Parameters:
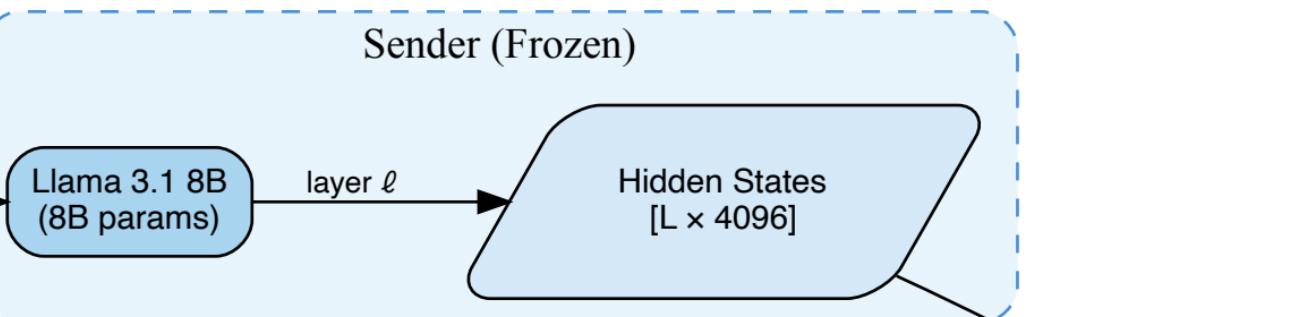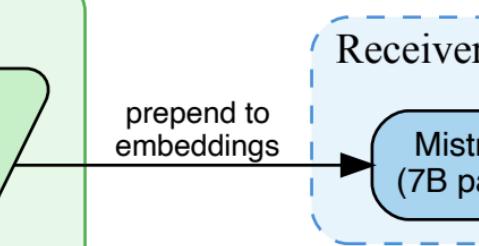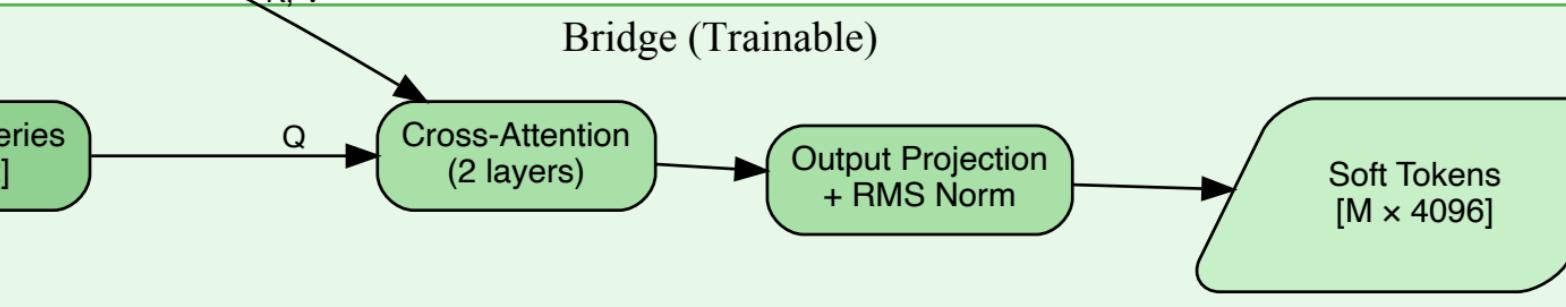Llama: 8B (frozen)
Mistral: 7B (frozen)
Bridge: 188K (trained)

Input Text
"Review: great movie..."

tokenize

Sender (Frozen)

Llama 3.1 8B
(8B params)

layer ℓ

Hidden States
[L × 4096]

K, V

Bridge (Trainable)

Learned Queries
[M × 512]

Q

Cross-Attention
(2 layers)

Output Projection
+ RMS Norm

Soft Tokens
[M × 4096]

prepend to embeddings

Receiver (Frozen)

Mistral 7B
(7B params)

generate

Classification
"positive"