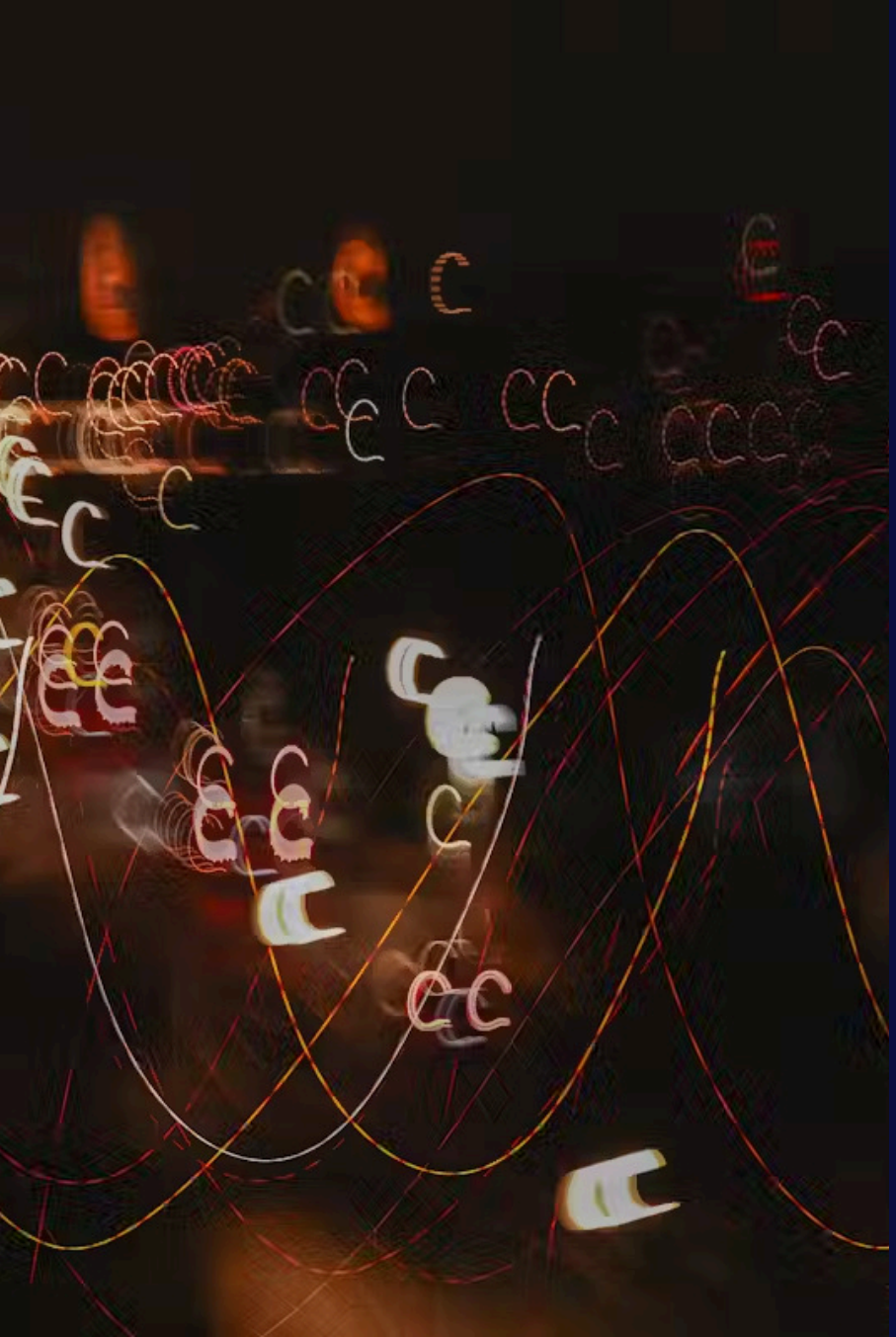
📄 Document Analysis Chatbot MVP Demo

This document outlines the features, technical innovations, and business value of our Document Analysis Chatbot MVP. Powered by Gemini 2.5 Flash, it offers intelligent document analysis, Vietnamese compliance checking, and significant cost optimization through implicit caching.

 **by Nguyen Hoang Nguyen (K18 HCM)**



Document Analysis Chatbot with Implicit Caching

Our intelligent document analysis system is powered by Gemini 2.5 Flash. It functions as a Vietnamese compliance checker for data protection laws, offering a cost-optimized AI chatbot with automatic caching. The system supports multi-format documents including PDF, DOC/DOCX, MD, and TXT.

The Problem We Solved

Business Pain Points

- Manual document review is time-consuming.
- Compliance checking requires legal expertise.
- High AI costs for document processing.
- Repetitive queries waste resources.
- Language barrier for Vietnamese regulations.

Our Solution

- Automated document analysis.
- Built-in Vietnamese legal knowledge.
- 50-90% cost reduction via smart caching.
- Legal compliance checking.

Key Features Demo

Interactive Chat

Ask questions, get detailed, contextual answers, compliant with data protection law.

Smart Summarization

Structured summaries, key point extraction, and technical term explanations.

Information Verification

Fact-checking against sources, confidence scoring, and evidence citations.

Legal Compliance Check

Vietnamese data protection law (Decree 13/2023), violation detection, risk assessment, and remediation.

Technical Innovation - Implicit Caching

Our implicit caching offers 50-90% cost reduction. Unlike traditional methods where every query incurs full cost, our system charges only for new content after the first query. This is achieved by Gemini 2.5 Flash automatically detecting repeated content when large documents and regulations are placed at the prompt's beginning, requiring no configuration.

Architecture Overview

```
# Smart Prefix Strategy
def create_common_prefix(document_content, data_protection_rules):
    prefix = f"""
    DOCUMENT CONTEXT: {document_content}
    LEGAL REGULATIONS: {data_protection_rules}
    You are a professional AI assistant...
    """
    return prefix

# Optimized Generation
def generate_with_implicit_cache(client, prefix, prompt):
    full_prompt = f"{prefix}\\n{prompt}"
    response = client.models.generate_content(full_prompt)
    # Automatic cache hit detection & cost tracking
    return response.text
```

The system design leverages a smart prefix strategy, combining document content and legal regulations into a common prefix. This optimized generation process, using Gemini 2.5 Flash, automatically detects cache hits, ensuring efficient cost tracking without manual configuration.

Live Demo Walkthrough

Document Upload

Upload sample documents, observe automatic text extraction and token count optimization.

Cache Optimization in Action

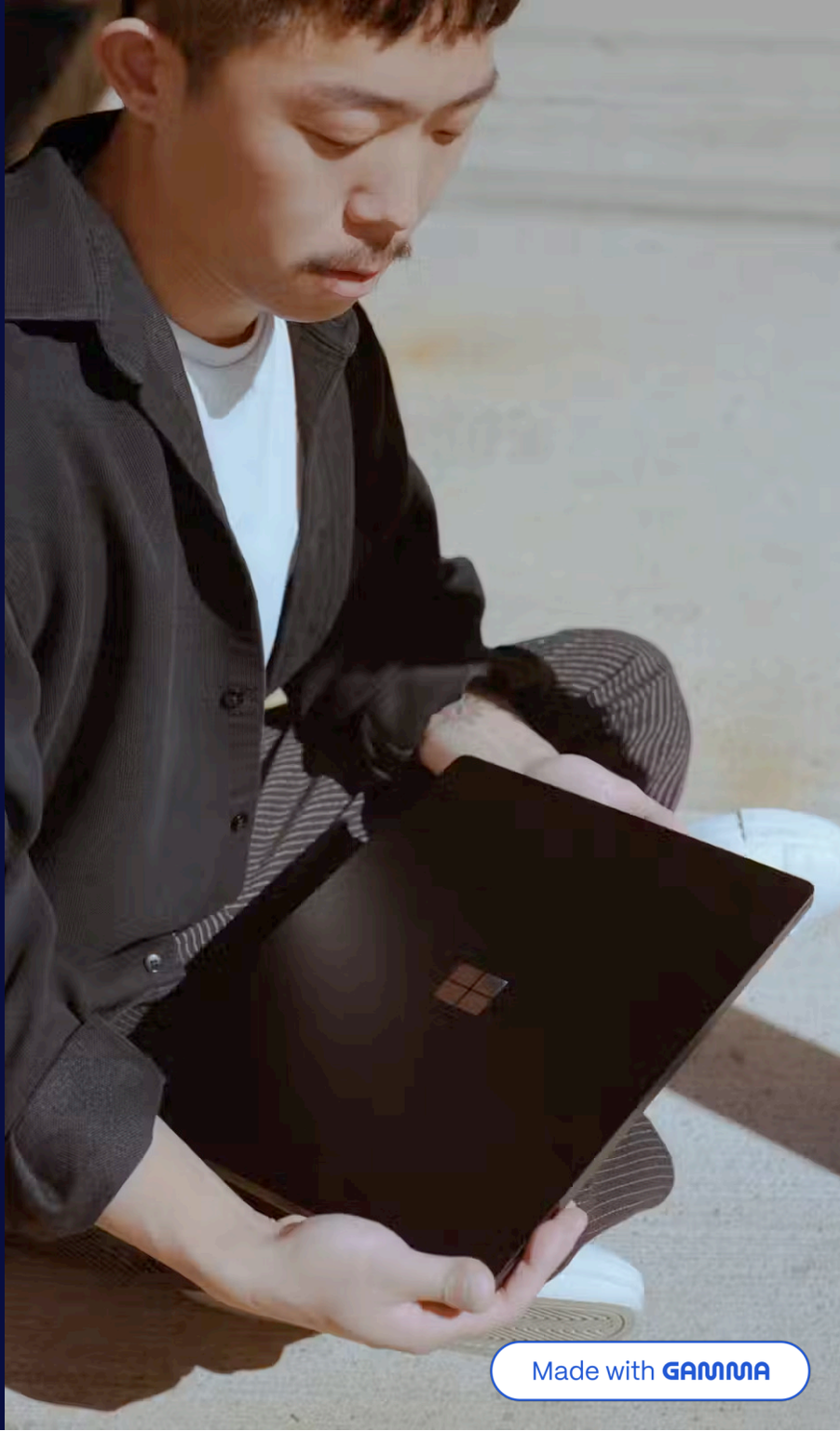
See full token usage on first query, cache hit notification, and real-time cost savings on subsequent queries.

Vietnamese Legal Analysis

Witness compliance checks against Decree 13/2023, detailed violation analysis, risk assessment, and recommendations.

Interactive Q&A

Engage in multi-turn conversations, asking complex document questions and receiving contextual answers with citations.



Results & Impact

Cost Efficiency

50-90% cost reduction via implicit caching, optimizing over 1024 tokens, and automatic scaling.

User Experience

Reduce response time, accurate answers with full document context, and Vietnamese language support.

Technical Achievement

Zero-configuration caching, multi-format support (PDF, DOC, TXT, MD), and scalable Streamlit architecture.

Business Value

Real-World Applications

- Legal Firms: Automated contract review, compliance, client Q&A.
- Enterprises: Policy analysis, regulatory compliance, employee queries.
- Government: Citizen service automation, document processing, multilingual support.

Cost Benefits

- Reduce legal review time by 70%.
- Lower AI processing costs by 80%.
- 24/7 availability vs. human experts.

Thank You for Listening

We appreciate your time today. Feel free to reach out with questions.

We look forward to discussing how our chatbot benefits your organization.