MedPaLM 2 Architecture

540B
Parameters

86.5%
MedQA Score
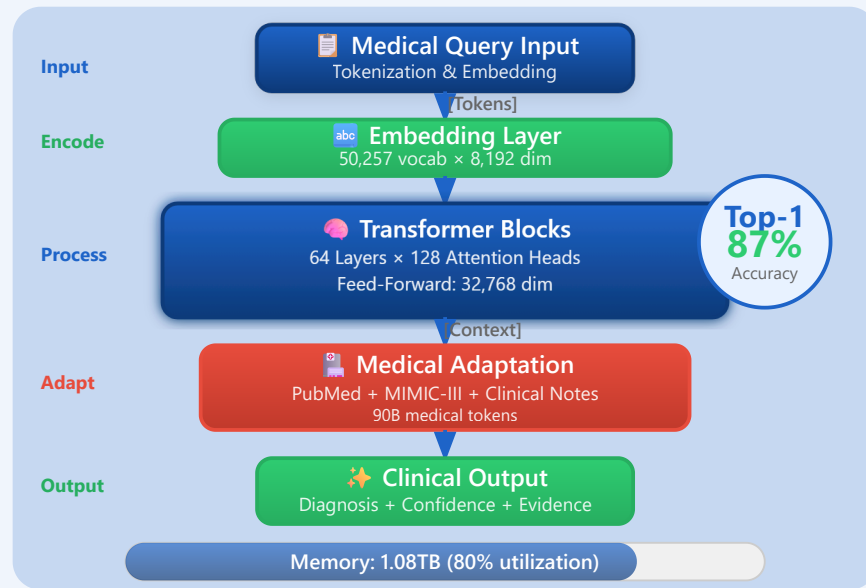
Flan-PaLM
Base Model

64
Layers

📋 Medical Query Input
Tokenization & Embedding

[Tokens]

🔤 Embedding Layer
50,257 vocab × 8,192 dim

🧠 Transformer Blocks
64 Layers × 128 Attention Heads
Feed-Forward: 32,768 dim

Top-1
87%
Accuracy

[Context]

🏥 Medical Adaptation
PubMed + MIMIC-III + Clinical Notes
90B medical tokens

✨ Clinical Output
Diagnosis + Confidence + Evidence

Memory: 1.08TB (80% utilization)

Input
Encode
Process
Adapt
Output

🎯 Instruction Tuning
Medical task-specific fine-tuning

💡 Chain-of-Thought
Step-by-step clinical reasoning

🚀 Few-Shot Learning
5-shot prompting for rare diseases

🛡️ Safety Alignment
Medical ethics & harm reduction

📚 Knowledge Base
UpToDate, PubMed integration

🔄 Continuous Learning
Weekly medical literature updates