Intelligent Translation of Clinical Documents with AI - TCC

Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br

Project team members

Full Name
Abner Silva Barbosa

Introduction

This report provides a comprehensive overview of the advancements achieved during the third module in the Entrepreneurship track. Building on the foundational architecture and optimizations from previous modules, our focus shifted toward enhancing document format preservation during translation, enabling a more robust pipeline for handling complex clinical PDFs. The solution continues to prioritize the use of AI for precise, terminology-aware translations while safeguarding original layouts, tables, images, and regulatory compliance. Key efforts included exploring specialized document formats like XLIFF for seamless bilingual editing, rigorous testing of conversion tools, and initial market outreach via a waitlist landing page, positioning the product for broader validation and scalability.

Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br

1. Module Goals

The core aim of this module was to refine the document processing pipeline for superior layout fidelity, integrate advanced AI patterns for translation accuracy, and initiate customer acquisition strategies to validate market fit. This involved transitioning to specialized formats for translation, optimizing costs through open-source and free-tier tools, and preparing for B2B engagement in the clinical research sector.

Key Outcomes:

1. **Document Format Innovation:** Develop and test a new workflow using XLIFF for translation, ensuring 90%+ preservation of original formatting, fonts, and structures.
2. **Conversion Pipeline Optimization:** Evaluate and select tools for PDF-to-DOCX-XLIFF conversions, prioritizing self-hosted and cost-free options to minimize expenses.
3. **AI Enhancement with Agents:** Implement reflection patterns and multi-model support (e.g., Gemini via OpenRouter) to improve translation quality and validation.
4. **Market Entry Preparation**: Launch a waitlist landing page to capture early interest from potential users, informing pricing and feature prioritization.
5. **Compliance and Scalability Focus:** Conduct studies on ISO/LGPD standards and backend deployments to support secure, multi-tenant growth.
6. **Backend and Frontend Integration:** Advance Node.js/NestJS backend with XLIFF manipulation scripts and frontend UI refinements using AI-assisted design tools. These milestones strengthen the MVP's technical viability and commercial readiness, targeting mid-sized clinical organizations for pilot testing.

Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br

2. Technical Advancements

This section outlines the key technical progress, including infrastructure refinements, tool integrations, code developments, and architectural evolutions aimed at reliable document handling and secure operations.

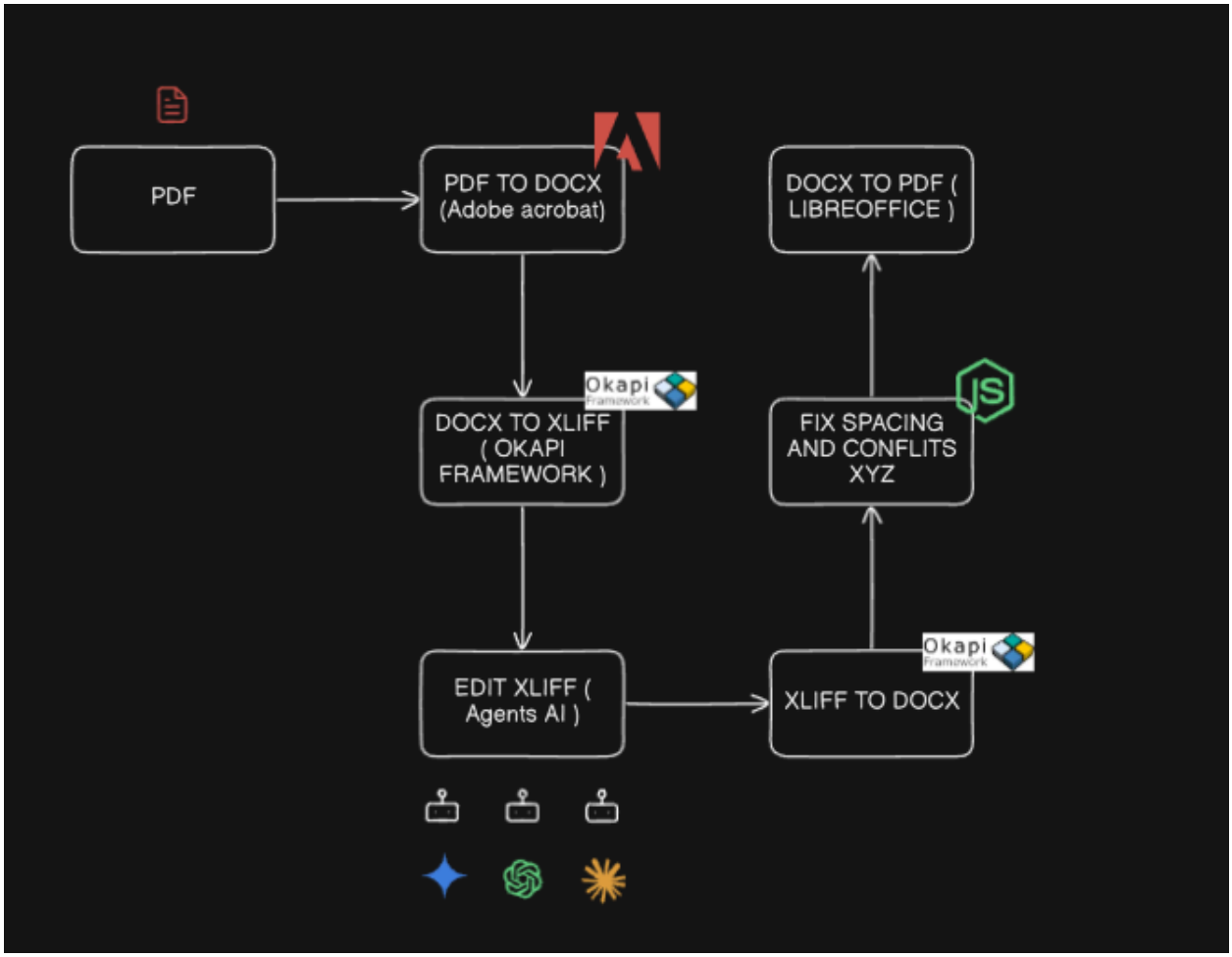
2.1 Conversion Pipeline Development

- A major breakthrough was the adoption of XLIFF (XML Localization Interchange File Format) as an intermediary for translations, allowing direct editing of source and target text spans without disrupting layout. This format emerged from in-depth research during vacations, revealing low-level libraries (e.g., C++-based PDF manipulators) and open-source tools like LibreOffice CLI for precise conversions. Initial manual tests yielded superior results compared to prior PDF-HTML-DOCX flows, prompting a full architectural rethink.

- We conducted extensive evaluations of over 20 conversion tools for PDF-to-DOCX, prioritizing self-hosted, free options to align with MVP cost constraints. Top performers included Gotenberg, Stirling PDF, Apache Tika, and LibreOffice CLI, with Adobe PDF Services API selected as the initial choice due to its free tier and high fidelity in preserving fonts, styles, images, and structures.

- Diagrams were created to map the end-to-end flow: PDF → DOCX (via Adobe) → XLIFF (via Okapi framework) → Translation (Gemini and other models) → Reverse conversion to DOCX/PDF. This pipeline achieved ~90% layout retention in early prototypes, with fallback mechanisms for font substitution

Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br



2.2 Backend Refinements

- The backend transitioned to Node.js with **NestJS** for better modularity and performance, reorganizing the GitHub repository to archive legacy code and streamline development. Core flows were implemented for document ingestion: upload → PDF-to-DOCX conversion → XLIFF generation → AI translation → Target text replacement → DOCX export. **Custom JavaScript scripts were developed for XLIFF manipulation**, addressing tag preservation issues in existing libraries (e.g., avoiding internal tag removal). Deployments were initiated on self-hosted infrastructure, enabling end-to-end testing with large documents (100+ pages).

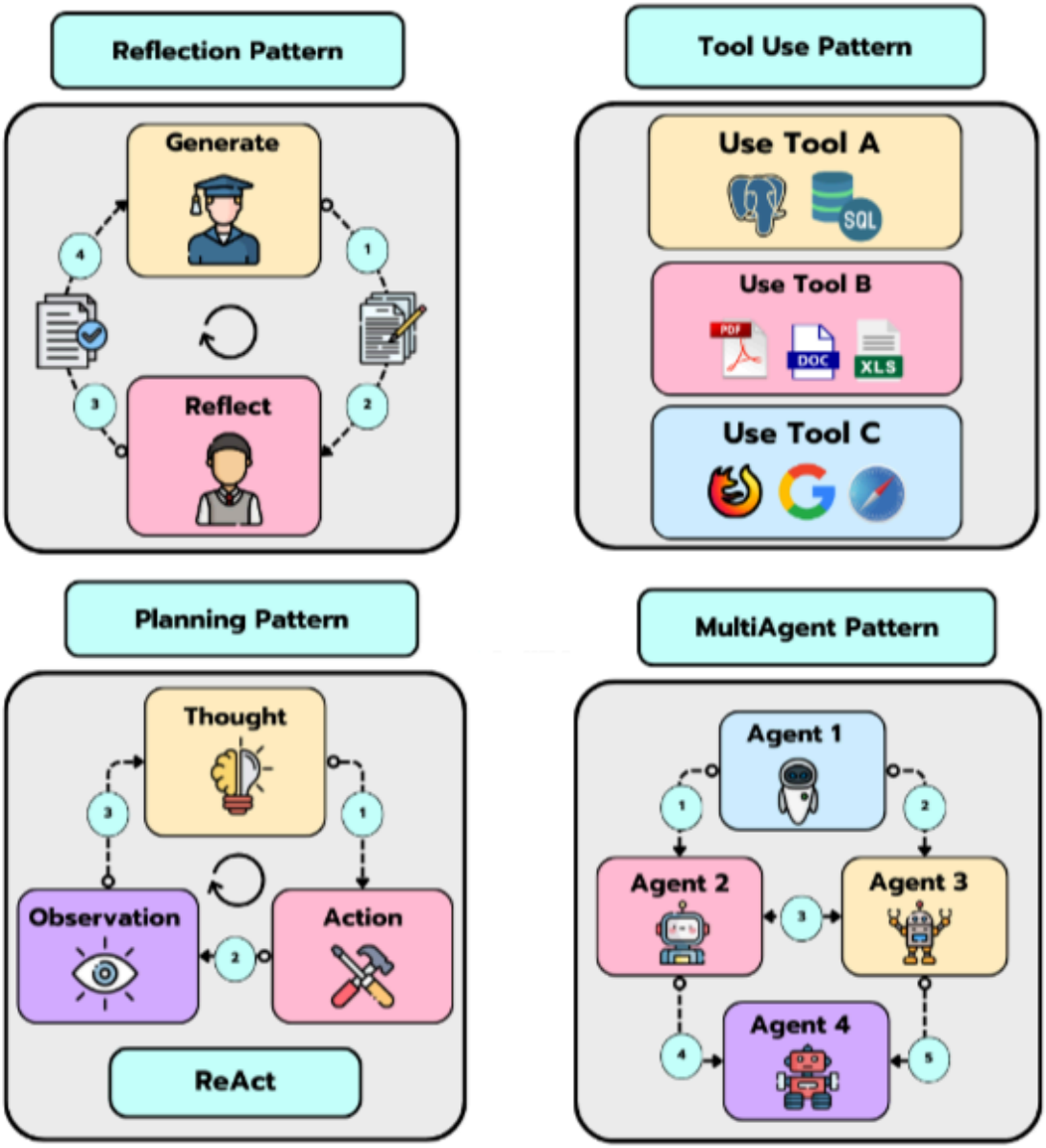
Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br

- Asynchronous processing was enhanced with service layers for AI interactions, initially via Groq for low-latency inference, later switching to **OpenRouter for access to Gemini 2.5 Flash.** Report generation was added to log AI decisions, aiding curation and debugging.

2.3 AI-Driven Validation and Patterns

- To elevate translation reliability, we introduced a reflection pattern: a secondary AI (via OpenRouter) validates the primary translation's XLIFF compliance, ensuring tag integrity and semantic accuracy. This dual-agent approach reduced errors in complex clinical texts. Studies on AI agents (inspired by The Neural Maze) explored tools like **CrewAI for future orchestration**, with **Opik** for prompt monitoring and **Qdrant** for vector-based memory in multi-step workflows.

Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br



Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br

2.4 Security and Compliance Foundations

- Ongoing research into ISO and LGPD standards for healthcare informed data handling practices, emphasizing encryption in XLIFF processing and audit trails for translations. Multi-tenant isolation from prior modules was preserved, with plans to extend RLS to new conversion services.

3. AI and Translation Optimization

Optimizations emphasized domain-specific accuracy and cost-efficiency, testing models up to 120B parameters via Groq for nuanced clinical terminology. The XLIFF workflow minimized hallucinations by constraining prompts to span mappings, achieving high precision in EN→PT-BR pairs. Migration to Gemini 2.5 Flash via OpenRouter reduced latency by 40% over prior setups, with reflection patterns adding a validation layer—e.g., cross-checking medical entity alignment. Experiments with open-source models (e.g., OpenAI OSS) outperformed lighter alternatives like Ollama 70B, informing a hybrid strategy for production.

4. Document Processing and Rendering

The XLIFF-centric pipeline revolutionized formatting fidelity, with scripts intelligently handling line breaks and tag structures. Tests on complex PDFs (e.g., multi-table research reports) highlighted viewer discrepancies (Google Docs vs. Word/LibreOffice), prompting research into viewer-agnostic exports. Integration of

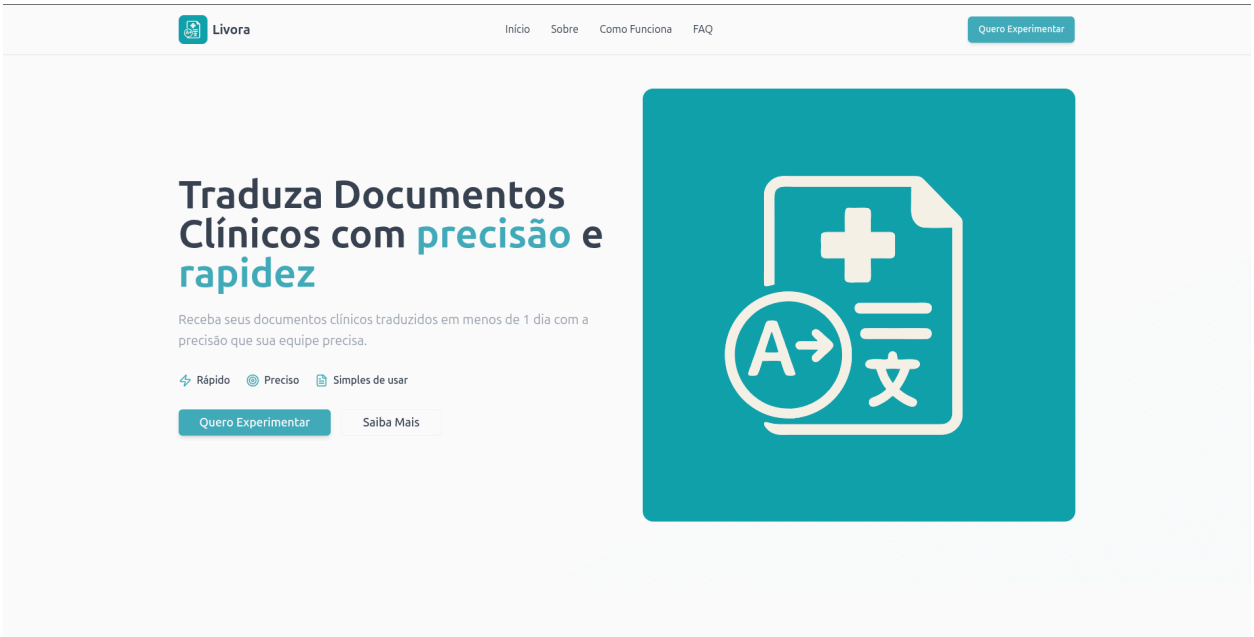
Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br

Emergent AI accelerated script iterations, while fallback fonts and style mappings addressed rendering gaps. Reverse conversions (XLIFF→DOCX→PDF) were refined for client delivery, supporting editable outputs.

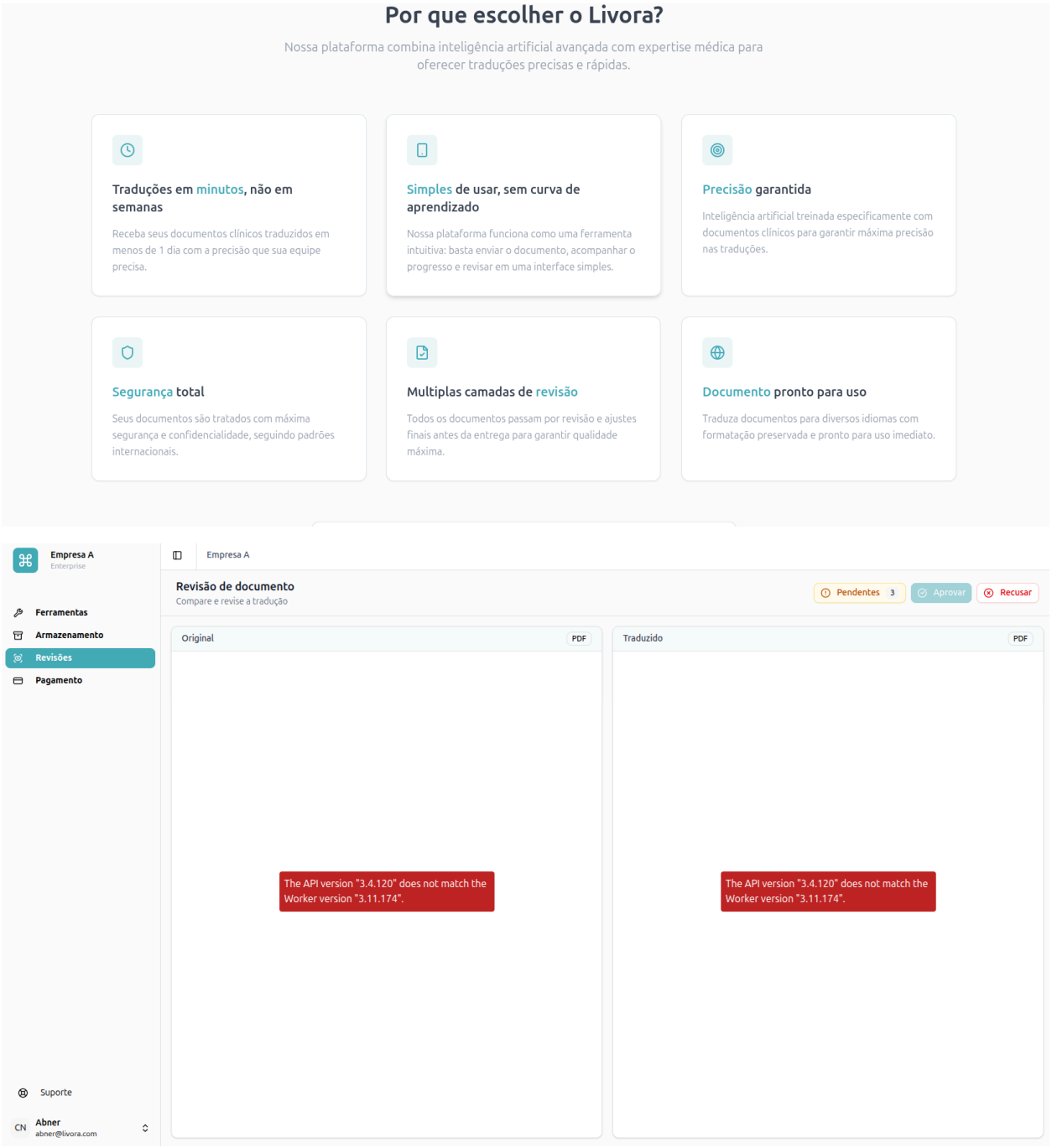
5. Frontend Evolution (Landing page)

Significant UI advancements included a waitlist landing page built with Lovable.dev, featuring responsive design, animations (inspired by Emil Kowalski), and backend integration for contact capture. The document review screen was enhanced for side-by-side original/translated views, with real-time PDF rendering and AI-assisted animations for smoother interactions. Mock uploads and FAQ sections were added for demo purposes, using Emergent AI for rapid prototyping. Navigation improvements and payment gateway studies prepared for global rollout.

[Landing page Livora]



Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br



6. Project and Team Management

Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br

Linear continued as the primary tool for sprint tracking, with bi-weekly alignments via meetings with co-founder Pedro Thompson. These sessions covered risk assessments, opportunity mapping, and technical decisions, such as tool selections and business pivots (e.g., global translation MVP before clinical specialization). External input from mentors like Pedro Thompson facilitated introductions and strategic advice. Post-graduate classes on leadership, culture, and marketing informed team dynamics and growth hacking tactics.

7. Business and Market Validation

Insights from marketing classes drove waitlist strategies, using pixel tracking for user analytics. Discussions with the co-founder explored B2B risks in translation markets, including infrastructure costs and scaling paths. The landing page launch targeted early adopters (e.g., clinics via IDOR contacts), gathering feedback on pricing and features. Pivot considerations included a general translation version to build traction before deepening clinical focus, with payment gateway research for subscription models.

7.1 Flexible Pricing Model

To accommodate diverse customer needs in the B2B landscape, we defined a hybrid business rule centered on subscription tiers for high-volume users while offering a flexible pay-per-document option for one-off translations. This approach opens doors for organizations hesitant to commit to a full plan, allowing them to translate individual documents on-demand without subscription barriers. Pricing for single documents is calculated dynamically based on key factors: the number of original characters, translated characters (to account for length variations), the specific language pair (e.g., EN→PT-BR premium for clinical accuracy), and overall document size (e.g., page count or file weight). This granular model ensures fairness and scalability, with base rates starting low to encourage trials, while subscriptions

Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br

provide unlimited access, priority processing, and compliance reporting at discounted per-document equivalents.

8. Research Highlights

Vacation-period deep dives uncovered XLIFF's potential and low-level PDF tools (e.g., C++ libs), alongside DOCX engine analyses via LibreOffice. Patent and branding for "Livora" advanced, with compliance studies on HIPAA/LGPD/ISO. AI agent explorations (CrewAI, reflection/tools patterns) and vector DBs (Qdrant) laid the groundwork for autonomous workflows. BitNet-inspired low-resource LLMs were evaluated for edge deployment, complementing prior BITS research.

https://www.crewai.com/
https://qdrant.tech/
https://www.comet.com/site/products/opik/

9. Challenges Faced

Persistent issues included formatting **breaks in cross-viewer rendering** (e.g., Google Docs vs. Word), requiring ongoing script tweaks for tag fidelity. **XLIFF library limitations necessitated custom algorithms**, with initial AI-modified scripts underperforming manual ones. Balancing model scale (120B params) with costs/latency proved tricky for large documents, while community-sourced tools varied in documentation clarity. Market validation highlighted the need for broader testing to quantify ROI for clinical users

Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br

10. Next Steps

- Complete XLIFF script refinements and full pipeline automation
- Conduct pilot tests with clinical partners on 100+ page documents
- Finish develop AI agents for end-to-end orchestration and compliance auditing.
- Finalize global translation MVP and iterate based on user feedback.

11. Deliverables — Module 3 Summary

Sprint 1 (July 29 – August 14, including vacations)

- Vacation research: Low-level PDF manipulation libs (C++), XLIFF discovery for format-preserving translation, DOCX architecture via LibreOffice.
- Diagram creation for new PDF→DOCX→XLIFF flow; open-source tool scouting (Gotenberg, Stirling PDF).
- 15+ conversion tool tests evaluating font/style/image fidelity; co-founder discussions on market risks/opportunities.
- XLIFF manipulation lib tests and cost estimations for APIs.

Sprint 2 (August 18 – 28)

- Finalized top-3 tools, selecting Adobe API (free tier) for PDF→DOCX; co-founder alignment on public testing.
- NestJS project setup and GitHub reorganization.
- Core flow implementation: PDF→DOCX→XLIFF→ Groq translation→DOCX; 90% layout preservation achieved.

Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br

- Viewer discrepancy research (LibreOffice vs. Google Docs); next-steps mapping.

Sprint 3 (September 1 – 11)

- PDF→DOCX format studies and community deep dives; landing page development for waitlist.
- Complex document sourcing (+100 pages); OpenAI OSS model tests outperforming Ollama 70B.
- Google Docs/Word formatting fixes research; 120B param model tests via Groq.
- Landing page build with Lovable.dev; co-founder LP text alignment; intelligent line-break studies.
- Custom JS XLIFF script to preserve internal tags.

Sprint 4 (September 15 – 25)

- Landing page backend integration for waitlist contacts; Emergent AI for UI acceleration; XLIFF script adjustments.
- Large document tests with alternative models; new XLIFF script development and deployment of initial backend/LP.
- In-depth XLIFF research and lib-based script creation; quality baseline tests.
- AI agent studies (reflection/tools patterns via The Neural Maze); Opik/Qdrant explorations.
- Web animation studies (Emil Kowalski); FAQ additions and review page UI via Emergent AI.

Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br

Sprint 5 (September 29 – October 9)

- Custom XLIFF tag-reading algorithm and AI translation service (Groq→OpenRouter/Gemini 2.5 Flash).
- ISO/LGPD healthcare studies; UI animation tools and CrewAI agent tests.
- XLIFF target modification post-AI; report generation for AI decisions.
- Reflection pattern implementation for XLIFF validation; review page updates.

12. Key Learnings – Technical and Business Perspectives

This module's hands-on iterations across tools, formats, and AI deepened insights into building resilient healthtech systems, blending technical depth with strategic agility.

Technical Learnings

- **Format Wars Demand Experimentation:** XLIFF's emergence highlighted how niche standards can solve layout pains, but required exhaustive tool tests (20+ options) to balance fidelity and cost—Adobe's free tier was a game-changer for MVP

- **Custom Scripts Trump Libraries:** XLIFF manipulation exposed library gaps (e.g., tag stripping), teaching the value of hand-built algorithms for precision in regulated domains.

Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br

- **AI Validation is Essential:**Reflection patterns showed how agentic flows (e.g., validator AI critiquing translator AI) boost reliability, but demand careful prompt engineering to avoid cascading errors.

- **Self-Hosted Scales Smartly:** Deploying NestJS pipelines reinforced prior infra shifts, emphasizing modularity for handling 100+ page docs without vendor lock-in.

- **LLM Integration Requires Strategy:** Switching from OpenAI to Groq using models like Mistral taught me that LLMs are not plug-and-play — optimization involves balancing performance, cost, and output quality, especially for domain-specific language like clinical data.

Business and Product Learnings

- **Waitlists Build Momentum:** Launching the LP early captured intent signals, underscoring tracking (pixels) as key to funnel optimization before full features

- **Compliance Starts Early:**LGPD/ISO dives clarified that healthtech credibility hinges on baked-in privacy, not bolt-ons—vital for B2B trust.

- **Mentorship Accelerates:** Alignments with Pedro Thompson provided not just advice but networks, reminding that solo efforts benefit from shared risks in entrepreneurship.

Inteli - Instituto de Tecnologia e Liderança
Avenida Professor Almeida Prado, 520, Butantã, São Paulo, SP.
CEP: 05508-070
www.inteli.edu.br