

개발자센터 파트너사를 위한 비교 · 도입 제안

라우터는 많습니다. 만들고 · 돌리고 · 파는 것까지 끝까지 안 꿈기는 라우터는?

하나의 OpenAI 호환 API · 단일 키로 100+ LLM을 자동 라우팅하고, 원화로 정산하며, 카페24 스토어로 유통합니다.

1 Key

100+ Models

OpenAI-Compatible

Auto Router

Provider Routing

Auto Fallback

AI 앱을 만드는 일은 '프론트'와 'LLM 운영' 둘 다입니다.



① 프론트 / 앱을 빨리 못 만든다

기획부터 화면·배포까지 리소스와 시간이 든다. PoC 한 번 돌리는 데도 며칠이 걸린다.



② LLM 운영은 '지옥'이다

멀티벤더 키·결제 분산, 갑작스런 장애·타임아웃, 예측 안 되는 달러 비용. 정산·세금계산서는 또 별개.

시장 지형

LLM 라우터는 이미 검증된 카테고리다.

글로벌은 종합 라우터로 표준화, 국내는 소수·대개 B2B.

그래서 진짜 차이는

라우터 자체가 아니라

운영 · 현지화 · 유통에서 갈린다.

둘을 따로 풀면 느리고 비쌉니다. 카페24는 **한 스택에서** 묶습니다.

한눈에

만들고 → 돌리고 → 판다, 한 스택에서.

① 만들고 베타

AI Space

자연어로 앱 · 프론트엔드를 빠르게 생성하는
AI 앱 제작 플랫폼.



② 돌리고 핵심

LLM Router

1키 · 100+ 모델 · 자동 라우팅 · 무중단 풀백
의 멀티모델 LLM 백엔드.



③ 판다

개발자센터 · 스토어

수많은 쇼핑몰에 앱을 유통하고 원화로 정산.
고객이 이미 여기 있다.



제작 · 운영 · 유통 · 정산을 카페24 한 스택에서. 본 제안의 주축은 LLM Router, AI Space는 결합 가치를 보여주는 보조 축(베타/예정)입니다.

개요

LLM Router란 — 1키로 통하는 LLM 게이트웨이

하나의 OpenAI 호환 API와 단일 키 `sk-cafe24-*` 로 약 100~125개 LLM을 호출하고 자동 라우팅하는 통합 게이트웨이.



- 1 Key
- 100+ Models
- OpenAI-Compatible
- Auto Router
- Provider Routing
- Auto Fallback

프레이밍

개발사가 라우터를 고를 때 봐야 할 7가지

 01 모델 폭 100+ · 15 패밀리 이상	 02 OpenAI 호환 락인 최소 · 드롭인	 03 자동 모델선택 오토라우팅	 04 장애 폴백 무중단 전환
 05 제어 깊이 화이트/블랙 · 우선순위 · Preset	 06 국내 정산 원화 · 세금계산서	 07 유통 / 지원 우리 고객이 어디 있나	이 7개 축이 다음 비교표의 기준입니 다. →

9개 축으로 정직하게 비교하면

비교 축	Cafe24 LLM Router	글로벌 종합 라우터	국내 타 라우터
모델 폭 (100+)	✓ 100~125 · 15+ 패밀리	✓ 다수	△ 제한적
OpenAI 호환 · 락인 최소	✓ 드롭인	✓	△
Auto Router (4전략 모델선택)	✓ cafe24/auto	△ 일부	—
Provider Routing (멀티경로)	✓ 화이트/블랙리스트	✓	△
장애 폴백 (무중단)	✓ on/off	✓	△
제어 깊이 (우선순위·Preset)	✓ 세밀	✓	△
국내 정산 (원화·세금계산서)	✓ 원화·세금계산서·카카오	—	✓
유통 / 지원 (생태계)	✓ 개발자센터·스토어	—	△
운영 부가 (캐시·BYOK·관측성·알림)	✓ 전부	△ 일부	—

※ 경쟁사 카테고리 라벨로만 표기 · 사실에 근거한 축만 비교 · 불명 항목은 "—". 글로벌 벤치마크 OpenRouter.

차별점 ① · 핵심

하나의 호출, 3개의 라우팅 엔진

cafe24/auto 한 줄이면 — 모델 선택 · 프로바이더 경로 · 장애 전환이 자동으로 맞물려 비용·성능·안정성을 동시에 잡습니다.

📍 엔진 1

Auto Router

어떤 **모델**이 최적인가

요청을 분석해 가장 효율적인 모델에 자동 분배. 요청 단위 오버라이드·화이트리스트 지원.

price latency throughput

balanced

🗺️ 엔진 2

Provider Routing

어떤 **프로바이더** 경로로

같은 모델이라도 다수 프로바이더 중 최적 경로로 라우팅. 화이트/블랙리스트로 경로를 직접 통제.

provider A → best path

🛡️ 엔진 3

Auto Fallback

멈추지 않는 **SLA**

장애·타임아웃 시 사전 정의된 대체 경로로 즉시 전환. 실패한 호출은 과금하지 않습니다(ZCI).

qwen3-72b → llama-3.3 →

deepseek

세 엔진이 한 번의 호출에서 맞물립니다 — **비용 최적화 · 성능 · 무중단**을 코드 변경 없이.

차별점 ②

한국 최적화 정산 — 국내 사업 운영, 그대로

 원화 종량제 매일 9시 최초고시환율 · 월 기본료 0원 · 약정 없음	 모델가 마진 0% 원천 공개가 그대로 + 충전 수수료 ~3% (+VAT)	 실패 호출 무과금 ZCI — 실패한 호출은 과금하지 않습니다
 세금계산서 · 환불 신용카드/계좌이체 · 자동충전 · 세금계산서(현금분)	 카카오톡 알림 잔액 부족 · 자동충전 실패 · 호출 에러 즉시 발송	 국내 지원 · 한국어 국내 정산·지원 체계, 한국어 문서와 응대

달러 · 해외카드 · 영문 지원이 아니라, **국내 사업 운영 그대로.**

프론트는 AI Space가, LLM 운영은 Router가

1 AI Space(베타)로 앱·프론트 생성
자연어로 화면을 빠르게

2 LLM Router에서 키 발급
sk-cafe24-* 단일 키

3 앱의 .env에 키·base_url 주입
드롭인, 코드 변경 최소화

4 cafe24/auto로 호출·폴백·한도
무중단 운영 시작

```
# .env (AI Space로 만든 앱에 주입)
LLM_ROUTER_API_KEY=sk-cafe24-*****
LLM_ROUTER_BASE_URL=https://llm-router.cafe24.com/v1

from openai import OpenAI; import os
client = OpenAI(
    api_key=os.environ["LLM_ROUTER_API_KEY"],
    base_url=os.environ["LLM_ROUTER_BASE_URL"])
resp = client.chat.completions.create(
    model="cafe24/auto", messages=[...])
```

둘 다 카페24 안에서 — 만들고, 돌리는 일을 한 번에.

OpenAI SDK 드롭인 — base_url·key만 교체

STEP 1

키 발급

sk-cafe24-*

STEP 2

base_url 교체

.../v1

STEP 3

호출

cafe24/auto

```
from openai import OpenAI
client = OpenAI(api_key="sk-cafe24-*****",
base_url="https://llm-router.cafe24.com/v1")
resp = client.chat.completions.create(
model="cafe24/auto",
messages=[{"role":"user",
"content":"이 상품 리뷰 100개를 핵심 3줄로 요약해줘"}],
)
print(resp.choices[0].message.content)
```

```
# cURL
curl .../v1/chat/completions \
-H "Authorization: Bearer sk-cafe24-*" \
-d '{"model":"cafe24/auto",
"stream":true}'
```

스트리밍(SSE) 지원

/v1/chat/completions · 직접 모델 지정도 가능: **claude-sonnet-4-6** 등

스토어 앱 시나리오

화면은 AI Space로, 모델은 Router로



상품상세 · 카피 생성
SEO 친화 설명·마케팅 카피 자동 작성



리뷰 요약 · 감성분석
수백 개 리뷰를 핵심 3줄로, 긍/부정 집계



CS 챗봇 · 자동응답
문의 분류·자동 답변, 취소 사유 분류



다국어 (글로벌몰)
상품·페이지 다국어 번역과 현지화



상품추천 · 검색 보강
의도 기반 추천, 시맨틱 검색 강화



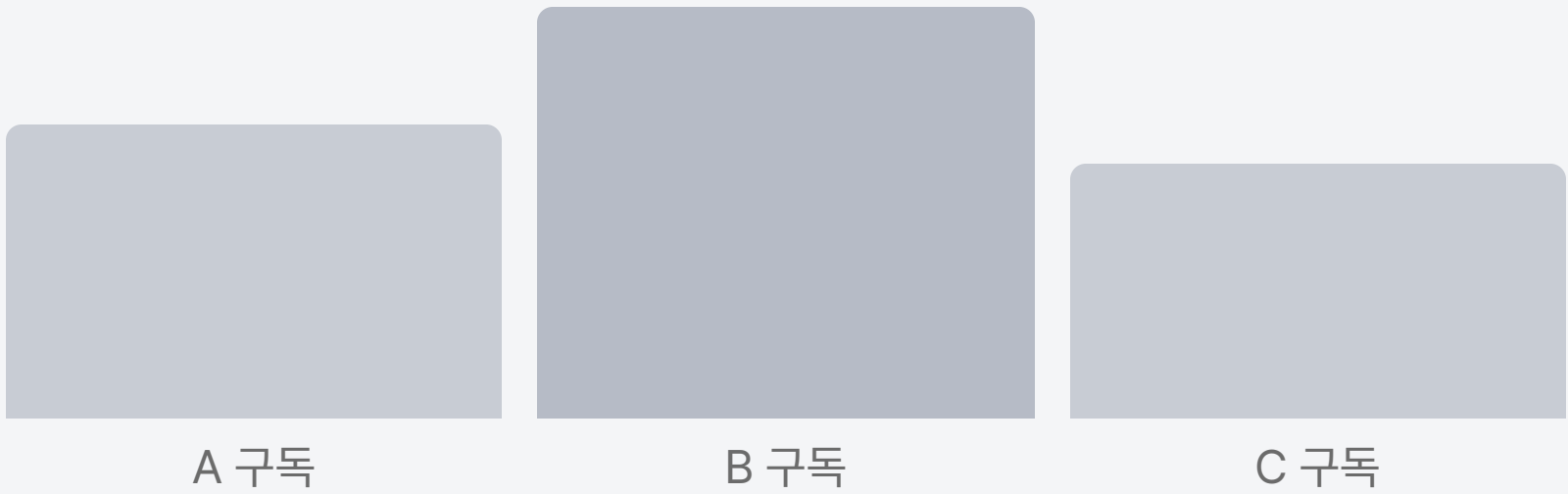
운영 리포트
주문·매출·문의 요약 리포트 자동 생성

✓ 각 카드: **AI Space로 화면, Router로 모델**. Auto Fallback으로 상점 실서비스 SLA를 보호합니다.

비용 감각

고정 구독 N개 vs 쓴 만큼 원화 1건

벤더별 구독 합산



고정비 다발 · 미사용분도 지불 · 달러 변동

※ 막대는 개념 비교 예시입니다. 실제 단가는 사용 모델·트래픽에 따라 달라집니다(임의 단가 아님).

쓴 만큼 원화 1건 청구

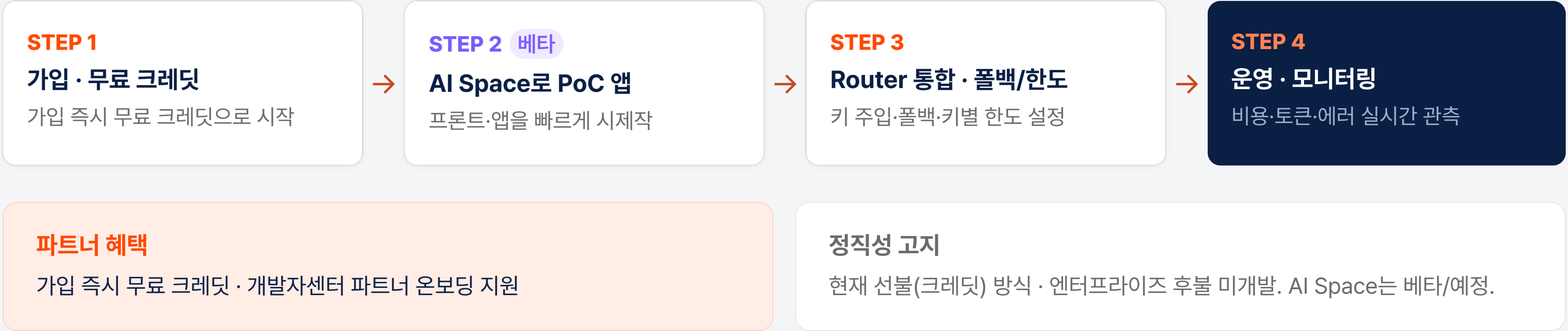


사용량 1건

초기 종량 유리 · 실패 무과금 · 작업별 저가 모델·캐시로 추가 절감

도입 절차 · 파트너 제안

4단계로 도입, 파트너로 함께



만들고, 돌리고, 파세요. 카페24에서.

비교는 끝났습니다 — 키 하나로 시작하세요.

무료 크레딧



AI Space로 앱 (베타)



5분 LLM 통합



폴백 켜기