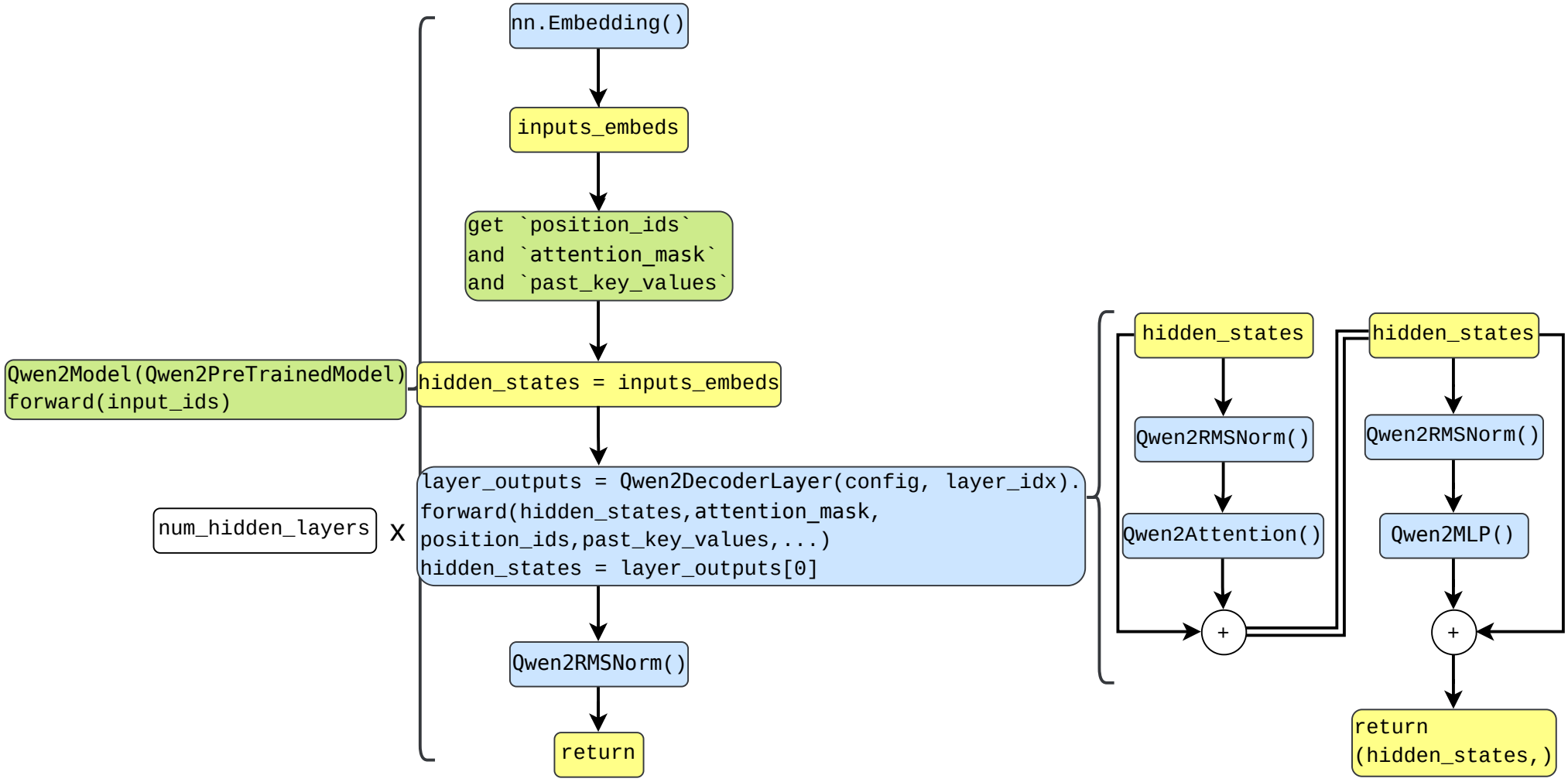

Qwen2RMSNorm():Copied from transformers.models.llama.modeling_llama.LlamaRMSNorm with Llama->Qwen2
Qwen2MLP(): Copied from transformers.models.mistral.modeling_mistral.MistralMLP with Mistral->Qwen2
Qwen2RotaryEmbedding(): Copied from transformers.models.mistral.modeling_mistral.MistralRotaryEmbedding with Mistral->Qwen2
rotate_half: Copied from transformers.models.llama.modeling_llama.rotate_half
apply_rotary_pos_emb: Copied from transformers.models.mistral.modeling_mistral.apply_rotary_pos_emb
repeat_kv: Copied from transformers.models.llama.modeling_llama.repeat_kv
Qwen2Attention: 与LlamaAttention类似
1.rotary_emb初始化：缺少对rope_scaling["type"]的判断，进而无法使用`LinearScalingRotaryEmbedding`或`DynamicNTKScalingRotaryEmbedding`，只能使用`RotaryEmbedding`
2.forward函数中：
2.1 没有`cache_position`参数
2.2 没有对`pretraining_tp > 1`的判断：关系到`q_proj`,`k_proj`,`v_proj`,`o_proj`
2.3 Llama: cos, sin = self.rotary_emb(value_states, position_ids)
    Qwen2: cos, sin = self.rotary_emb(value_states, seq_len=kv_seq_len)
2.4 attention_mask