1 AI Integration Strategy for Adaptive Dialogue

To answer our second research question— **What integration strategies for STT, TTS, and conversational AI enable adaptive, pedagogically-aligned dialogue?**— we designed and specified a hybrid, service-oriented integration strategy. This approach is the cornerstone of our architecture, as it directly addresses the core conflict between the non-functional requirements of two-tier latency (NFR1) and cost efficiency (NFR3), and the functional requirements for high-fidelity, adaptive AI interaction (FR1, FR2, FR3, FR4, FR5).

1.1 The Hybrid AI Integration Model

Our integration strategy is founded on a hybrid processing model that balances on-device and cloud computation. This model is visually specified in Figure 1.

This architecture is defined by two distinct processing loops that run in parallel:

- **Loop 1: The Immersion Loop (<200ms).** For latency-critical, simple interactions (e.g., a simple greeting, a "yes/no" answer), we leverage a lightweight on-device AI SDK. This local processing is designed to meet our sub-200ms latency target (NFR1), preserving VR immersion.
- **Loop 2: The Quality Loop (<1.5s).** For complex conversational tasks, the user's speech is simultaneously streamed to high-performance cloud services. This loop is responsible for high-accuracy streaming STT (FR1), phoneme-level pronunciation analysis (FR4), complex dialogue generation (FR2), adaptive difficulty (FR5), and natural TTS (FR3). This loop is engineered to complete within our 1.5-second target (NFR1).

1.2 Justification: Architectural Trade-offs

The choice of a hybrid architecture is a deliberate compromise, designed to resolve the competing non-functional requirements. We evaluated this model against the two primary alternatives, with the trade-offs summarized in Table 1.

As the analysis shows:

- **A Cloud-Centric Model** fails our primary latency requirement (NFR1). Even simple feedback would incur a full network round-trip, far exceeding the sub-200ms threshold required to maintain VR immersion.
- **An On-Device-Centric Model** fails our functional requirements. The computational constraints of current mobile VR hardware make it unfeasible to run the large-scale models required for high-accuracy STT (FR1), complex dialogue generation (FR2), and natural, prosodic TTS (FR3).

Therefore, our hybrid integration strategy is the only specified solution that satisfies all competing requirements. It uses on-device processing to satisfy the immediate immersion loop, while strategically delegating complex tasks to the cloud to ensure pedagogical quality.

1.3 Enabling Pedagogically-Aligned Dialogue

Our integration strategy is driven by pedagogical requirements derived from Second Language Acquisition (SLA) theory, not just technical feasibility. The "Conversation Service" (Figure 1) is the core component that enables this. Figure 2 specifies the interaction flow for this service.

The integration of AI systems is fundamentally driven by three pedagogical objectives that form the conversational loop:

- **1. Goal-Oriented Conversation (Constrained Dialogue)**

  The conversational AI (LLM) is not a generic chatbot. Dialogue generation (FR2) is deliberately "constrained by scenario graphs" aligned with specific user goals (Sec 3.1) and pedagogical principles [2]. This structured approach ensures that the conversation remains focused on the learning objective, providing a controlled environment for the student to engage in meaningful interaction, which aligns directly with Long's Interaction Hypothesis [4].

- **2. Adaptive Difficulty (Implementation of Krashen's i+1)**

  The system is designed to be dynamically adaptive. The Conversation Service manages the LLM and interfaces with the User Service to retrieve the learner's profile and progress. This enables continuous adaptive difficulty adjustment (FR5). By ensuring that the conversational input is slightly above the learner's current competence level, the system directly implements Krashen's "Comprehensible Input (i+1)" hypothesis [1], thereby maximizing learning potential without overwhelming the student.

- **3. The STT-TTS Interaction Loop (Output Hypothesis)**

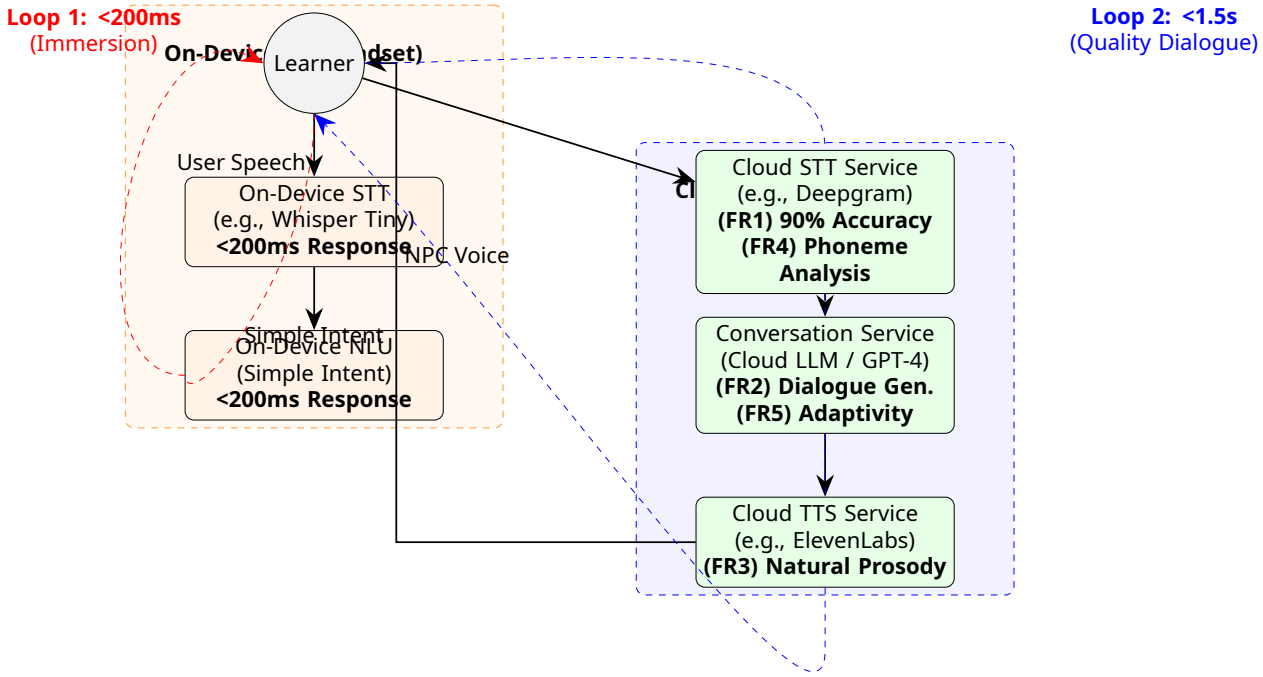

Figure 1: The Hybrid AI Integration Model, showing the two-tier latency loops (NFR1) by distributing STT, NLU, LLM, and TTS workloads.

Table 1: Comparison of Architectural Integration Strategies

Criteria	Cloud-Centric Model	On-Device-Centric Model	MundoVR Hybrid Model (Our Choice)
Latency (NFR1)	**FAILS** (High latency)	**PASS** (Low latency)	**PASS** (Two-tier latency)
AI Quality (FR1-5)	**PASS** (High quality)	**FAILS** (Low quality)	**PASS** (High quality via cloud)
Immersion	**FAILS** (Broken by lag)	**PASS** (Maintained)	**PASS** (Maintained by <200ms loop)
Scalability (NFR2)	High (Cloud-native)	N/A (Device-bound)	High (Cloud components)
Cost (NFR3)	High (All tasks)	Low (No cloud)	**Optimized** (On-device for 40%)

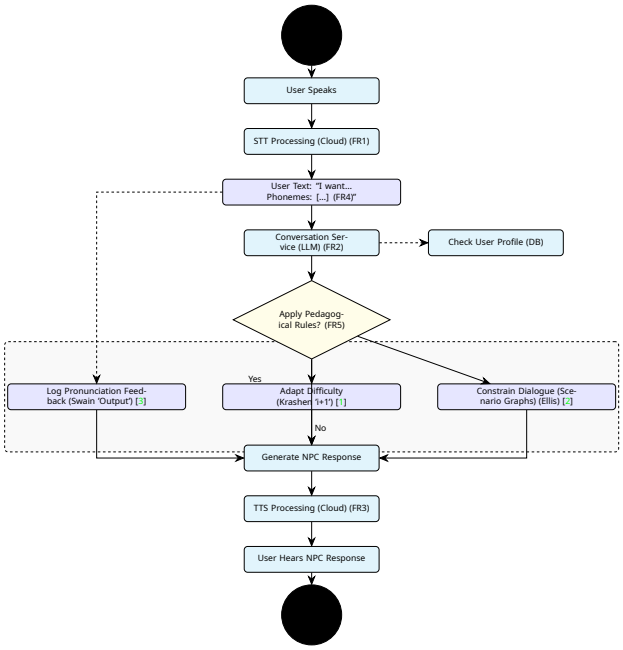

Figure 2: Activity Diagram of the Adaptive Pedagogical Loop, illustrating how SLA theories [1, 3, 4] are integrated into the conversational flow to answer RQ2.

The STT and TTS services form the technical basis of the learning loop. The TTS provides natural, prosodic output (FR3), serving as a comprehensible and realistic language model for the learner. The STT service, crucially, performs phoneme-level pronunciation analysis (FR4) on the student's speech. This analysis generates data for immediate, actionable feedback on the user's *production*. This mechanism strongly supports Swain's Output Hypothesis [3], which posits that the act of producing language (output) pushes learners to notice gaps in their knowledge, thus driving language acquisition.

In summary, our integration strategy answers RQ2 by specifying a decoupled, hybrid architecture that manages technical trade-offs while ensuring that the STT, TTS, and LLM components are interconnected to serve a pedagogically-grounded, adaptive conversational loop.

References

[1] S. D. Krashen, *Principles and Practice in Second Language Acquisition*. Oxford: Pergmon Press, 1982. [Online]. Available: http://www.sdkhen.com/content/books/principles_and_practice.pdf.

[2] R. Ellis, "Principles of instructed second language acquisition," *CAL Digest*, 2008. [Online]. Available: https://www.cal.org/resource/principles-of-instructed-second-language-acquisition/.

[3] M. Swain, "Three functions of output in second language learning," in *Principles and Practice in Applied Linguistics: Studies in Honour of H. G. Widdowson*, G. Cook and B. Seidlhofer, Eds., Oxford: Oxford University Press, 1995, pp. 125–144.

[4] M. H. Long, "The role of the linguistic environment in second language acquisition," in *Handbook of Second Language Acquisition*, W. C. Ritchie and T. K. Bhatia, Eds., San Diego, CA: Academic Press, 1996, pp. 413–468.