BÁO CÁO TÍNH NĂNG CHIẾN LƯỢC KỸ THUẬT

Phần Mềm Dịch Thuật Offline

CSV Translator Pro

Sử dụng mô hình AI TranslateGemma của Google

Hoạt động hoàn toàn offline – Hỗ trợ 55 ngôn ngữ

Kiến trúc Client-Server hiện đại

Mục đích:	Trình hội đồng phê duyệt kế hoạch thực hiện
Quy mô triển khai:	1000 người dùng
Mô hình AI:	TranslateGemma (4B / 27B tham số)
Công nghệ:	FastAPI + React + PyTorch + CUDA
Tổng ngân sách dự kiến:	789.800.000 VNĐ

Ngày lập: Ngày 8 tháng 2 năm 2026

Mục lục

1 Tóm tắt điều hành (Executive Summary)

1.1 Bối cảnh và nhu cầu

Trong bối cảnh hội nhập quốc tế, đơn vị thường xuyên phải xử lý khối lượng lớn văn bản đa ngôn ngữ, đặc biệt là **tiếng Ả Rập**. Việc dịch thủ công tốn nhiều thời gian và nguồn lực, trong khi các dịch vụ dịch thuật trực tuyến (Google Translate, DeepL) **không đáp ứng yêu cầu bảo mật** vì dữ liệu phải gửi qua Internet.

1.2 Giải pháp đề xuất

Triển khai hệ thống **CSV Translator Pro** — phần mềm dịch thuật AI **hoàn toàn offline**, đáp ứng:

- **Bảo mật tuyệt đối**: Dữ liệu không rời khỏi mạng nội bộ
- **Dịch hàng loạt**: Xử lý file CSV chứa hàng nghìn dòng văn bản
- **Chất lượng cao**: Sử dụng mô hình AI tiên tiến của Google
- **Đa ngôn ngữ**: Hỗ trợ 55 ngôn ngữ, đặc biệt tốt với tiếng Ả Rập

1.3 Điểm nhấn chiến lược

Tiêu chí	Giá trị
Tổng ngân sách	**789.800.000 VNĐ** (bao gồm phần cứng + triển khai + dự phòng 10%)
Thời gian triển khai	**6 tuần** (từ khảo sát đến nghiệm thu)
Số người dùng	**1000 người** (mở rộng được)
Mô hình AI	**TranslateGemma-27B** (27 tỷ tham số, chất lượng production)
GPU Server	**NVIDIA A100 80GB** (chạy full precision BF16)
Khả năng offline	**100%** (sau khi tải model lần đầu)

Khuyến nghị của đội dự án

Đề xuất hội đồng **phê duyệt kế hoạch** triển khai với cấu hình GPU NVIDIA A100 80GB để đảm bảo:

- Chất lượng dịch thuật tối ưu (mô hình 27B full precision)
- Hiệu năng cao phục vụ 1000 người dùng đồng thời
- Khả năng mở rộng trong tương lai
- Độ tin cậy enterprise với bảo hành dài hạn

2 Tính năng chính của hệ thống

2.1 Tổng quan 3 chức năng dịch thuật

1. DỊCH FILE CSV HÀNG LOẠT

Upload file CSV → Chọn cột cần dịch → Chọn ngôn ngữ nguồn/đích → Theo dõi tiến trình → Tải file kết quả

Ứng dụng: Dịch danh sách tàu thuyền, báo cáo tình báo, dữ liệu Thuraya...

2. DỊCH VĂN BẢN TRỰC TIẾP

Nhập văn bản → Chọn ngôn ngữ → Nhấn Dịch → Nhận kết quả ngay

Ứng dụng: Dịch tin nhắn, email, đoạn văn ngắn cần xử lý nhanh

3. DỊCH VĂN BẢN TỪ ẢNH (OCR)

Upload ảnh → AI nhận dạng chữ → Dịch tự động → Hiển thị kết quả

Ứng dụng: Dịch ảnh chụp tài liệu, biển báo, ảnh màn hình

Hình 1: Ba chức năng dịch thuật chính của CSV Translator Pro

2.2 Chi tiết tính năng dịch file CSV

Đây là tính năng **chiến lược** của hệ thống, cho phép xử lý hàng loạt dữ liệu:

Bảng 1: Quy trình dịch file CSV

Bước	Thao tác	Chi tiết kỹ thuật
1	Upload file CSV	Hỗ trợ file lớn, mã hóa UTF-8, tự động detect header
2	Chọn cột cần dịch	Mặc định cột "Text", có thể tùy chỉnh
3	Chọn ngôn ngữ	55 ngôn ngữ hỗ trợ (Ả Rập, Việt, Anh, Trung, Nhật...)
4	Bắt đầu dịch	Server xử lý bất đồng bộ, không block giao diện
5	Theo dõi tiến trình	Progress bar cập nhật mỗi giây, hiển thị % và số dòng
6	Tải kết quả	File CSV mới với cột "Translated_Text" thêm vào

Hiệu năng ước tính với GPU A100 80GB

Số dòng CSV	Thời gian ước tính	Tokens/giây
100 dòng	2–3 phút	40–70
500 dòng	8–12 phút	40–70
1,000 dòng	15–22 phút	40–70
5,000 dòng	1–2 giờ	40–70

2.3 Danh sách 55 ngôn ngữ hỗ trợ

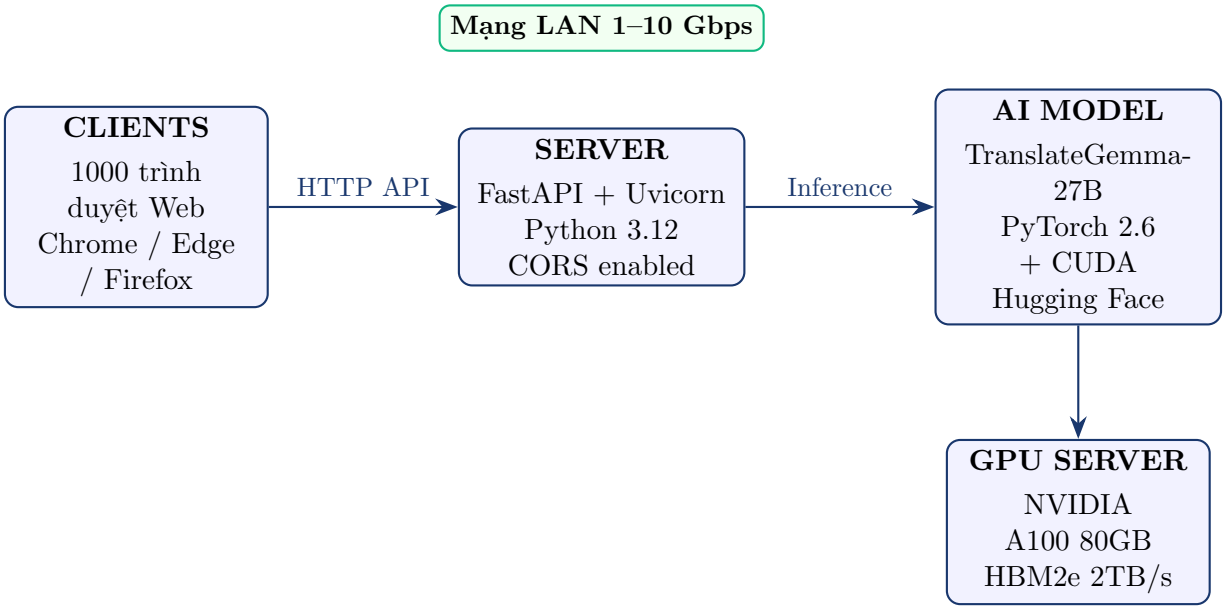
Mô hình TranslateGemma hỗ trợ dịch qua lại giữa **55 ngôn ngữ**, bao gồm:

Bảng 2: Các ngôn ngữ được cấu hình sẵn trong hệ thống

Mã	Ngôn ngữ	Mã	Ngôn ngữ	Mã	Ngôn ngữ
`ar`	Tiếng Ả Rập	`vi`	Tiếng Việt	`en`	Tiếng Anh
`zh`	Tiếng Trung	`ja`	Tiếng Nhật	`ko`	Tiếng Hàn
`ru`	Tiếng Nga	`de`	Tiếng Đức	`fr`	Tiếng Pháp
`es`	Tiếng Tây Ban Nha	`pt`	Tiếng Bồ Đào Nha	`it`	Tiếng Ý
`nl`	Tiếng Hà Lan	`pl`	Tiếng Ba Lan	`tr`	Tiếng Thổ Nhĩ Kỳ
`th`	Tiếng Thái	`cs`	Tiếng Séc	*... và 38 ngôn ngữ khác*	

3 Kiến trúc kỹ thuật hệ thống

3.1 Kiến trúc tổng quan Client-Server



Hình 2: Sơ đồ kiến trúc hệ thống CSV Translator Pro

3.2 Stack công nghệ

Bảng 3: Công nghệ sử dụng trong hệ thống

Thành phần	Công nghệ	Mô tả
Backend (Server)		
Framework web	FastAPI	Framework Python hiệu năng cao, async native
ASGI Server	Uvicorn	Xử lý requests đồng thời, production-ready
ML Framework	PyTorch 2.6	Deep learning framework của Meta
Model Loading	Hugging Face Transformers	Quản lý và chạy mô hình AI
Quantization	bitsandbytes	Lượng tử hóa 4-bit/8-bit giảm VRAM
Frontend (Client)		
UI Library	React 19.2	Component-based UI framework
Build Tool	Vite 7	Build tool siêu nhanh, Hot Module Replacement
AI Model		
Mô hình chính	TranslateGemma-27B-IT	27 tỷ tham số, Google research
CUDA	12.4	GPU acceleration cho NVIDIA

3.3 API Endpoints

Bảng 4: Danh sách API của hệ thống

Method	Endpoint	Chức năng
GET	`/api/languages`	Lấy danh sách 55 ngôn ngữ hỗ trợ
POST	`/api/translate-text`	Dịch một đoạn văn bản
POST	`/api/translate-image`	OCR + dịch văn bản từ ảnh
POST	`/api/upload`	Upload CSV và bắt đầu dịch hàng loạt
GET	`/api/status/{job_id}`	Kiểm tra tiến trình dịch CSV
GET	`/api/download/{job_id}`	Tải file CSV đã dịch
GET	`/api/health`	Kiểm tra trạng thái hoạt động server

4 Mô hình AI TranslateGemma

4.1 Giới thiệu mô hình

TranslateGemma là dòng mô hình dịch thuật mới nhất của Google, được xây dựng trên nền tảng kiến trúc **Gemma 3**. Đây là mô hình **mã nguồn mở** (Open Source) có thể chạy hoàn toàn offline sau khi tải về.

Bảng 5: Thông số mô hình TranslateGemma-27B-IT

Thông số	**Giá trị**
Tên mô hình	`google/translategemma-27b-it`
Số tham số	27 tỷ (27B)
Kiến trúc	Decoder-only Transformer (Gemma 3)
Số lớp (Layers)	46
Hidden Dimension	4096
Attention Heads	64 (Grouped-Query Attention)
Context Length	2,048 tokens
Số ngôn ngữ hỗ trợ	55 ngôn ngữ
Framework huấn luyện	JAX + ML Pathways (Google TPU)
Dữ liệu huấn luyện	4.3 tỷ tokens (SFT) + 10.2 triệu tokens (RLHF)
Giấy phép	Gemma License (mã nguồn mở)

4.2 So sánh chất lượng dịch thuật

Bảng 6: Benchmark so sánh các phiên bản TranslateGemma

Benchmark	**4B**	**12B**	**27B**	**Ghi chú**
WMT24++ MetricX ↓	5.32	3.60	**3.09**	Thấp hơn = tốt hơn
WMT24++ COMET ↑	81.6	83.5	**84.4**	Cao hơn = tốt hơn
Vistra MetricX ↓	2.57	2.08	**1.57**	Đánh giá đa ngôn ngữ

Lợi ích của mô hình 27B

Mô hình TranslateGemma-27B cho **chất lượng dịch thuật vượt trội**:

- Giảm 42% lỗi dịch so với phiên bản 4B (MetricX 5.32 → 3.09)
- Xử lý tốt các ngữ cảnh phức tạp, thuật ngữ chuyên ngành
- Đặc biệt hiệu quả với tiếng Ả Rập (ngôn ngữ khó)

- Hỗ trợ OCR từ ảnh với độ chính xác cao

4.3 Yêu cầu VRAM theo phương pháp lượng tử hóa

Bảng 7: Yêu cầu VRAM GPU cho mô hình 27B

Phương pháp	**VRAM Model**	**VRAM Thực tế**	**Ghi chú**
FP16 / BF16 (Full)	~54 GB	~58–62 GB	Chất lượng tốt nhất, cần A100/H100
INT8 (8-bit)	~27 GB	~30–34 GB	Giảm nhẹ chất lượng, cần A6000 48GB
NF4 (4-bit)	~14 GB	~18–22 GB	Giảm chất lượng, chạy được RTX 4090

5 Cấu hình phần cứng đề xuất

5.1 Cấu hình Server với GPU A100 80GB

Đây là cấu hình được **khuyến nghị** cho triển khai production với 1000 người dùng:

Bảng 8: Chi tiết cấu hình Server GPU NVIDIA A100 80GB

Thành phần	Thông số kỹ thuật	SL	Thành tiền (VNĐ)
GPU	NVIDIA A100 80GB PCIe (HBM2e, 2TB/s)	1	450.000.000
CPU	AMD EPYC 7313 (16C/32T, 3.0GHz)	1	55.000.000
Mainboard	Supermicro H12SSL-i (SP3 Socket)	1	25.000.000
RAM	DDR4-3200 ECC REG 128GB (8x16GB)	1 bộ	35.000.000
SSD	NVMe Gen4 2TB (Samsung PM9A3 Enterprise)	1	12.000.000
PSU	1200W 80+ Platinum Redundant	1	8.000.000
Case	4U Rackmount Server Chassis	1	10.000.000
Cooling	Hệ thống tản nhiệt Server	1	5.000.000
UPS	3000VA Online Double Conversion	1	25.000.000
	Tổng chi phí phần cứng		**625.000.000**

Tại sao chọn NVIDIA A100 80GB?

- **VRAM 80GB HBM2e**: Chạy mô hình 27B với Full Precision BF16 — không cần quantization, chất lượng dịch tối ưu
- **Tensor Cores thế hệ 3**: Tăng tốc inference AI lên 312 TFLOPS
- **Băng thông 2TB/s**: Xử lý batch lớn không bị bottleneck
- **ECC Memory**: Đảm bảo độ chính xác tính toán
- **Thiết kế Datacenter**: Hoạt động 24/7, bảo hành enterprise

5.2 Yêu cầu máy khách (Client)

Máy khách **không cần cấu hình cao** vì toàn bộ AI chạy trên Server:

Bảng 9: Yêu cầu phần cứng/phần mềm máy khách

Thành phần	**Tối thiểu**	**Khuyến nghị**
CPU	Intel Core i3 / AMD Ryzen 3	Intel Core i5 trở lên
RAM	4 GB	8 GB
Màn hình	1280×720	1920×1080 (Full HD)
Trình duyệt	Chrome / Edge / Firefox $\geq$ 90	Chrome mới nhất
Mạng	Kết nối LAN/WiFi đến Server	LAN Gigabit

6 Chi phí dự toán tổng hợp

6.1 Chi phí phần cứng

Bảng 10: Dự toán chi phí phần cứng Server

Hạng mục	Chi phí (VNĐ)
GPU NVIDIA A100 80GB PCIe	450.000.000
CPU AMD EPYC 7313	55.000.000
Mainboard Supermicro H12SSL-i	25.000.000
RAM 128GB DDR4 ECC	35.000.000
SSD NVMe 2TB Enterprise	12.000.000
PSU 1200W Platinum + Case + Cooling	23.000.000
UPS 3000VA Online	25.000.000
Tổng phần cứng	**625.000.000**

6.2 Chi phí triển khai và đào tạo

Bảng 11: Dự toán chi phí triển khai

Hạng mục	Chi phí (VNĐ)
Triển khai phần mềm (cài đặt, cấu hình, kiểm thử 2 tuần)	30.000.000
Đào tạo quản trị viên (2–3 người, 2 ngày)	5.000.000
Đào tạo người dùng (1000 người, 10 đợt)	50.000.000
Tài liệu hướng dẫn (biên soạn, in ấn)	3.000.000
Hỗ trợ kỹ thuật ban đầu (1 tháng sau nghiệm thu)	5.000.000
Tổng triển khai & đào tạo	**93.000.000**

6.3 Tổng hợp ngân sách

Bảng 12: Tổng hợp ngân sách dự án

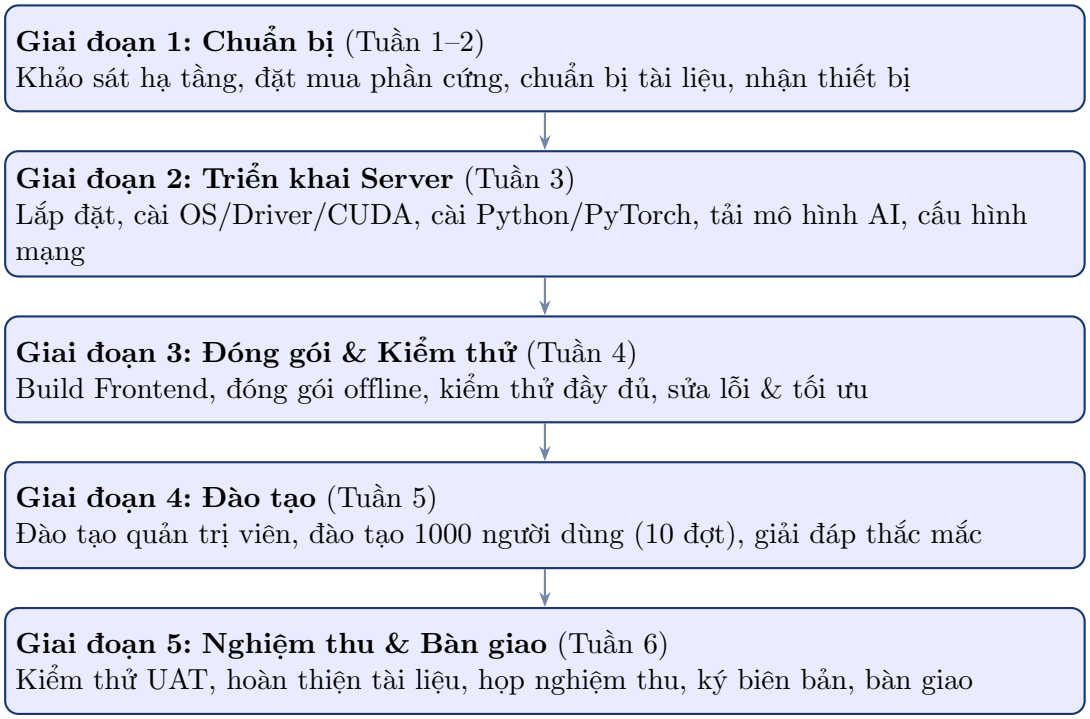
Hạng mục	Chi phí (VNĐ)
Phần cứng Server (A100 80GB + hệ thống)	625.000.000
Triển khai & Đào tạo	93.000.000
Dự phòng (10%)	71.800.000
TỔNG CỘNG	**789.800.000**

Phân tích chi phí theo đầu người sử dụng

- Tổng chi phí: **789.800.000 VNĐ** cho 1000 người sử dụng
- Chi phí trung bình: **789.800 VNĐ/người** (đầu tư 1 lần)
- Chi phí phần mềm: **Miễn phí** (mã nguồn mở Google)
- Chi phí vận hành: Chỉ điện năng + bảo trì thường xuyên
- So sánh: Dịch vụ dịch thuật trung bình **200.000 VNĐ/1000 từ**

7 Kế hoạch thời gian triển khai

7.1 Tổng quan 5 giai đoạn (6 tuần)



Hình 3: Sơ đồ 5 giai đoạn triển khai trong 6 tuần

7.2 Nhân sự triển khai

Bảng 13: Đội ngũ triển khai dự án

STT	Vai trò	Trách nhiệm	Số lượng
1	Quản lý dự án (PM)	Điều phối, báo cáo tiến độ, giao tiếp các bên	1
2	KTV Phần cứng	Lắp đặt Server, kiểm tra thiết bị	1
3	KTV Hệ thống	Cài đặt OS, Driver, CUDA	1
4	KTV Phần mềm	Cài đặt ứng dụng, cấu hình, debug	1–2
5	KTV Mạng	Cấu hình mạng, firewall, IP	1
6	QA/Tester	Kiểm thử chức năng, hiệu năng	1
7	Đào tạo viên	Đào tạo người dùng cuối	1–2
		Tổng cộng	**8–10 người**

8 Đánh giá rủi ro và phương án giảm thiểu

8.1 Ma trận rủi ro

Bảng 14: Đánh giá các rủi ro chính

ID	Rủi ro	Xác suất	Mức	Phương án giảm thiểu
R1	Chậm giao hàng GPU A100	TB	Cao	Đặt hàng sớm, có 2–3 nhà cung cấp dự phòng
R2	GPU lỗi hoặc không tương thích	Thấp	TB	Kiểm tra kỹ trước mua, yêu cầu bảo hành
R3	Mạng LAN không ổn định	Thấp	Thấp	Khảo sát và nâng cấp mạng trước triển khai
R4	Mất điện đột ngột	Thấp	TB	UPS 3000VA + yêu cầu generator dự phòng
R5	Người dùng khó tiếp cận	TB	Thấp	Đào tạo kỹ, tài liệu đơn giản, video hướng dẫn
R6	Bảo mật: Truy cập trái phép	Thấp	TB	Firewall, chỉ mở trong LAN, không public

8.2 Phương án dự phòng GPU

Nếu không mua được A100 80GB

Trong trường hợp GPU A100 80GB không khả dụng hoặc vượt ngân sách:

- **Phương án B**: RTX A6000 48GB + INT8 quantization (chi phí ∼300 triệu, chất lượng giảm 5–10%)
- **Phương án C**: RTX 4090 24GB + NF4 4-bit quantization (chi phí ∼100 triệu, chất lượng giảm 15–20%)
- **Phương án D**: Sử dụng mô hình TranslateGemma-4B thay vì 27B (chất lượng giảm 40%, nhưng chạy được trên GPU 8GB)

9 Bảo mật và tuân thủ

9.1 Cam kết bảo mật dữ liệu

Bảng 15: Các biện pháp bảo mật hệ thống

Biện pháp	Chi tiết
Hoạt động offline	Sau khi cài đặt, hệ thống hoạt động 100% offline, không cần Internet
Dữ liệu nội bộ	Tất cả dữ liệu dịch thuật xử lý trong mạng LAN, không gửi ra ngoài
Firewall	Chỉ mở port 8000 trong mạng nội bộ, không public ra Internet
Không lưu log dịch	Hệ thống không lưu trữ nội dung văn bản đã dịch (chỉ log kỹ thuật)
Xóa dữ liệu tạm	File CSV upload được xóa tự động sau 7 ngày
Không cần đăng nhập	Truy cập trong LAN (có thể thêm authentication nếu cần)

Điểm mạnh về bảo mật

So với dịch vụ dịch thuật online (Google Translate, DeepL):

- Dữ liệu KHÔNG rời khỏi mạng đơn vị
- Không phụ thuộc Internet — hoạt động trong mọi điều kiện
- Không lo ngại điều khoản sử dụng của bên thứ ba
- Toàn quyền kiểm soát hệ thống và dữ liệu

10 Kết luận và đề xuất

10.1 Tóm tắt dự án

Bảng 16: Tóm tắt thông tin dự án

Tiêu chí	Nội dung
Tên dự án	CSV Translator Pro – Phần mềm dịch thuật AI Offline
Mục tiêu	Dịch thuật đa ngôn ngữ offline, bảo mật tuyệt đối
Quy mô	1000 người dùng
Tổng ngân sách	**789.800.000 VNĐ**
Thời gian triển khai	6 tuần
Công nghệ	TranslateGemma-27B + FastAPI + React
GPU đề xuất	NVIDIA A100 80GB (Full Precision BF16)
Khả năng offline	100% sau khi triển khai

10.2 Đề xuất hội đồng

ĐỀ XUẤT PHÊ DUYỆT

Kính đề xuất Hội đồng **phê duyệt kế hoạch thực hiện** dự án CSV Translator Pro với các nội dung:

1. **Phê duyệt ngân sách**: 789.800.000 VNĐ (bao gồm dự phòng 10%)
2. **Phê duyệt cấu hình**: GPU NVIDIA A100 80GB + Server workstation
3. **Phê duyệt timeline**: 6 tuần triển khai
4. **Phê duyệt đội ngũ**: 8–10 nhân sự triển khai
5. **Chỉ đạo phối hợp**: Các phòng ban hỗ trợ nhân sự, mạng, điện...

10.3 Lợi ích khi triển khai

- **Tiết kiệm thời gian**: Dịch tự động hàng nghìn dòng CSV trong vài phút thay vì hàng ngày thủ công
- **Bảo mật tuyệt đối**: Dữ liệu không bao giờ rời khỏi mạng nội bộ
- **Chất lượng cao**: Mô hình AI 27 tỷ tham số của Google
- **Chi phí 1 lần**: Không phí hàng tháng như dịch vụ online

- **Hoạt động 24/7**: Không phụ thuộc Internet hay dịch vụ bên ngoài
- **Mở rộng**: Có thể nâng cấp phục vụ nhiều người dùng hơn

Kính trình Hội đồng xem xét và phê duyệt

Ngày lập: Ngày 8 tháng 2 năm 2026