sigmoid				/			mul		\ /		mul -> [6, 1, 13696]	mlp Linear -> [6, 1, 4096] 13696->4096	/	dropout	/ Add

```

```

Revision #2
Created 2025-01-11 09:44:04 UTC by Colin
Updated 2026-04-29 07:34:22 UTC by Colin