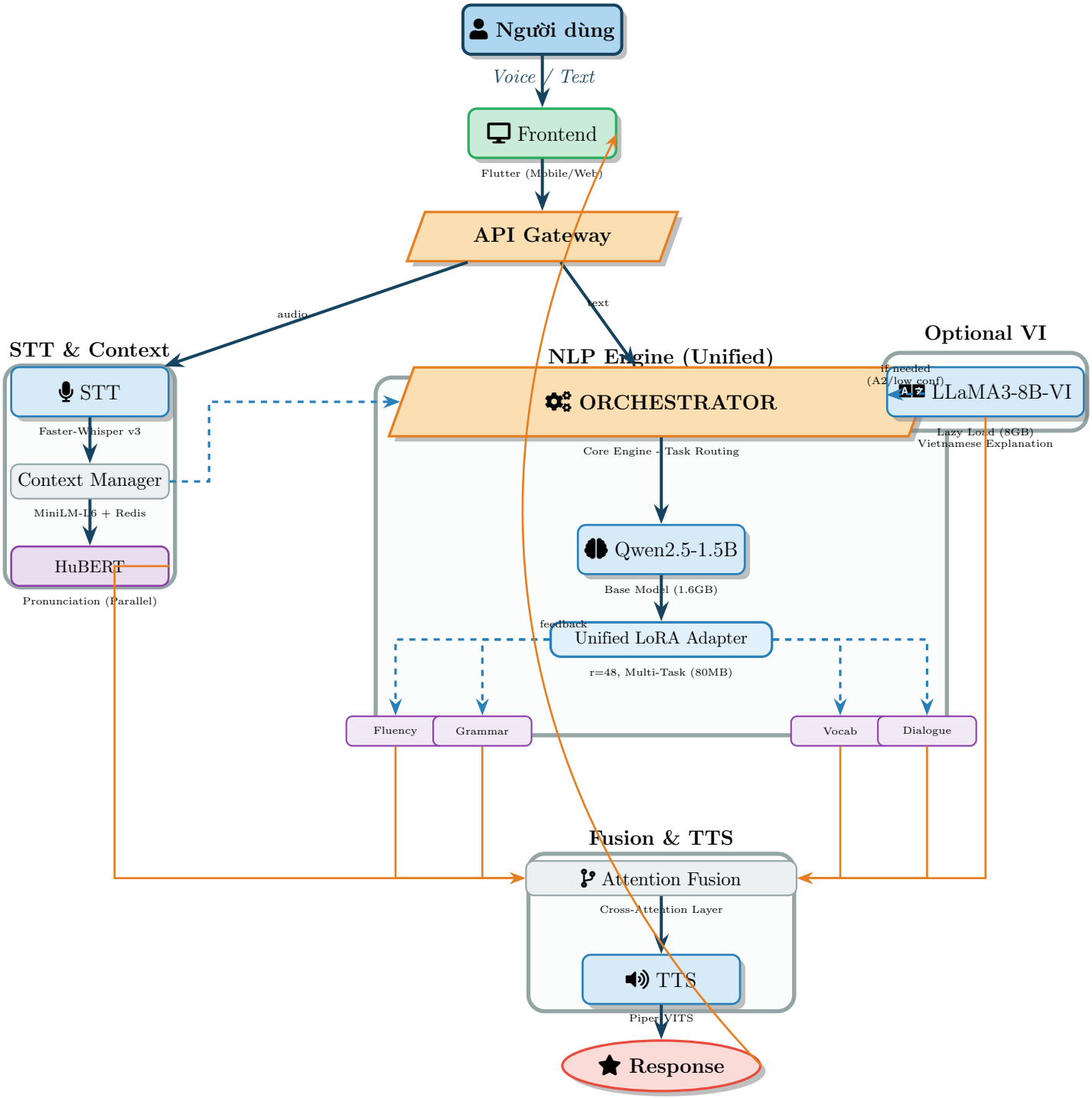
KIẾN TRÚC HỆ THỐNG LEXILINGO v2.0

AI Hỗ trợ Học Tiếng Anh - Optimized Architecture



LEGEND:

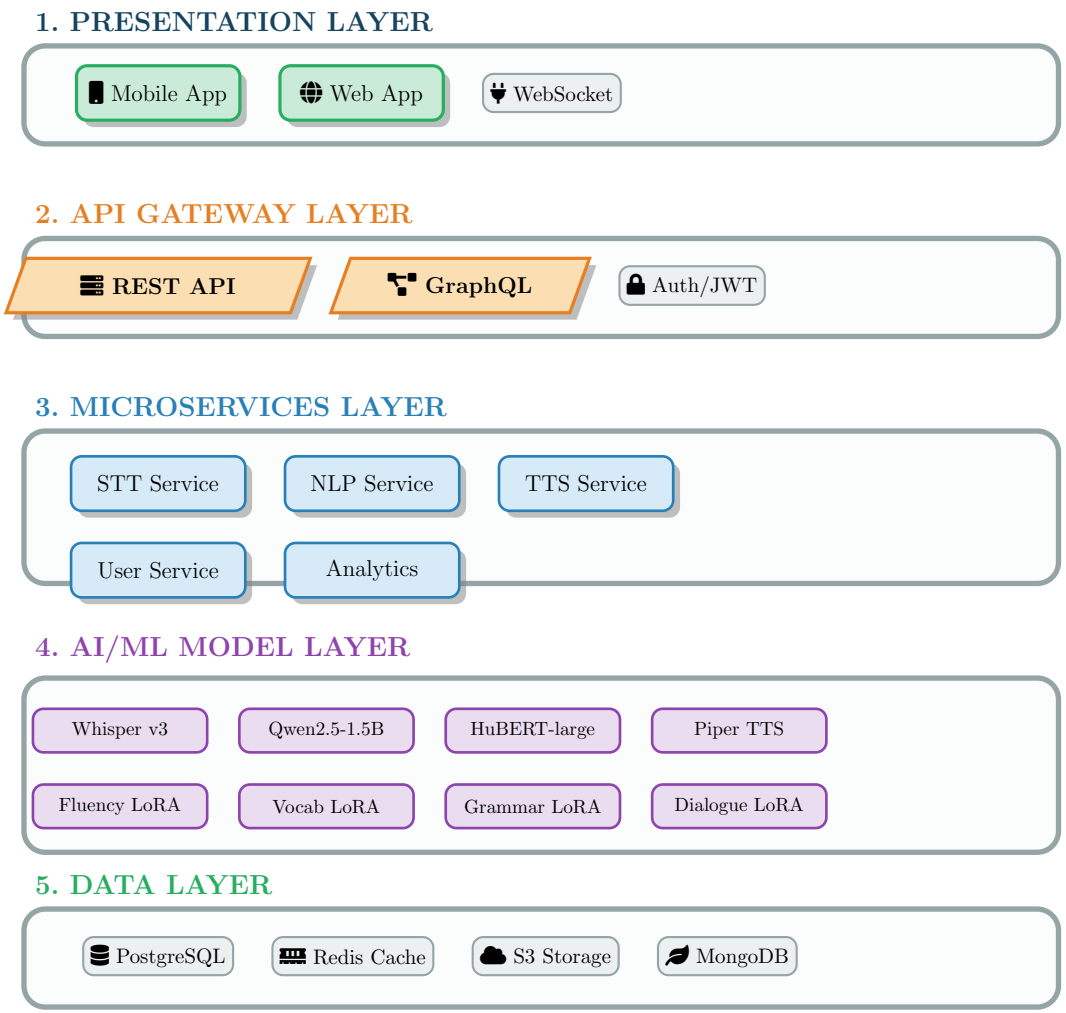
- Main flow
- Data
- Feedback

MODELS v2.0:
STT: Faster-Whisper
• 244MB, <100ms
Context: MiniLM-L6
• 22MB, 15ms
NLP: Qwen2.5-1.5B
• Base: 1.6GB
• Unified LoRA: 80MB
• All tasks: 100-150ms
VI: LLaMA3-8B (lazy)
• 8GB, load on demand
Pronunciation: HuBERT
• 960MB, parallel
TTS: Piper VITS
• 30-60MB

IMPROVEMENTS:
✓ Latency: -56% (350ms)
✓ Memory: -60% (4.8GB)
✓ Cache hit: 45%
✓ Switching: <1ms
✓ Lazy loading enabled

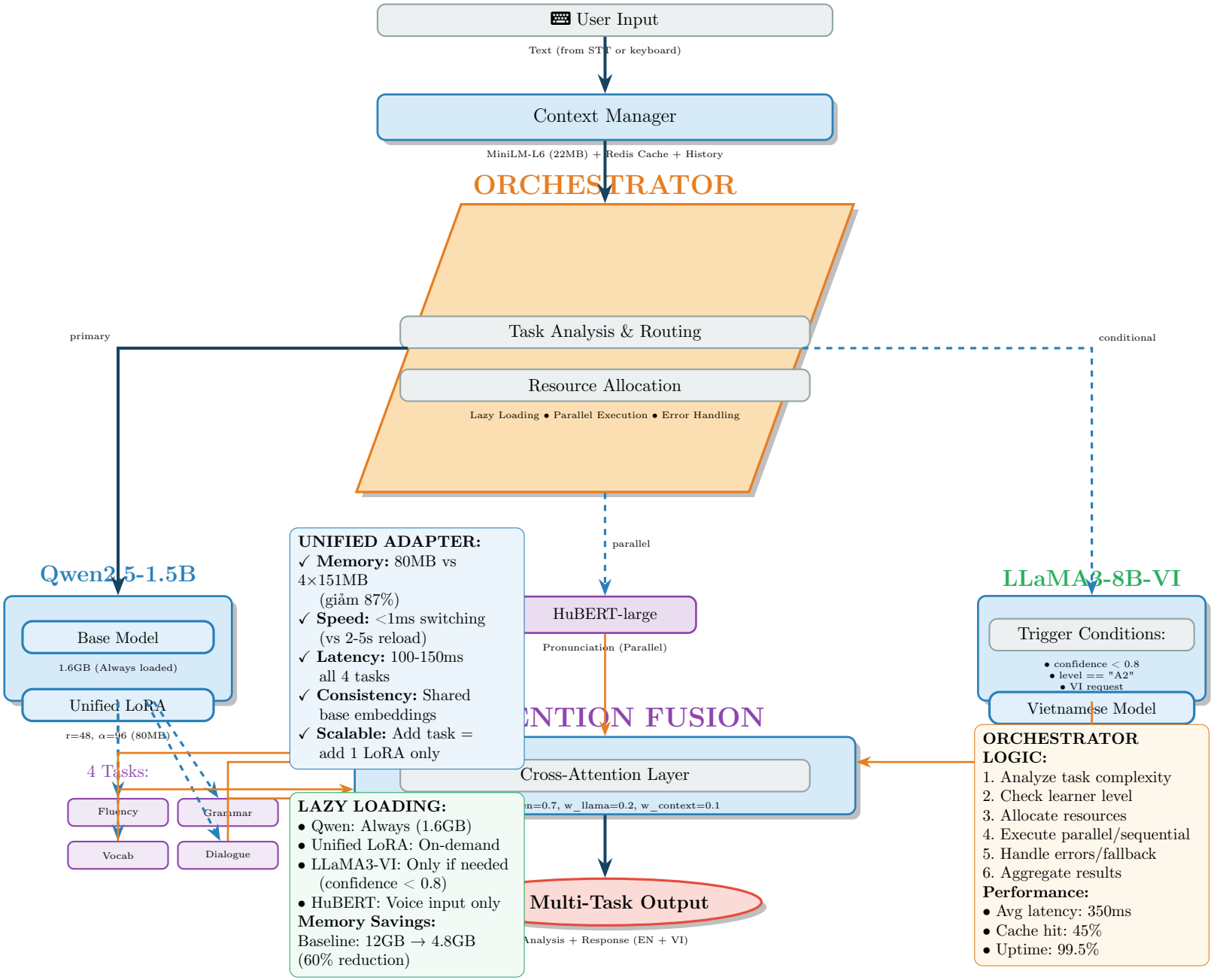
KIẾN TRÚC 5 TẦNG

Phân tầng chi tiết hệ thống



UNIFIED NLP ENGINE v2.0

Kiến trúc Multi-Task Optimized


User Input
Text (from STT or keyboard)
Context Manager
MiniLM-L6 (22MB) + Redis Cache + History
ORCHESTRATOR
Task Analysis & Routing
Resource Allocation
Lazy Loading • Parallel Execution • Error Handling
primary
conditional
parallel
Qwen2.5-1.5B
Base Model
1.6GB (Always loaded)
Unified LoRA
r=48, α=96 (80MB)
4 Tasks:
Fluency
Grammar
Vocab
Dialogue
UNIFIED ADAPTER:
✓ Memory: 80MB vs 4×151MB (giảm 87%)
✓ Speed: <1ms switching (vs 2-5s reload)
✓ Latency: 100-150ms all 4 tasks
✓ Consistency: Shared base embeddings
✓ Scalable: Add task = add 1 LoRA only
HuBERT-large
Pronunciation (Parallel)
ENTION FUSION
Cross-Attention Layer
n=0.7, w_llama=0.2, w_context=0.1
LLaMA3-8B-VI
Trigger Conditions:
• confidence < 0.8
• level == "A2"
• VI request
Vietnamese Model
ORCHESTRATOR LOGIC:
1. Analyze task complexity
2. Check learner level
3. Allocate resources
4. Execute parallel/sequential
5. Handle errors/fallback
6. Aggregate results
Performance:
• Avg latency: 350ms
• Cache hit: 45%
• Uptime: 99.5%
LAZY LOADING:
• Qwen: Always (1.6GB)
• Unified LoRA: On-demand
• LLaMA3-VI: Only if needed (confidence < 0.8)
• HuBERT: Voice input only
Memory Savings:
Baseline: 12GB → 4.8GB (60% reduction)
Multi-Task Output
Analysis + Response (EN + VI)