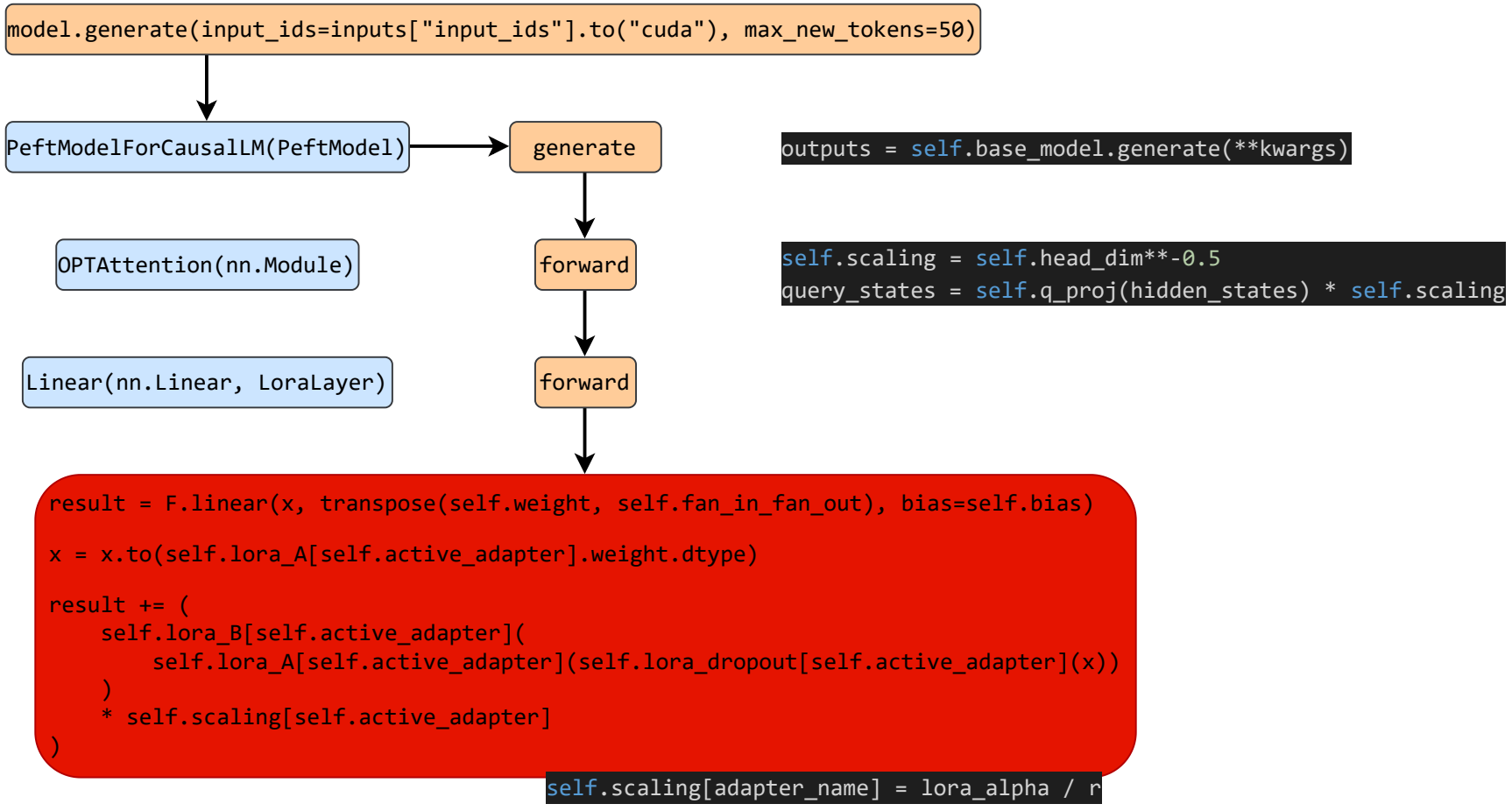
model.generate(input_ids=inputs["input_ids"].to("cuda"), max_new_tokens=50)
PeftModelForCausalLM(PeftModel)
generate
outputs = self.base_model.generate(**kwargs)
OPTAttention(nn.Module)
forward
self.scaling = self.head_dim**-0.5
query_states = self.q_proj(hidden_states) * self.scaling
Linear(nn.Linear, LoraLayer)
forward
result = F.linear(x, transpose(self.weight, self.fan_in_fan_out), bias=self.bias)
x = x.to(self.lora_A[self.active_adapter].weight.dtype)
result += (
self.lora_B[self.active_adapter](
self.lora_A[self.active_adapter](self.lora_dropout[self.active_adapter](x))
)
* self.scaling[self.active_adapter]
)
self.scaling[adapter_name] = lora_alpha / r