

UNIFIED LLM GATEWAY

MonoRouter

하나의 API 키로 모든 주요 LLM을 호출하는 통합 게이트웨이

MonoRouter는 OpenAI·Anthropic·Google의 최신 언어 모델을 단일 엔드포인트와 단일 인증 키로 호출하는 API 게이트웨이입니다. 모델마다 다른 SDK와 계정, 결제 수단을 따로 관리할 필요 없이, 발급받은 키 하나로 모든 모델을 사용할 수 있습니다. 모든 호출은 크레딧 종량제로 정산되며, 키 단위로 사용량과 한도를 실시간으로 관리합니다.

연결 방식 | 기준 도메인 · <https://monogpt.kr>

Universal API

Bearer / x-api-key

</api/monorouter/v1>

OpenAI 호환 단일 엔드포인트. **model 값만 바꾸면**
OpenAI·Claude·Gemini를 동일한 형식으로 호출합니다.
Chat·Responses·Embeddings 지원.

ChatGPT API

Bearer

</api/monorouter/v1/openai>

OpenAI 규격을 그대로 사용하는 호환 경로. 기존 OpenAI SDK
의 base URL만 교체하면 됩니다.

Anthropic API

x-api-key

</api/monorouter/v1/anthropic>

Claude Messages 규격 호환. anthropic-version 헤더를 포함
해 Anthropic SDK와 동일하게 연동합니다.

Gemini API

x-goog-api-key

</api/monorouter/v1/gemini>

Google generateContent 규격 호환. 텍스트 생성과 임베딩
(embedContent)을 지원합니다.

핵심 기능

- **단일 키, 모든 모델** — Universal 키 하나로 4가지 연결 방식의 모든 모델을 호출합니다.
- **크레딧 종량제** — 사용한 만큼만 차감되며, 남은 크레딧은 이월됩니다.
- **키별 한도·사용량 관리** — 키마다 크레딧 한도와 사용 가능 모델을 지정하고, 사용 내역을 실시간으로 확인합니다.
- **즉시 비활성화** — 노출이 의심되는 키를 콘솔에서 바로 정지할 수 있습니다.
- **투명한 정산** — 요청·토큰·차감 크레딧을 건별로 기록해 비용을 명확히 추적합니다.

시작하기

- 1 **MonoGPT 로그인** 후 상단 메뉴에서 MonoRouter 콘솔로 이동합니다.
- 2 연결 방식을 선택하고 **크레딧 한도를 지정해 API 키를 발급**합니다.
- 3 발급된 키를 코드의 인증 헤더에 넣어 **곧바로 호출**합니다.
- 4 콘솔에서 키별 **사용량·잔여 크레딧**을 확인하고 관리합니다.

OpenRouter 키 발급 — 동일 콘솔에서 사용 한도를 지정해 OpenRouter 키(sk-or-v1)를 발급받아 openrouter.ai의 모델을 직접 호출할 수 있습니다. 발급한 키는 언제든지 다시 확인(보기)할 수 있습니다.

요금 안내

Biz-Pro 월 구독과 **Enterprise** 연간 라이선스로 제공합니다. 자세한 요금제와 도입 절차는 마지막 면을 참고해 주세요.

주식회사 모노플로우

대표 이기문 | 사업자등록번호 555-88-02700 | 통신판매업번호 2023-전주덕진-0858
본사 전북특별자치도 전주시 덕진구 오공로 33-6, 5층 금융혁신 공유오피스 9호실
연구소 서울특별시 서초구 반포대로22길 19-8 (서초동, 삼원빌딩) 3층

02-6956-8056

sales@monoflow.kr
소개 monogpt.kr/about
도입문의 monogpt.kr/contact

**QUICK START**

연동 샘플

아래 예시는 **curl** 기준입니다. 키는 인증 헤더의 **YOUR_KEY** 자리에 발급받은 값을 넣으세요. OpenAI·Anthropic·Google 공식 SDK를 사용하는 경우, **base URL만 위 연결 방식 경로로 교체**하면 기존 코드를 그대로 활용할 수 있습니다.

1 Universal API

/api/monorouter/v1

```
# 단일 엔드포인트 — model 값만 바꾸면 OpenAI · Claude · Gemini 모두 호출
curl https://monogpt.kr/api/monorouter/v1/chat/completions \
-H "Authorization: Bearer YOUR_KEY" \
-H "Content-Type: application/json" \
-d '{"model": "gpt-5.5", "messages": [{"role": "user", "content": "안녕하세요"}]}'
```

2 ChatGPT API

/api/monorouter/v1/openai

```
curl https://monogpt.kr/api/monorouter/v1/openai/v1/chat/completions \
-H "Authorization: Bearer YOUR_KEY" \
-H "Content-Type: application/json" \
-d '{"model": "gpt-5.5", "messages": [{"role": "user", "content": "안녕하세요"}]}'
```

3 Anthropic API

/api/monorouter/v1/anthropic

```
curl https://monogpt.kr/api/monorouter/v1/anthropic/v1/messages \
-H "x-api-key: YOUR_KEY" \
-H "anthropic-version: 2023-06-01" \
-H "Content-Type: application/json" \
-d '{"model": "claude-haiku-4.5", "max_tokens": 128, "messages": [{"role": "user", "content": "안녕하세요"}]}'
```

4 Gemini API

/api/monorouter/v1/gemini

```
curl "https://monogpt.kr/api/monorouter/v1/gemini/v1beta/models/gemini-3.5-flash:generateContent" \
-H "x-goog-api-key: YOUR_KEY" \
-H "Content-Type: application/json" \
-d '{"contents": [{"role": "user", "parts": [{"text": "안녕하세요"}]}]}'
```

5 OpenRouter

openrouter.ai · 별도 키(sk-or-v1)

```
# MonoGPT 콘솔에서 발급한 OpenRouter 키로 openrouter.ai를 직접 호출
curl https://openrouter.ai/api/v1/chat/completions \
-H "Authorization: Bearer YOUR_OPENROUTER_KEY" \
-H "Content-Type: application/json" \
-d '{"model": "openai/gpt-5.4-nano-chat", "messages": [{"role": "user", "content": "안녕하세요"}]}'
```



AVAILABLE MODELS

이용 가능 모델

MONOROUTER 기본 모델 Universal · ChatGPT · Anthropic · Gemini API

제공사	모델	모델 ID	비고
OPENAI	GPT-5.5	gpt-5.5	플래그십 범용
	GPT-5.4	gpt-5.4	고성능 범용
	GPT-5.4 Mini	gpt-5.4-mini · -mini-chat	경량·고속
	GPT-5.4 Nano	gpt-5.4-nano · -nano-chat	초경량
	GPT-5 Nano Chat	gpt-5-nano-chat	최저비용 대화
ANTHROPIC	Claude Opus 4.8	claude-opus-4-8	최상위 추론
	Claude Opus 4.7 · 4.6	claude-opus-4-7 · -4-6	고급 추론
	Claude Sonnet 4.6	claude-sonnet-4-6	성능·비용 균형
	Claude Haiku 4.5	claude-haiku-4.5	경량·고속
	Claude Fable 5	claude-fable-5	지원 예정
GOOGLE	Gemini 3.1 Pro	gemini-3.1-pro-preview	대용량 멀티모달
	Gemini 3.5 Flash	gemini-3.5-flash	고속 범용
	Gemini 3.5 Flash (Search)	gemini-3.5-flash-search	실시간 검색 결합
	Gemini 2.5 Flash Lite	gemini-2.5-flash-lite	초경량

임베딩 모델

text-embedding-3-large

text-embedding-3-small

text-embedding-ada-002

gemini-embedding-002

OPENROUTER 이용 가능 모델 별도 키(sk-or-v1) · openrouter.ai 경유

OPENAI	GPT-5.5	openai/gpt-5.5	ANTHROPIC	Claude Opus 4.5	anthropic/claude-opus-4.5
	GPT-5.4	openai/gpt-5.4		Claude Sonnet 4.5	anthropic/claude-sonnet-4.5
	GPT-5.4 Mini Chat	openai/gpt-5.4-mini-chat		Claude Haiku 4.5	anthropic/claude-haiku-4.5
	GPT-5.4 Nano Chat	openai/gpt-5.4-nano-chat	GOOGLE	Gemini 3.1 Pro	google/gemini-3.1-pro-preview
	GPT-5.2 Chat	openai/gpt-5.2-chat		Gemini 3.1 Flash Lite	google/gemini-3.1-flash-lite
	GPT-5.2 Pro Chat	openai/gpt-5.2-pro-chat		Gemini 2.5 Pro	google/gemini-2.5-pro
	GPT-4o	openai/gpt-4o		Gemini 2.5 Flash Lite	google/gemini-2.5-flash-lite
	GPT-4o Mini	openai/gpt-4o-mini	XAI	Grok 4	x-ai/grok-4
			PERPLEXITY	Sonar Pro	perplexity/sonar-pro

모델 목록은 지속적으로 업데이트되며, 일부 모델은 제공사 정책·지역에 따라 제공이 제한될 수 있습니다. Claude Fable 5는 미국 제재 해제 이후 제공될 예정입니다. 플랜·키 설정에 따라 계정별로 이용 가능한 모델이 달라질 수 있으며, 최신 목록과 모델별 단가는 MonoRouter 콘솔에서 확인하실 수 있습니다.

PRICING

요금 안내

도입 규모에 맞춰 월 구독(Biz-Pro)과 연간 라이선스(Enterprise)를 제공합니다. 아래 금액은 모두 부가가치세 (VAT)가 포함된 가격입니다.

Biz-Pro

팀·부서 단위 · 월 구독

550,000 원 / 월

VAT 포함

팀 단위로 MonoRouter API와 통합 AI 기능을 월 단위로 이용하는 구독형 요금제입니다.

Enterprise

시범 도입 · 연간

2,200,000 원 / 년

VAT 포함

기관 도입을 검토하는 단계에 적합한 1년 라이선스입니다. 본격 도입 전 시범 운영에 권장합니다.

기관 평균 · 가장 많이 선택

Enterprise

기관 표준 · 연간

22,000,000 원 / 년

VAT 포함 · 기관 평균 선택가

기관 전체 도입에 적합한 규모로, 도입 기관이 가장 많이 선택하는 구성입니다.

이용 기간 Enterprise 라이선스는 발급일로부터 **1년** 유효합니다.

연장 연장 구매가 가능하며, 연장 시 **잔여 크레딧의 유효기간이 연장 시점 기준 1년**으로 갱신됩니다.

정산 크레딧 종량제 — 사용한 만큼만 차감되고, 잔여 크레딧은 이월됩니다.

위 금액은 표준 구성 기준이며, 도입 규모와 모델 구성에 따라 견적이 달라질 수 있습니다. 자세한 사항은 도입문의로 안내해 드립니다.