

SyncLLM™ : 중앙 AI 모델 게이트웨이 및 비용·보안 관리 플랫폼

사내에서 호출하는 모든 AI 요청을 한 곳에서 관리하여, 개인정보 유출을 막고 중복 질문 캐싱과 모델 자동 분배로 AI 사용료를 아껴주는 관문 플랫폼입니다.

EXECUTIVE SUMMARY · 도입 핵심 방향

복잡한 수작업과 부서 간 대기 시간을 줄이고, 현업 담당자가 쉽고 안전하게 업무를 처리할 수 있는 실무 중심 환경을 제안합니다.

01. 한눈에 보는 도입 기대 지표

Key Highlights

비용 절감

유사 질문 캐시 재사용
중복 토큰 과금 방지

자동 분기

난이도별 모델 라우팅
경량 모델 + 고성능 모델 조합

사전 차단

개인정보 자동 마스크
이름·연락처·계좌번호 보호

예산 한도

부서별 사용량 상한 관리
예산 초과 시 알림 및 제어

02. 도입 부서 주요 고민 및 SyncLLM™ 해결 방안 조건표

Requirement & Solution Mapping

업무 영역	현업에서 자주 겪는 고민	SyncLLM™ 해결 방법
AI 호출 비용 관리	전 직원이 상용 AI API를 쓰기 시작하면 매달 요금이 얼마나 나올지 예측하기 어려움	비슷한 질문은 저장된 답변(시맨틱 캐시)으로 바로 응답하고, 부서별 월 예산 한도를 설정
모델 선택 효율화	단순 요약이나 인사말 작성에도 비싼 최고급 모델만 호출해 비용이 낭비됨	일상적인 작업은 사내 경량 모델(vLLM)로, 복잡한 추론만 상위 모델로 자동 분배
개인정보·기밀 보호	직원이 고객 실명이나 전화번호, 내부 코드를 프롬프트에 그대로 입력해 유출될 우려	외부 전송 전에 주민번호·전화번호·이메일 등 민감정보를 자동으로 가림(마스킹) 처리
서비스 연속성 확보	특정 AI 서비스(OpenAI 등)에 일시적인 장애가 나면 사내 AI 기능이 모두 멈춤	주력 모델 응답이 지연되면 예비 모델(Claude, 사내 모델 등)로 자동 전환해 중단 방지

Quick Access

온라인 문서 및 체험 안내

기술 문서 포털(<https://doc.empasy.com/syncllm/>) 및 통합 포털(<https://sync.empasy.io>)에서 상세 정보를 확인하실 수 있습니다.

라이브 데모: <https://llm.empasy.io>

도입 배경 및 기존 업무 방식과의 비교

현재 업무에서 반복되는 불편함을 짚어보고, 도입 후 업무 흐름이 어떻게 간편해지는지 비교합니다.

Current Pain Points

01. 기존 업무 환경에서 자주 겪는 4가지 어려움

왜 기존 방식으로는 업무 속도와 품질을 동시에 높이기 어려울까요?

- 부서마다 각자 API 키를 발급받아 사용하다 보니, 전사적으로 AI 비용이 어디에 얼마나 쓰이는지 파악하기 어렵습니다.
- 비슷한 내용의 질문이나 안내 문구 생성을 매번 새로 호출하여 불필요한 토큰 비용이 반복해서 나갑니다.
- 현업 직원이 실수로 고객 개인정보나 미공개 계약 내용을 외부 AI에 입력해도 이를 사전에 걸러낼 장치가 없습니다.
- 출근 시간 등 특정 시간대에 요청이 몰리면 외부 API 호출 한도(Rate Limit)에 걸려 응답이 멈추는 일이 생깁니다.

02. 기존 처리 방식(AS-IS) vs SyncLLM™ 도입 후(TO-BE) 흐름 비교

Process Transformation

기존 수작업 중심 방식 (AS-IS)

반복 수작업 · 대기 발생

1 개별 API 직접 호출

시스템·부서마다 외부 AI에 직접 연결

2 민감정보 그대로 전송

고객명·연락처가 여과 없이 외부로 전달

3 중복 질문 반복 과금

같은 질문에도 매번 고가 모델 비용 지불

4 월말 요금 폭증 확인

사후 청구서를 받고서야 비용 초과 인지

SyncLLM™ 도입 후 (TO-BE)

간편 처리 · 안전 검증

1 SyncLLM 단일 관문 통과

모든 AI 요청이 중앙 게이트웨이로 인입

2 개인정보 자동 가림 처리

연락처·주민번호 등 민감정보 즉시 마스킹

3 캐시 확인 및 맞춤 모델 연결

유사 질문은 즉시 반환, 난이도별 모델 분배

4 실시간 비용·예산 대시보드

부서별 사용량 확인 및 한도 내 안전 운영

03. 핵심 추진 방향 요약

Key Direction

방향 01 · 쉬운 사용성

복잡한 명령어 없이 직관적 처리

개발 지식이 없는 현업 담당자도 친숙한 웹 화면과 일상적인 언어로 원하는 작업을 바로 실행할 수 있습니다.

방향 02 · 안전한 통제

사람이 확인하고 언제든지 복구

중요한 변경은 담당자가 눈으로 확인한 뒤 반영하며, 변경 직전 상태를 저장해 두어 실수가 생겨도 바로 되돌립니다.

방향 03 · 비용 효율화

반복 인건비 및 AI 요금 절약

수작업 공수를 줄이고, 캐시 재사용과 사내 경량 모델을 활용해 운영 비용 부담을 낮춥니다.

목표 시스템 구성도 (3-Layer 아키텍처)

사용자 화면층 · 애플리케이션 및 AI 엔진층 · 데이터 및 인프라층의 3개 계층으로 분리된 안정적인 구조입니다.

01. SyncLLM™ 3-Layer 통합 아키텍처 다이어그램

Target Architecture



02. 기존 사내 시스템과의 유연한 연동 구조

Seamless Integration

연동 구분	지원 방식 및 표준 규격	도입 고객사 이점
단독 도입 및 단계적 확장	필요한 제품 1종만 먼저 단독 도입 가능하며, 추후 타 Sync 제품군과 표준 MCP로 즉시 연결	초기 도입 부담 없이 가장 급한 업무부터 순차 적용 가능
사내 기간제(ERP/SSO/DB) 연동	표준 REST API(/api/v1/...), OAuth2/JWT 통합 인증, JDBC 표준 DB 연결 지원	기존 운영 중인 사내 시스템을 갈아엎지 않고 그대로 연결해 활용

핵심 작동 프로세스 및 4대 주요 기능

실제 업무가 어떤 순서로 처리되는지와 실무에서 바로 체감할 수 있는 4가지 핵심 기능을 안내합니다.

01. 4단계 업무 처리 프로세스 (Step-by-Step)

End-to-End Workflow

STEP 01

요청 접수 및 개인정보 자동 마스킹

소요시간: 0.01초 내

사내 시스템에서 AI 요청이 들어오면 주민번호, 전화번호, 계좌번호 등 민감한 정보를 먼저 찾아 안전하게 가립니다.

STEP 02

유사 질문 캐시 확인 (비용 0원 응답)

소요시간: 0.05초 내

이전에 들어왔던 비슷한 질문인지 벡터 유사도로 확인해, 이미 검증된 답변이 있으면 외부 AI를 부르지 않고 바로 돌려줍니다.

STEP 03

작업 난이도에 맞는 모델로 자동 연결

소요시간: 실시간

새로운 질문이라면 단순 요약·분류는 사내 경량 모델로 보내고, 복잡한 분석만 고성능 상용 모델로 연결해 비용을 아낍니다.

STEP 04

토큰 사용량 기록 및 예산 한도 관리

소요시간: 실시간

호출에 사용된 토큰 수와 비용을 부서별로 기록하며, 정해진 월 예산 한도에 가까워지면 관리자에게 미리 알려줍니다.

02. SyncLLM™ 4대 핵심 기능 상세

Core Capabilities

핵심 기능 01

의미 기반 유사 질문 캐시 (Semantic Cache)

토씨까지 똑같지 않아도 질문의 뜻이 비슷하면 저장된 답변을 재사용하여, 응답 속도를 높이고 중복 과금을 줄여줍니다.

핵심 기능 02

사내 GPU + 외부 API 하이브리드 분배

아모레 아이리스 등 대규모 환경에서 검증된 방식으로, 일상적인 생성 작업의 대부분은 사내 GPU(vLLM)에서 추가 과금 없이 처리합니다.

핵심 기능 03

외부 전송 전 개인정보 자동 가림 (PII Masking)

프롬프트 안에 고객 실명이나 연락처가 포함되어 있으면 외부 API로 나가기 전에 가명 기호로 바꾸어 개인정보 유출을 예방합니다.

핵심 기능 04

OpenAI 호환 표준 주소 및 무중단 우회

기존 프로그램의 접속 주소(Base URL)만 SyncLLM으로 바꾸면 바로 적용되며, 특정 모델 장애 시 예비 모델로 자연스럽게 넘어갑니다.

주요 화면 구성 및 기존 방식과의 비교

실무 담당자가 사용하는 주요 화면 구성과 기존 수작업·타사 도구 대비 달라지는 점을 정리했습니다.

01. 실무자용 주요 화면 및 메뉴 구성

Key Screens & Menus

화면 01

llm.empasy.io

전사 토큰 비용 및 절감액 대시보드

오늘 사용한 토큰 비용과 캐시 재사용으로 아낀 금액, 모델별 호출 비중을 한눈에 보여줍니다.

화면 02

/llm/routing

모델 라우팅 및 예비(Fallback) 설정

업무별로 기본 사용할 AI 모델과 장애 시 대신 응답할 예비 모델 순서를 손쉽게 지정합니다.

화면 03

/llm/security-audit

개인정보 마스킹 및 보안 감사 로그

어떤 요청에서 개인정보가 감지되어 가려졌는지 이력을 확인하고 보안 필터 규칙을 관리합니다.

화면 04

/llm/budget

부서·프로젝트별 예산 한도 관리

팀별 일간·월간 최대 사용 금액을 설정하고 80% 도달 시 알림을 받도록 관리합니다.

02. 기존 운영 방식 대비 SyncLLM™ 비교표

Before vs After Comparison

비교 항목	기존 수작업 · 일반 도구 환경	SyncLLM™ 도입 환경
AI 사용 요금 관리	호출하는 만큼 무제한 과금 (예산 통제 어려움)	시맨틱 캐시 재사용 + 사내 경량 모델 분배로 비용 절감
고객 개인정보 보호	담당자 주의에만 의존 (프롬프트 내 유출 위험)	게이트웨이에서 연락처·주민번호 등 자동 마스킹 후 전송
외부 AI 장애 대응	특정 AI 서버 장애 시 사내 AI 기능 전체 중단	응답 지연 감지 시 사내 모델 또는 타사 모델로 자동 전환
기존 시스템 연동	모델을 바꿀 때마다 각 프로그램 코드 수정 필요	OpenAI 호환 표준 규격으로 코드 수정 없이 중앙에서 모델 변경

보안 관리 대책 및 운영 비용(TCO) 절감 방안

기업 내부 데이터를 안전하게 보호하는 보안 체계와 AI-인프라 운영 비용을 합리적으로 줄이는 방안입니다.

SECURITY & COST OPTIMIZATION

사내 폐쇄망·전용 클라우드 설치로 데이터 유출 걱정을 덜고, 사내 모델·캐시 활용으로 고정비 안에서 경제적으로 운영합니다.

01. 엔터프라이즈 3대 보안 및 안전 관리 대책

Security & Governance

보안 대책 01

정규식 + 문맥 결합 개인정보 필터

주민등록번호, 휴대전화, 카드번호, 이메일, 사내 기밀 키워드를 전송 전에 감지하여 안전하게 처리합니다.

보안 대책 02

외부 AI 학습 데이터 활용 차단 및 사내 모델 우선

민감한 고객 데이터가 포함된 작업은 외부로 보내지 않고 사내 네트워크 안의 자체 GPU 모델로만 처리하도록 고정할 수 있습니다.

보안 대책 03

API 키 중앙 금고(Vault) 암호화 관리

개발자 PC나 소스코드에 상용 API 키를 넣지 않고, SyncLLM에 AES-256으로 암호화해 등록한 뒤 내부 권한 토큰만 배포합니다.

02. 운영 비용(TCO) 및 리소스 절감 비교

Cost Optimization Analysis

비용 항목	기존 방식 (수작업 · 외부 유료 API 전면 의존)	SyncLLM™ 최적 운영안 (자동화 · 캐시/사내GPU 결합)
반복 질의·안내 문구 비용	동일·유사 질문도 매번 외부 API 토큰 비용 발생	시맨틱 캐시 및 야간 사전 생성으로 추가 호출 비용 대폭 감소
단순 요약·분류 작업 비용	고비용 플러그인 모델 일괄 사용	일반 작업(80%+)은 사내 GPU 경량 모델로 처리해 고정비 내 운영
보안 심의 및 관리 인건비	서비스마다 개별 보안 필터 및 로깅 개발	단일 관문 적용으로 전사 AI 보안·감사 기준 일괄 충족

03. 고객사 환경에 맞춘 3가지 배포(설치) 옵션

Deployment Options

옵션 A · 사내 구축형

온프레미스 (폐쇄망)

외부 인터넷과 분리된 고객사 자체 서버 및 사내 GPU 환경에 독립 설치하여 보안 규정을 준수합니다.

옵션 B · 전용 클라우드

고객사 전용 VPC

AWS, Azure, NCP 등 고객사 전용 격리 네트워크 안에 컨테이너(Docker/K8s) 기반으로 신속하게 구축합니다.

옵션 C · 하이브리드

사내 DB + 클라우드 연동

민감한 고객 원장은 사내에 두고, 필요한 연산 기능만 암호화 통신으로 연결해 초기 구축 부담을 낮춥니다.

단계별 도입 추진 일정(WBS) 및 산출물 안내

사전 검증(PoC)으로 실효성을 먼저 확인한 뒤 본 시스템 구축과 현업 이관까지 안전하게 진행합니다.

01. 3단계 표준 도입 일정 및 주요 과업 (WBS)

Implementation Roadmap

추진 단계	소요 기간	주요 수행 과업 내용	제공 산출물
1단계: 사전 검증 (PoC)	1~2주차 (2주)	SyncLLM 게이트웨이 설치 · 주요 업무 프롬프트 캐시 적중률 및 개인정보 마스킹 테스트	비용 절감 시뮬레이션 결과서
2단계: 본 시스템 구축	3~6주차 (4주)	사내 GPU(vLLM) 및 외부 상용 API 연결 · 부서별 예산 한도 및 보안 필터 정책 적용	라우팅 정책 정의서 · 보안 점검표
3단계: 오픈 및 안정화	7~8주차 (2주)	전사 AI 서비스 엔드포인트 전환 · 캐시 유사도 임계값 미세 조정 및 운영자 교육	운영 매뉴얼 · 월간 정산 리포트 양식

02. 단계별 품질 점검 기준 (Quality Gate)

Quality Assurance

GATE 01 · 착수/설계 요구사항·아키텍처 점검 고객사 업무 환경과 연동 대상 API, 보안 요건을 사전 확인하고 설계서를 확정합니다.	GATE 02 · 실데이터 연동 실환경 동작 검증 가짜 더미 데이터가 아닌 실제 테스트 DB와 연동하여 기능 동작과 데이터 정합성을 확인합니다.	GATE 03 · 통합 테스트 성능·권한·복구 리허설 역할별 접근 권한, 예외 상황 시 스냅샷 복원, 동시 접속 응답 속도를 종합 점검합니다.	GATE 04 · 변화관리/이관 현업 실습 교육 및 오픈 운영자와 현업 담당자가 직접 화면을 조작해 보는 실습 교육 후 안정화 지원에 돌입합니다.
---	--	---	---

03. 전담 수행 조직 및 역할 분담

Project Organization

역할 구분	주요 담당 업무	고객사 협업 사항
PM / 솔루션 아키텍트	전체 일정·품질 관리, 사내 시스템 연동 아키텍처 및 보안 설계 총괄	도입 목표 범위 확정 및 주간 진척 점검
도메인·AI 엔지니어	SyncLLM™ 코어 구축, 업무 규칙 및 프롬프트·모델 최적화, 인터페이스 개발	연동 대상 시스템 API/DB 명세 공유
QA / 변화관리 전담	실환경 시나리오 검증, 운영자·현업 사용자 매뉴얼 작성 및 대면 실습 교육 진행	현업 시범 사용자(UAT) 참여 및 검수

도입 기대 효과(ROI), 주요 활용 사례 및 기술 지원

도입 전후의 실질적인 업무 개선 효과와 안정적인 운영을 뒷받침하는 3단계 기술 지원 체계입니다.

01. 도입 전·후 기대 효과 비교 (ROI)

Expected Benefits

측정 항목	도입 전 (기존 방식)	SyncLLM™ 도입 후	개선 효과
월간 외부 LLM API 과금	모든 요청을 상용 고가 API로 직접 호출	캐시 응답 + 사내 경량 모델 우선 분배	토론 비용 50~60% 절감
반복 질문 평균 응답 속도	평균 2~4초 (매번 전체 추론 수행)	0.05초 내외 (시맨틱 캐시 적중 시 즉시 반환)	사용자 체감 속도 향상
신규 AI 모델 교체 작업	각 시스템 소스코드 수정 및 재배포 (수일)	관리 콘솔에서 라우팅 설정 변경 (즉시)	유지보수 공수 단축

02. 추천 도입 대상 및 주요 활용 시나리오

Use Cases & References

활용 분야 01

전사 AI 도입을 준비하는 중견·대기업

여러 부서의 AI 활용을 장려하면서도 월별 예산 한도와 개인정보 보호 기준을 한 곳에서 안전하게 관리

활용 분야 02

대규모 고객 상담·영업 지원 시스템

출근 시간대 집중되는 일일 브리핑과 반복 상담 질문을 캐시와 사내 GPU로 분산해 비용과 속도 동시 확보

03. 3단계 전담 기술 지원 및 유지보수 체계 (SLA)

Maintenance & Support

LEVEL 1 · 헬프데스크

15분 내 초동 응답

온라인·유선 기술 지원 창구를 통해 사용법 문의 및 설정 변경을 빠르게 안내합니다.

LEVEL 2 · 전담 엔지니어

1시간 내 원인 분석·복원

연동 이상 발생 시 로그 분석과 스냅샷 원상 복원을 통해 업무 지연을 최소화합니다.

LEVEL 3 · 코어 연구소

정기 점검 및 기능 패치

분기별 성능·보안 정기 점검 리포트를 제공하고 코어 엔진 업데이트를 지원합니다.

CONTACT & LIVE DEMO

주식회사 엠파시 (Empasy Inc.) | 살아 있는 소프트웨어는 엠파시가 만듭니다

서울특별시 금천구 가산디지털1로 205-27 · 도입 문의: poh@empasy.com / 0507-1360-8169

공식 홈페이지: <https://empasy.io>

통합 데모 포털: <https://sync.empasy.io>

기술 문서 포털: <https://doc.empasy.com>