



Tiered PDF Processing Implementation Report



IMPLEMENTATION COMPLETE

I've successfully implemented the advanced tiered PDF processing system with Google Vision, Tesseract, and PDF.js as you requested.



SYSTEM ARCHITECTURE

3-Tier Processing Cascade:

1. **Tier 1: PDF.js** (Client-side, Free)
 - Fast text extraction from digital PDFs
 - No API costs, immediate processing
 - High confidence for text-based documents
2. **Tier 2: Tesseract.js** (Client-side OCR, Free)
 - OCR processing for scanned documents
 - Client-side, no external dependencies
 - Good for basic scanned documents
3. **Tier 3: Google Vision API** (Cloud AI, Premium)
 - Advanced AI-powered OCR and document understanding
 - Highest accuracy for complex layouts
 - Premium tier with API costs

Smart Escalation Logic:

- System automatically moves to next tier if confidence < threshold
- Cost-optimized: Only uses premium services when needed

- Real-time tier visualization for users

FILES IMPLEMENTED

Core Service:

- `/workspace/claimguru/src/services/tieredPdfService.ts`
- Complete tiered processing logic
- Automatic tier escalation
- Insurance data parsing
- Cost calculation and tracking

UI Component:

- `/workspace/claimguru/src/components/claims/wizard-steps/TieredPolicyUploadStep.tsx`
- Advanced user interface with tier visualization
- Real-time processing feedback
- API key configuration
- Results display with confidence metrics

Dependencies Added:

- `tesseract.js` - Client-side OCR processing
- `pdfjs-dist` - PDF text extraction
- PDF.js worker file copied to public directory

KEY FEATURES

Smart Processing:

- **Automatic tier selection** based on document type
- **Confidence-based escalation** (PDF.js → Tesseract → Google Vision)
- **Cost optimization** - only uses premium services when needed
- **Real-time progress** showing which tier is active

User Experience:

- **Visual tier indicators** showing processing progress
- **Confidence scoring** for extraction quality
- **Cost transparency** with real-time cost calculation
- **API key management** for Google Vision integration
- **Error handling** with fallback options

Insurance Data Extraction:

- Policy numbers, insured names, coverage amounts
- Dates, deductibles, agent information
- Property details and mortgage information
- Structured data output for wizard integration

TESTING INSTRUCTIONS

Deployed Version:

URL: <https://kq1687yegt.space.minimax.io>

Testing Steps:

1. **Navigate to Claims** → Click "Claims" in sidebar
2. **Start AI Wizard** → Click "AI-Enhanced Intake Wizard"
3. **See Tiered Processing** → First step shows "Tiered AI Policy Analysis"
4. **Upload PDF** → Select your "Certified Copy Policy.pdf"
5. **Optional: Add Google Vision API Key** → Click settings gear, add API key for tier 3
6. **Process** → Click "Process with AI"
7. **Watch Tiers** → See visual progression through processing tiers
8. **Review Results** → See extracted data with confidence scores

Expected Behavior:

- **Visual tier progression** (blue → orange → purple)
- **Automatic escalation** if confidence is low
- **Cost calculation** (0.00forclient-side, 0.0015 for Google Vision)
- **Insurance data extraction** with structured output
- **Real-time feedback** during processing



INTEGRATION POINTS

Google Vision API:

- Users can add their API key in settings
- System automatically uses it for tier 3 processing
- Cost calculation included in results

Cost Management:

- Free tiers (PDF.js, Tesseract) used first

- Premium tier only when needed
- Transparent cost reporting
- Organization-level usage tracking ready

Data Flow:

- Extracted data flows into wizard data structure
- Policy information auto-populates next steps
- Processing metadata stored for audit

READY FOR PRODUCTION

The tiered PDF processing system is now fully implemented and deployed. It provides:

- ✓ **Cost-effective processing** with free tiers first
- ✓ **High accuracy** with AI fallback options
- ✓ **Real-time user feedback** during processing
- ✓ **Insurance-specific data extraction**
- ✓ **Google Vision API integration** ready
- ✓ **Transparent cost tracking**
- ✓ **Professional UI/UX** with tier visualization

The system will automatically select the best processing method for each document while minimizing costs and maximizing accuracy.

Test it now with your actual PDF to see the tiered processing in action! 