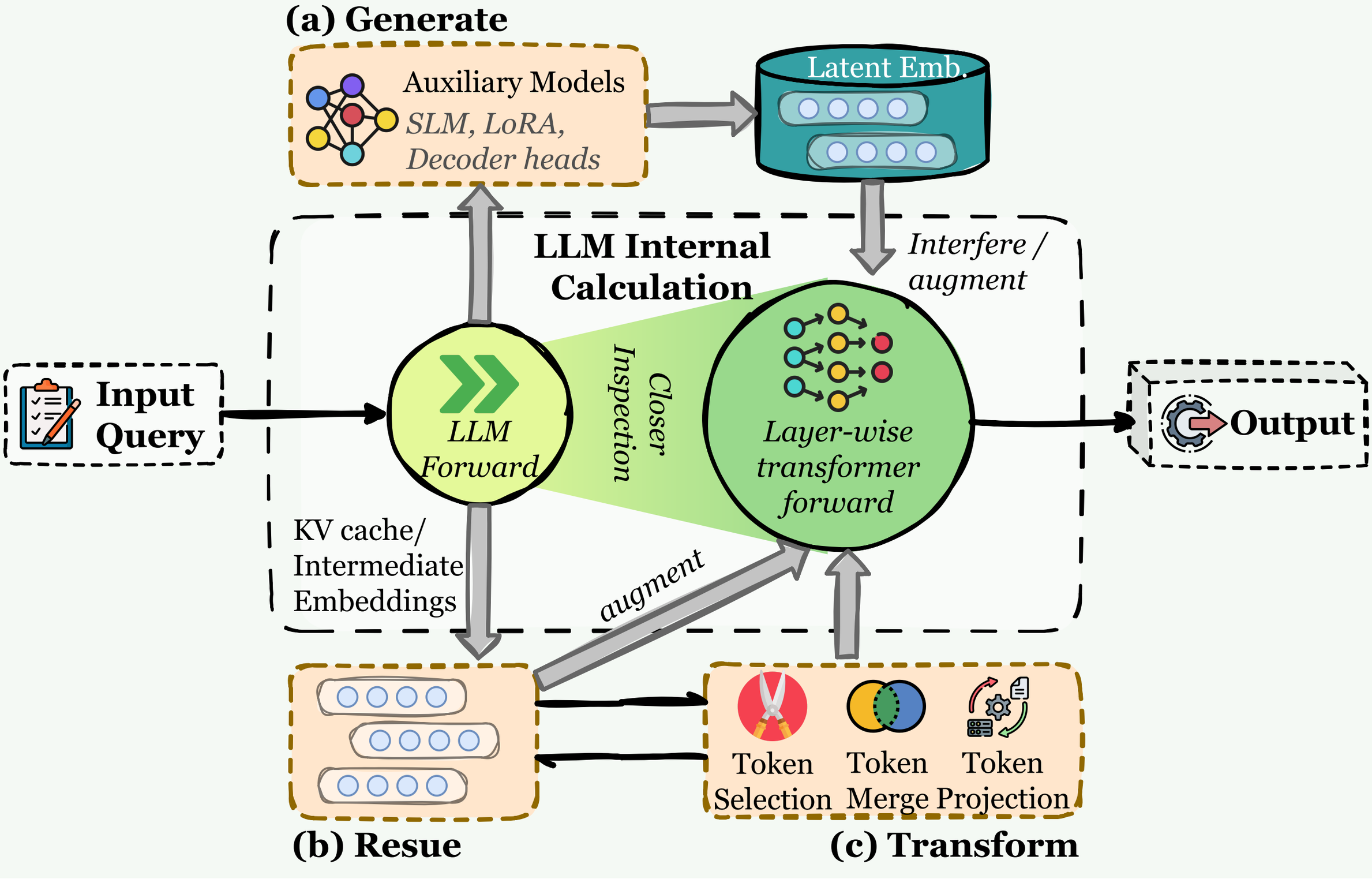
(a) Generate
Auxiliary Models
SLM, LoRA,
Decoder heads
Latent Emb.
LLM Internal Calculation
Interfere / augment
Input Query
LLM Forward
Closer Inspection
Layer-wise transformer forward
Output
KV cache/ Intermediate Embeddings
augment
Token Selection
Token Merge
Token Projection
(b) Resue
(c) Transform