Roadmap

Creating a comprehensive roadmap for the full deployment of **EchoHilf**, a voice chatbot solution that incorporates advanced features such as **SpeechToText (STT)**, **Retrieval Augmented Generation (RAG)** technology, human-in-the-loop for quality assurance, and **TextToSpeech (TTS)**, involves detailed planning across several stages.
This roadmap also considers potential implementation partners and provides a rough cost estimation framework.

Phase 1: Planning and Requirements (1 month)

- **Objective Setting**: Define specific goals for EchoHilf, including functionalities, user experience expectations, and privacy compliance objectives.
- **Requirements Gathering:** Collaborate with internal stakeholders and potential users to gather detailed requirements.
- **Implementation Partners:**
    - For STT and TTS: As the first implementation, to address these functionalities, state-of-the-art models can be implemented. This would bring as partners the API vendors, such as OpenAI and Cohere. This partnership will change, as we are considering switching to a local model, for better customization and privacy issues.
    - For RAG Technology: Collaboration with AI and NLP research institutes or companies with expertise in machine learning models.
    - Cost Estimation: Preliminary analysis suggests a range from $5,000 to $10,000 for initial development, depending on the complexity and the chosen partners.

Phase 2: Design and Prototype (2-3 months)

- System Architecture Design: Design the technical architecture, ensuring it supports all components securely and efficiently.
- Infrastructure Planning: Determine the infrastructure needed for telephony integration, web chatbot deployment, and SMS services.
- Prototype Development: Develop an alpha product to test with selected customers basic functionalities and privacy features.
- User Testing and Feedback: Conduct early user testing to gather feedback and refine the prototype.
- Cost Estimation: Requirements engineering costs are estimated between $10,000 and $15,000, including partner services and internal resource allocation.

Phase 3: Development and Integration (4 months)

- Development Sprints: Organize development work into sprints, focusing on different components of EchoHilf.
- Integration: Seamlessly integrate STT, RAG, human-in-the-loop oversight, and TTS technologies.
- Telephony and Web Integration: Integrate the chatbot with telephone services and the website, ensuring seamless user experiences across both channels.

- Privacy Measures Implementation: Incorporate encryption, anonymization, and secure data handling practices from the start.
- Implementation Partners: Engage with cybersecurity firms for privacy protection solutions and legal consultants for compliance verification.
- Cost Estimation: Development and integration costs are estimated between $30,000 and $40,000, including partner services and internal resource allocation.

Phase 4: Compliance and Security Testing (2 months)

- Compliance Audits: Work with legal partners to ensure EchoHilf meets all relevant data protection regulations.
- Security Testing: Perform comprehensive security testing, including penetration testing and vulnerability assessments, possibly with partners specializing in cybersecurity.
- Cost Estimation: Approximately $15,000 to $20,000 for external testing and auditing services.

Phase 5: Training and Deployment (1-2 months)

- Staff Training: Conduct training sessions for internal teams, focusing on system operation, privacy protection practices, and user support.
- Deployment: Roll out EchoHilf in stages, starting with a controlled launch to monitor performance and gather user feedback.
- Implementation Partners: Training organizations specialized in technology and privacy law education.
- Cost Estimation: Training and initial deployment costs can range from $10,000 to $15,000.

Phase 6: Post-Deployment Support and Optimization

- Continuous Improvement: Use feedback and usage data to refine and enhance EchoHilf, adding new features and improving privacy protection.
- Cost Estimation: Ongoing costs are estimated at $5,000 to $10,000 per month, depending on the scale of operations and enhancements.

Total Estimated Cost: $75,000 to $110,000

Additional variable costs for the infrastructure

The monthly cost for a server with adequate CPU and memory is estimated at €58,40/month.
For the API calls to the models, the calculations are based on specific token usage per call.

For GPT-4 API calls, each call is estimated to involve 150 tokens for input and 30 tokens for output. The cost per token for input is $10,00 per 1 million tokens, and for output, it's $30,00 per 1 million tokens. Combining these costs, the estimated cost per GPT-4 API call is approximately $0,0024.

Additionally, for Cohere API calls, with an estimated total of 180 tokens per call (combining input and output), and a cost of $0,10 per 1 million tokens, the cost per call is approximately $0,000018. When considering a higher cap scenario, where the costs are multiplied by 10 for increased usage, the estimated cost per user question adjusts to about $0,00018.

These costs reflect the calculations based on the specified token usage and prices per million tokens for both GPT-4 and Cohere, providing a framework for estimating the operational costs associated with each API call under the outlined conditions.