

FLAN-T5 Architecture

ENCODER

DECODER

Self-Attention

Feed-Forward

Layer Norm

*Input: Article Text
(up to 512 tokens)*

**Cross-
Attention**



Masked Self-Attn

Cross-Attention

Feed-Forward

Layer Norm

*Output: Summary
(typically 50-150 tokens)*

Model Sizes:

Small (80M)

Base (250M)

Large (780M)

XL (3B)

XXL (11B)