Assumptions & Trade-offs

Assumptions

1. **PDF Format**: Assumed clean text extraction from PDFs (not scanned images)
2. **Document Structure**: Assumed consistent section headers and table formats
3. **Page Sampling**: Classified documents using first 5 + last 2 pages rather than full content
4. **Single File Processing**: Built for manual trigger per file (scalable to folder monitoring)
5. **Data Quality**: Assumed OpenAI extracts 95%+ accuracy with proper prompting

Trade-offs

Accuracy vs Speed

- **Chosen**: Sampled pages for classification (faster)
- **Alternative**: Full document analysis (more accurate but slower)
- **Rationale**: Classification rarely needs full document; saves tokens and time

Cost vs Reliability

- **Chosen**: gpt-4o-mini for classification, gpt-4o for extraction
- **Alternative**: gpt-4o for everything (more accurate)
- **Rationale**: Classification is simple; extraction needs precision

Simplicity vs Flexibility

- **Chosen**: Code nodes for routing instead of complex IF/SWITCH logic
- **Alternative**: Native n8n routing nodes
- **Rationale**: Code is clearer, easier to debug, more maintainable

Error Handling

- **Current**: Basic try-catch with console logging
- **Future**: Comprehensive error logging to Logs sheet with retry logic
- **Rationale**: Time constraint; focused on happy path first

Future Enhancements

1. Batch processing with Google Drive folder watch
2. Advanced error handling and retry mechanisms
3. Support for scanned PDFs (OCR integration)
4. Data validation rules before sheet insertion
5. Email notifications on completion/errors
6. Multi-language support
7. Lease extraction (third branch, time permitting)