BizRouter

ENTERPRISE LLM GATEWAY · PRODUCT BRIEF

AI 도입의 속도와 통제를, 동시에.

하나의 API와 한 장의 원화 청구서로 전 세계 주요 LLM을 안전하게 운영하는 국내 유일의 기업용 LLM 라우터. 요청마다 가장 적합한 모델을 고르는 스마트 라우팅이 비용을 최대 96% 낮추고, 모델 평가가 우리 업무에 맞는 모델을 30분에 찾아내고, Search가 실제 데이터를 찾아 답하며, CLI 한 단어로 코딩 에이전트까지 연결됩니다. 최상위 오픈웨이트 모델은 광주 국가 AI 데이터센터의 자체 GPU에서 국외 전송 없이 운영합니다.

```
route   bizrouter/route · 전략=비용 절감 · 실측 절감 93~96%
eval    우리 업무 사례로 후보 모델 비교 · 결과를 라우팅 정책으로
search  법령 · 공시 · 통계 · 웹 검색 55종 출처 · 공개 벤치 정답률 29→84%
serve   bizrouter/glm-5.3 · 광주 국가 AI 데이터센터 B300 · 요청 데이터 국외 전송 0건
200 OK · 확정 원가 KRW 회신 · 선택 근거 감사 로그 저장됨
```

01 — WHY BIZROUTER

모델은 계속 바뀝니다.
운영 체계는 흔들리면 안 됩니다.

부서마다 따로 발급한 프로바이더 키, 달러로 흩어진 청구서, 어디로 무엇이 나갔는지 알 수 없는 프롬프트. 기업의 AI 도입이 멈추는 지점은 모델 성능이 아니라 **운영과 통제**입니다. BizRouter는 전 세계 주요 LLM 앞에 단 하나의 관문을 세웁니다.

BEFORE — 직접 계약·직접 호출

- **프로바이더별 달러 결제** · 법인카드·환율·증빙이 팀마다 제각각
- **키 난립** · 누가 어떤 키로 얼마를 쓰는지 집계 불가
- **통제 부재** · 개인정보·금칙어가 그대로 외부 모델로
- **모델 교체 = 코드 수정 · 재배포**

AFTER — BIZROUTER 하나로

- **원화 통합 결제** · 청구서 한 장, 국내 B2B 정산(후불·세금계산서·거래 증빙)
- **조직 단위 키 관리** · 키·부서·프로젝트별 사용량과 한도
- **정책 게이트웨이** · 마스킹·금칙어·모델 통제를 요청마다 강제
- **모델 교체 = 설정 변경** · 재배포 없이 즉시 전환

12개
글로벌·국내 LLM 프로바이더를 하나의 API로

2줄
기존 OpenAI 호환 코드에서 바꿀 설정

100%
모든 요청·라우팅 결정에 감사 로그 기록

지원 프로바이더

OpenAI · Anthropic · Google Gemini · xAI Grok · OnPod · Perplexity · Upstage · Novita · Fireworks · z.AI · vLLM 자체 호스팅

모델 카탈로그는 실시간으로 갱신됩니다. 최신 목록은 bizrouter.ai/models에서 확인하세요.

```
# 기존 OpenAI 호환 코드에서 두 줄만 바꾸면 끝
client = OpenAI(
  base_url="https://api.bizrouter.ai/api/v1",
  api_key="sk-br-v1-...",
)
```

02 — SIX THINGS ONLY BIZROUTER HAS

게이트웨이는 많습니다.
이 여섯 가지는 BizRouter에만 있습니다.

키를 모아 주는 프록시는 어디에나 있습니다. BizRouter는 **모델을 고르고, 검증하고, 사실을 찾아 답하고, 개발자의 터미널까지 이어지는** 운영 체계이고, 최상위 오픈웨이트 모델을 국내 GPU에서 직접 운영합니다. 여섯 가지 차별점은 서로 맞물려 돕니다 — 평가 결과가 라우팅 정책이 되고, 라우팅이 Search와 CLI에 그대로 붙고, 모든 요청이 한 장의 원화 청구서와 감사 로그로 모입니다.

01 · SMART ROUTER p.04

93~96% ↓

요청마다 계산해서 고르는 모델

고정 규칙표가 아닙니다. 요청의 작업 유형·복잡도를 분석하고 후보의 비용·지연·안정성·적합도를 점수로 합산해 선택합니다. 응답을 바꾸지 않는 검증 모드부터 켤 수 있습니다.

02 · MODEL EVALUATION p.07

30분

우리 업무 사례로 모델 비교

공개 벤치마크가 아니라 우리 회사가 실제로 받는 질문으로 여러 모델을 한 번에 돌려 품질·비용·속도를 비교합니다. 결과가 그대로 스마트 라우팅 설정이 됩니다.

03 · SEARCH p.08

29% → 84%

웹을 넘어 한국 데이터까지

법령·공시·실거래·통계 등 55종 출처를 모델이 직접 조회합니다. 스마트 라우터의 공개 벤치마크 400문항 정답률이 29%에서 84%로. 데이터 조회 요금은 ₩0.

04 · CLI & AGENTS p.11

한 단어

bizrouter claude — 설정 파일 무변경

Claude Code·Codex·OpenCode·Hermes 명령 앞에 bizrouter 한 단어. 그 실행의 모델 호출과 과금만 BizRouter를 거치고, 에이전트에는 콘솔과 Search가 MCP 도구로 붙습니다.

05 · KRW BILLING p.12~14

청구서 1장

원화 정산과 요청 단위 통제

12개 프로바이더를 원화 청구서 한 장·세금계산서·후불로. 마스킹·금칙어·모델 통제·한도가 모든 요청에 강제되고, 라우팅 결정까지 감사 로그에 남습니다.

06 · K-OPENWEIGHT p.10

국외 전송 0건

최상위 오픈웨이트를 국내 GPU에서

GLM 5.3·DeepSeek V4.1 Flash·Qwen3.8의 가중치를 광주 국가 AI 데이터센터의 자체 B300에 올려 직접 운영합니다. 원 개발사 서버와 통신 0, 해외 API보다 빠르고 최대 1/24 비용.

여섯 가지가 하나의 흐름으로

01 **평가로 고른다** — 우리 사례로 후보 모델을 비교해 라우팅 정책에 반영

› 02 **라우터가 선택한다** — 요청마다 정책 안에서 가장 적합한 모델을 계산 · 국내 추론 모델만 허용할 수도

› 03 **Search가 찾아 답한다** — 사실이 필요한 질문은 55종 출처를 조회해 근거와 함께

› 04 **CLI로 개발자까지** — 코딩 에이전트도 같은 정책·같은 과금·같은 로그로

› 05 **한 장으로 정산한다** — 원화 청구서 1장 · 감사 로그 100% · 세금계산서

03 — SMART ROUTING ENGINE

모든 요청마다, 지금 가장 적합한 모델을 계산해서 선택합니다.

국내 첫 스마트 라우팅 엔진입니다. 고정 규칙표가 아니라, 요청 하나하나의 성격을 분석하고 후보 모델의 **비용·지연·안정성과 작업 적합도**를 점수로 합산해 호출 대상을 결정합니다. 코드에서 바꿀 것은 모델 이름 하나, `bizrouter/route`뿐입니다.

01
후보 구성
모델×프로바이더 조합과 최근 운영 지표, 키 상태를 모읍니다

›

02
적격성 게이트
정책·권한·키·요청 호환성을 먼저 검사해 부적격 후보를 제외합니다

›

03
요청 분석
작업 유형과 복잡도를 계산합니다. 분석은 모델을 직접 고르지 않습니다

›

04
점수 산정
운영 기준 점수에 요청 적합도 보정을 합산합니다

›

05
선택·호출
최저 점수 후보를 호출하고, 일시 오류면 다음 후보로 전환합니다

routing-decision.log

```
request   POST /v1/chat/completions · model=bizrouter/route
analyze   task=coding · complexity=high · 분석 결과는 점수 보정에만 사용
gate      후보 축소 — 정책 범위·모델 차단 목록·키 상태·요청 호환성 통과분만
score     비용·지연·안정성 penalty + 처리량·키 여유 bonus + 작업 적합도 보정
flip      Model A 선택 — 복잡한 coding 요청 적합도가 비용 차이를 상쇄
200 OK    x-bizrouter-routed-model 회신 · 선택 근거 감사 로그 저장됨
```

✔ 적격성이 점수보다 먼저

정책, 권한, 키 상태, 요청 호환성을 통과하지 못한 후보는 아무리 저렴하고 빨라도 점수 계산에 들어오지 못합니다. 가격이 보안을 이기는 일은 구조적으로 불가능합니다.

✔ 분석과 선택의 분리

요청 분석은 작업 유형과 복잡도만 판단하고, 최종 선택은 통제된 점수 계산이 수행합니다. 결과가 재현 가능하고 예측 가능합니다.

✔ 낮은 점수가 이깁니다

점수는 비용 함수입니다. 비용·지연·실패 위험은 penalty로, 처리량과 키 여유는 bonus로 반영해 후보를 정렬합니다.

✔ 모든 선택은 설명 가능

선택된 모델과 확정 원가는 API 응답에도 함께 담깁니다. "왜 이 모델이었는가"에 로그와 응답 양쪽으로 답할 수 있습니다.

04 — COST CONTROL BY ROUTING

모든 요청이 가장 비싼 모델로 갈 이유는 없습니다.

LLM 모델 간 토큰 단가는 수십 배까지 차이 납니다. 단순 요약과 복잡한 코드 생성이 같은 프리미엄 모델로 가는 순간 비용은 구조적으로 새기 시작합니다. 스마트 라우터는 **요청마다 필요한 만큼의 모델**을 골라 총비용을 관리합니다 — 품질 가드레일 안에서.

운영 기준 점수 L1 · 낮을수록 우선

-	예상 요청 비용	penalty
-	최근 응답 지연	penalty
-	최근 실패율	penalty
+	처리량	bonus
+	프로바이더 키 여유	bonus

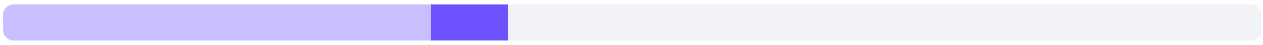
요청 적합도 보정 L2

±	작업 유형 적합도	coding·요약·번역…
±	복잡도 수용력	요청 난이도 대비
±	복잡도 정합	과소·과대 배정 방지

명백히 부적합한 모델은 싸고 빠르다는 이유만으로 선택되지 않습니다.

같은 두 후보, 요청이 순위를 뒤집습니다 (예시)

단순 요약 요청 → 저렴한 Model B 선택 — B가 1순위



복잡한 코드 생성 요청 → 적합도가 비용 차이를 상쇄, Model A 선택 — A가 1순위



운영 기준 점수(비용·지연·안정성) · 요청 적합도 보정 — 막대가 짧을수록(점수가 낮을수록) 우선

운영 전략 4종 — 무엇을 우선할지는 조직이 정합니다

네 가지 전략 모두 같은 신호를 사용하며, 달라지는 것은 가중치뿐입니다.

- **균형** · 비용·지연·안정성을 함께
- **비용 절감** · 예상 요청 비용 최우선
- **빠른 응답** · 지연·처리량 최우선
- **안정성 우선** · 성공률·키 여유 최우선

고객 상담은 빠른 응답, 야간 배치는 비용 절감 — 서비스별 Route로 분리 운영할 수 있습니다.

장애 시에도 원칙은 무너지지 않습니다

일시 오류가 나면 다음 순위 후보로 자동 전환하되, 그 후보도 동일한 적격성 검사를 다시 통과해야 합니다. 적격 후보가 없으면 임의 우회 없이 제외 사유와 함께 명시적으로 알립니다.

도입 경로 — 프로덕션 트래픽으로 검증한 뒤에 켭니다

STEP 1 · FIXED
고정 운영
정책의 기본 모델로만 응답합니다. 자동 선택은 꺼진 상태로 시작합니다.

STEP 2 · AUTO + PREVIEW
리스크 0 검증
응답은 기존 모델 그대로, 자동 선택 결과는 로그로만 기록해 실제 트래픽으로 검증합니다.

STEP 3 · AUTO
자동 라우팅
요청별 1순위 모델·프로바이더를 실제 호출에 적용합니다.

05 — PROVEN IN PRODUCTION

최신 플래그십과 완성도는 거의 그대로, 비용만 최대 96% 낮게.

한국어 실무 과제 57개를 Claude Fable 5 · GPT-5.6 Sol · Claude Opus 4.8 고정과 스마트 라우팅으로 똑같이 돌렸습니다. 스마트 라우팅은 정답이 명확한 과제에서 **객관 체크 31/32**, 품질은 최고 플래그십의 **94%**를 유지하면서 총비용은 플래그십 고정 대비 93~96% 낮았습니다.

93~96%↓
총비용 절감 (플래그십 고정 대비)

94%
품질 유지율 (최고 플래그십 대비 · 3사 블라인드 심사)

31 / 32
객관 체크 통과 (유닛테스트 · JSON · 정답 일치)



어떻게 쟀나

코딩·문서·분석·번역 등 한국어 실무 과제 57개. 같은 프롬프트를 플래그십 고정과 스마트 라우팅에 각각 보내 결과·비용을 기록했습니다. 게이트웨이 기본 설정, 2026년 7월 실측.

누가 채점했나

정답이 있는 과제는 유닛테스트·JSON 스키마·정답 일치로 기계 채점. 나머지는 서로 다른 3사 모델이 답의 출처를 모른 채 블라인드로 심사한 점수를 평균했습니다.

왜 싸지나

과제의 절반 이상은 플래그십이 필요 없습니다. 스마트 라우터가 요청마다 필요한 만큼의 모델을 고르고, 어려운 요청에만 비싼 모델을 쓰기 때문입니다.

결과는 정책·모델 풀·단가에 따라 달라집니다. 그래서 스마트 라우팅은 **응답을 바꾸지 않는 검증 모드**로 먼저 켜고, 실제 트래픽에서 절감폭을 확인한 뒤에 자동 라우팅으로 넘어가도록 설계했습니다. 상세 실측 근거는 bizrouter.ai/smart-routing.

06 — MODEL EVALUATION

어떤 모델이 우리 일에 맞는지, 30분이면 알 수 있습니다.

공개 벤치마크 점수는 우리 업무의 점수가 아닙니다. 모델 평가는 **우리 회사가 실제로 받는 질문**을 넣고 여러 모델에 똑같이 시켜 품질·비용·속도를 한 표로 비교합니다. 코드 한 줄 없이 콘솔에서 시작하고, 결과는 그대로 스마트 라우팅 정책으로 옮겨집니다 — 평가가 곧 운영 설정이 됩니다.

01
실제 사례 준비
호출 기록·CSV·직접 입력으로 우리 회사가 실제로 쓰는 질문을 모읍니다

›

02
모든 모델에 같은 일
후보 모델 전부에게 똑같은 사례를 보내 답변을 받습니다

›

03
같은 기준으로 채점
평가 모델이 정해진 기준표대로 모든 답변에 점수를 매깁니다

›

04
좋은 모델을 골라 적용
품질·비용·속도 결과를 보고 스마트 라우팅 설정에 반영합니다

30분
사례 준비부터 결과표까지 · 콘솔에서 클릭으로

8종
업무 템플릿 · 고르면 채점 기준이 미리 채워짐

3가지
사례 입력 경로 · 호출 기록 재사용 · CSV · 직접 입력

사례는 이미 있습니다

BizRouter를 거친 최근 30일의 실제 호출 기록을 평가 세트로 바로 재사용합니다. 새 사례를 정해진 범위·시간·건수로 수집하거나, CSV를 올리거나, 직접 적어도 됩니다. 모든 후보에게 똑같은 사례가 가야 공정한 비교가 됩니다.

채점은 다른 제조사 모델이

업무별 채점 기준표가 미리 채워지고, 평가 모델이 그 기준대로 모든 답변에 점수를 매깁니다. 후보와 다른 제조사의 평가 모델을 권장해 자기 답에 후한 점수를 주는 일을 막습니다. 후보에는 지금 쓰는 모델·저단가 모델·최신 모델을 섞어 넣는 것을 권합니다.

업무 템플릿 8종 — 무엇을 보는가

고객 문의 응대 · 정확성·지시 준수·말투
캐릭터 대화 · 일관성·맥락·몰입감
요약 · 누락·왜곡·분량
문서 기반 답변(RAG) · 근거 안에서만 답하는지
번역 · 뜻 보존·자연스러움
분류·태깅 · 정해진 값 정확히 반환
정보 추출 · 필드 정확도·JSON 형식
코드 · 요구사항 충족·실행 가능

후보	품질	비용	속도	판정
지금 쓰는 모델	9.1	100%	3.2초	기준선
후보 A (중급)	9.0	22%	1.4초	품질 동급 · 비용 1/5 → 라우팅 정책의 기본 후보로
후보 B (경량)	8.1	9%	0.9초	분류·추출처럼 정답이 명확한 업무에만

위 표는 이해를 돕기 위한 예시입니다. 실제 결과표는 사례별 답변·점수·비용·응답 시간을 모두 담고, 콘솔에서 바로 스마트 라우팅 정책으로 옮길 수 있습니다.
bizrouter.ai/evals

07 — BIZROUTER SEARCH

웹 검색을 넘어,
실제 데이터까지 찾아 답합니다.

「근로기준법 60조 원문」「삼성전자 최근 매출」「인천공항 지연편」 — 모델은 기억으로 답하거나 틀린 값을 자신 있게 말합니다. Search는 **모델에게 묻는 대신 실제 데이터를 찾아 답하는** API입니다. 질문을 이해하고, 필요한 출처를 고르고, 실제 값을 조회하고, 여러 출처를 교차 확인한 뒤 출처와 함께 답합니다. 웹 검색 API가 아닙니다.

01 **질문 이해**
무엇을 찾아야 하는지, 데이터가 필요한 질문인지 먼저 판단합니다

› 02 **출처 선택**
55종 출처 중 이 질문에 맞는 것만 고릅니다

› 03 **실제 조회**
모델이 출처를 도구로 호출해 지금의 값을 가져옵니다

› 04 **교차 확인**
여러 출처가 있으면 맞춰 보고, 못 찾은 것은 밝힙니다

› 05 **출처와 답**
답과 함께 어떤 출처를 몇 번 조회했는지 돌려줍니다

```
POST /v1/search

request   { "query": "근로기준법 60조 연차 유급휴가 조문" }
select    law(국가법령정보) 선택 · 다른 54종은 이 질문과 무관해 제외
lookup    law · get_law_article · 148ms · ok
200 OK    조문 원문 + 출처 · bizrouter_search { sources, calls, rounds, routed_model }
```

지금 쓰는 모델에 그대로 붙습니다

전용 엔드포인트 `POST /v1/search`로 부르거나, 기존 Chat Completions 코드에서 모델 이름 뒤에 `:search`만 붙이면 됩니다. 어떤 채팅 모델이든 되고, 스마트 라우터와도 결합됩니다. 데이터가 필요 없는 질문은 평소처럼 답합니다.

조회는 무료, 통제는 그대로

데이터 조회에는 별도 요금이 없고 모델 토큰만 과금됩니다. Search 요청도 다른 요청과 똑같이 마스킹·금칙어·한도·감사 로그를 거칩니다. 출처가 느리거나 죽어도 답이 멈추지 않도록 시간 예산과 자동 재시도가 걸려 있습니다.

들어 있는 출처 — 55종, 계속 늘어납니다

법령 · 판례 · 국회 의안 | 기업 공시 · 재무제표 | 사업자 등록 확인 | 부동산 실거래 · 건축물대장 | 국가통계 · 한국은행 금리 | 항공편 · 대중교통 · 지도 | 논문 · 임상시험 | 보건 · 의약품 | 웹 검색 · 뉴스 | 날씨 · 대기질 | 미국 주식 · 환율 | 공공조달 (나라장터) | 에너지 · 소상공인 · 창업 | 유튜브 · X 트렌드

출처는 자체 데이터 플랫폼 agent.store에서 운영합니다. 새 출처가 오르면 코드 변경 없이 Search에 편입됩니다. 최신 목록은 bizrouter.ai/docs/search · 소개와 실측은 bizrouter.ai/search

08 — SEARCH, PROVEN ON PUBLIC BENCHMARKS

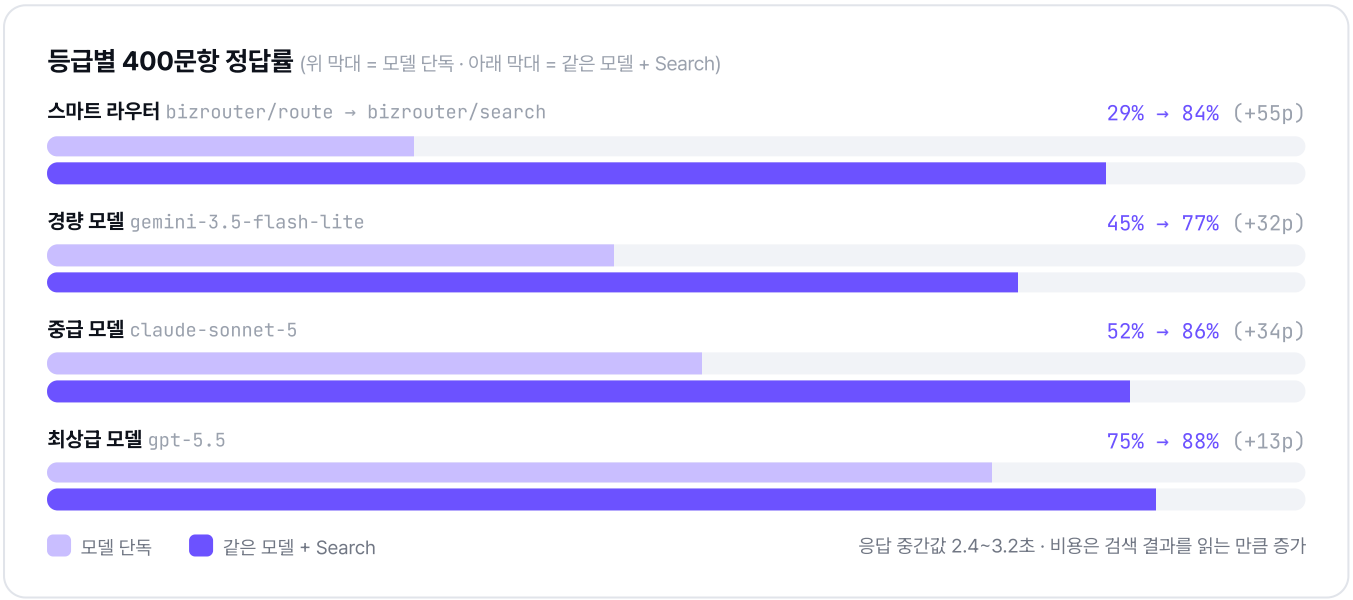
같은 모델, 같은 질문, 같은 채점.
달라진 것은 :search 한 단어.

SimpleQA·KoSimpleQA·FreshQA·FRAMES 각 100문항을 경량·중급·최상급 모델과 스마트 라우터에 같은 순서로 냈습니다. 모델 단독과 Search를 켠 것의 차이만 재도록 프롬프트·채점 모델·기준을 고정했고, 빈 답은 오답으로 셌습니다. 네 등급 모두 올랐고, **스마트 라우터는 정답률 29%에서 84%**로 — 최상급 모델 단독(75%)을 문항당 ~~₩6.2~~에 넘었습니다.

29% → 84%
스마트 라우터 400문항 정답률 · 정답 2.9배 · 오답 77% 감소

50% → 84%
네 등급 평균 정답률 · 경량·중급·최상급·라우터 모두 상승

~~₩6.2~~
라우터+Search 문항당 비용 · 최상급 단독은 ~~₩23.7~~에 75%



SimpleQA
짧은 사실 문답(영어). 답이 하나로 정해진 질문 — 기억으로 답하면 가장 자주 틀립니다. OpenAI 공개.

KoSimpleQA
한국 역사·문화·제도·인물 사실 문답. 해외 모델이 특히 약한, 한국 기업에 가장 현실적인 시험. NAVER 공개.

FreshQA
「지금 ~는?」처럼 답이 계속 바뀌는 질문과 잘못된 전제의 질문. 학습 시점 이후를 모르는 한계를 드러냅니다. Google 공개.

FRAMES
여러 문서를 찾아 조합해야 풀리는 어려운 질문. 검색 한 번으로는 안 되고 이어서 추론해야 합니다. Google 공개.

정답률 (모델 단독 → + Search)	SimpleQA	KoSimpleQA	FreshQA	FRAMES
스마트 라우터	12 → **97**	45 → **89**	42 → **77**	16 → **75**
경량 gemini-3.5-flash-lite	23 → **92**	60 → **87**	61 → **73**	37 → **57**
중급 claude-sonnet-5	22 → **97**	55 → **91**	70 → **77**	60 → **80**
최상급 gpt-5.5	61 → **97**	87 → **93**	72 → **79**	79 → **82**

2026년 9월 16일 실측 · 각 벤치 100문항 고정 추출 · OpenAI SimpleQA 채점 방식을 별도 채점 모델(gpt-5.4)로 적용 · 응답 시간은 검색이 한 번 더 들어가 몇 초 늘고, 최상급 모델은 자료를 종합하는 추론 단계가 더 들어가 비용 증가폭이 큽니다. 결과는 모델 풀·정책·시점에 따라 달라질 수 있습니다. 티어·벤치별 응답 시간과 비용까지 담은 상세 리포트: bizrouter.ai/search

09 — K-OPENWEIGHT · DOMESTIC INFERENCE

최상위 오픈웨이트 모델을, 대한민국에서 운영합니다.

GLM 5.3·DeepSeek V4.1 Flash·Qwen3.8 27B 같은 최상위 오픈웨이트 모델의 **가중치를 받아 광주 국가 AI 데이터센터에 둔 BizRouter 자체 B300 GPU에서 직접 돌립니다.** 요청도, 답도, 로그도 국경을 넘지 않고, 원 개발사 서버와는 어떤 통신도 없습니다. 「어느 나라 모델인가」 때문에 검토조차 못 했던 조직을 위해, 왜 안전한지를 파일 형식·방화벽·해시·감사 로그 수준에서 누구나 직접 확인할 수 있게 적었습니다.

0건
요청 데이터 국외 전송 · 게이트웨이(서울) → 마스킹(인천) → 추론(광주) → 로그(서울)

2.23s → 0.43s
GLM 5.3 응답 시간 · 개발사 공식 API 대비 국내 추론 · 같은 프롬프트 교차 10회 중앙값

최대 1/24
코딩 작업 비용 · Claude Opus 5 종량제 대비 GLM 5.3 Flash · 카탈로그 정가 기준

01 **요청**
고객사 · 국내 어디든 · OpenAI 호환 형식 그대로

› 02 **게이트웨이 · 서울**
AWS 서울 리전 · 인증 · 허용 목록 · 금칙어

› 03 **개인정보 처리 · 인천**
자체 H100 · 마스킹을 켠 조직만 거칩니다

› 04 **추론 · 광주**
국가 AI 데이터센터 · 자체 B300 · 가중치는 이 디스크에

› 05 **응답 · 기록 · 서울**
사용량·감사 로그도 서울 · 국외 구간 없음

기술적으로 왜 안전한가

- ✔ 가중치는 `safetensors` — 숫자 배열만 담는 형식이라 파일이 코드를 실행하는 단계가 없습니다.
- ✔ 추론 노드는 공인 IP가 없고 들어오는 연결 0, 나가는 길은 NAT 한 곳 — 원 개발사 서버로 연결을 열 프로그램 자체가 없습니다.
- ✔ 내려받은 파일은 개발사가 공개한 SHA-256과 대조합니다. 한 바이트라도 다르면 폐기.
- ✔ 원 개발사는 요청·응답·로그 어느 것도 볼 수 없습니다. 볼 수 있는 쪽은 고객사와, 조직 설정에 따라 BizRouter 감사 로그뿐입니다.

해외 API와 무엇이 다른가

개발사 API를 직접 부르면 요청이 그 회사의 해외 서버로 가고, OpenAI 최대 30일·Anthropic 30일·Gemini API 최대 55일 남용 감시용으로 보관되며, 미국 CLOUD Act로 저장 위치와 무관하게 미국 정부가 요구할 수 있습니다. 개인정보 국외 이전 고지·동의도 따라붙습니다.

K-OpenWeight는 추론 노드에 아무것도 남지 않고, 감사 로그(서울)는 조직이 켜고 끕니다. 국외 이전이 없으니 절차 자체가 없고, 같은 모델을 개발사 공식 API로 부를 때보다 응답도 빠릅니다(DeepSeek V4.1 Flash 0.83s → 0.47s).

지금 국내 추론으로 제공하는 모델 — 이름 앞의 BIZROUTER/ 가 표시입니다

- GLM 5.3 · Z.ai · 광주 B300
- GLM 5.3 Flash · Z.ai · 광주 B300
- DeepSeek V4.1 Flash · DeepSeek · 광주 B300
- Qwen3.8 27B · Alibaba Qwen · 광주 B300
- Qwen3 Embedding 4B · Alibaba Qwen · 인천 H100

개발사 이름은 그대로 적습니다 — 바뀐 것은 추론이 일어나는 장소입니다. 조직 관리자는 모델 허용 목록을 국내 추론 모델로만 제한할 수 있습니다. 개발사 공개 벤치마크, 사내 개발자 수 기준 월·연 비용 시뮬레이터, 응답 시간 실측 원본: bizrouter.ai/k-openweight

명령 앞에 bizrouter 한 단어.
코딩 에이전트가 BizRouter로 옵니다.

개발자는 이미 Claude Code, Codex, OpenCode, Hermes를 씁니다. BizRouter CLI는 그 습관을 건드리지 않습니다. **설정 파일은 한 줄도 바꾸지 않고**, 그 실행의 모델 호출과 과금만 BizRouter를 거치게 합니다. 사용량·모델 정책·감사 로그는 콘솔에 그대로 남습니다.

```
terminal

$ npm install -g bizrouter
$ bizrouter login  // 브라우저에서 8자리 코드를 승인 — 한 번만
$ bizrouter claude   // Claude Code 가 평소처럼 뜹니다. 달라진 건 과금·정책·로그의 행선지뿐
$ bizrouter codex --model openai/gpt-5.6-sol
$ bizrouter search "한국은행 기준금리 최근 결정"  // Search 를 터미널에서
```

한 번의 승인으로 콘솔이 명령이 됩니다

터미널의 코드를 브라우저에서 승인하면 이 기기용 세션과 API 키가 함께 발급됩니다. API 키 발급, 사용량·청구 조회, 스마트 라우팅 정책, 감사 로그까지 콘솔에서 클릭하던 일을 `bizrouter keys · usage · routing · audit`로 합니다.

에이전트가 콘솔과 Search를 도구로 씁니다

로그인한 뒤 실행한 Claude Code·Codex·OpenCode에는 BizRouter MCP 서버가 함께 연결됩니다. 에이전트가 키를 만들고 정책을 읽고, 최신 법령·공시가 필요하면 모델의 기억 대신 Search로 실제 데이터를 찾아 답합니다.

권한은 넓어지지 않습니다

API 키는 예전처럼 모델 호출에만 쓰입니다. 콘솔 작업은 승인한 사람의 웹 콘솔 권한(owner/member)과 정확히 같은 범위에서만 됩니다. 2차 인증과 콘솔 IP 허용 목록도 그대로 적용되고, 세션은 콘솔 「로그인된 기기」에서 언제든 끊을 수 있습니다.

에이전트가 지키는 규칙

조회는 자유, 만들기·바꾸기·삭제는 사용자 확인이 있어야만 실행됩니다. 전체 API 키를 파일·커밋·대화에 남기지 않고, 결제·크레딧은 명시적으로 요청받았을 때만 건드립니다. 규칙은 에이전트용 매뉴얼로 함께 배포됩니다.

도구	실행 시 하는 일	설정 파일
Claude Code	BizRouter 주소·인증을 이번 실행에만 주입. 사용자 설정에 다른 게이트웨이가 있어도 항상 BizRouter로. 모델 선택창은 쓸 수 있는 Claude 모델로 자동 채움	변경 없음
Codex CLI	BizRouter provider와 모델·추론 강도를 실행 옵션으로 지정. 요금 계약이 없는 옵션은 이번 실행에서만 기본값으로	변경 없음
OpenCode · Hermes	키가 쓸 수 있는 모델 전체를 실제 한도와 함께 주입. 세션 전체에 Search를 켜는 옵션 제공	변경 없음

npm 패키지 bizrouter · Node.js 18 이상 · 런타임 의존성 없음. 문서 bizrouter.ai/docs/cli · 소개 bizrouter.ai/cli

11 — INTERNAL CONTROLS

쓰게 하되, 새지 않게.
계층별 내부 통제 체계.

BizRouter의 통제는 안내문이 아니라 **게이트웨이에서 강제되는 정책**입니다. 조직 → API 키 → 요청, 세 계층에서 예산·모델·접근을 각각 통제하고 모든 예외는 즉시 차단됩니다.

통제 대상	조직 수준	API 키 수준	요청 수준
예산·한도	선불 크레딧 잔액 또는 후불 월 사용 한도. 초과 즉시 차단, 임계값 도달 시 알림	키별 누적·일별·월별 한도와 경보 임계값 (기본 80%), 지정 수신자 알림	확정 원가를 요청 단위로 기록·회신
모델 접근	조직 차단 목록 — 금지 모델은 어떤 키로도 호출 불가	키별 허용 목록 — 지정한 모델만 호출 가능	스마트 라우팅 허용 모델 제약 헤더 — 기존 정책을 좁히는 방향으로만 동작
호출량	조직 기본 분당 요청·버스트 한도	키별 분당 요청·버스트 한도 재정의	초과 시 표준 429 응답
네트워크	—	키별 IP 허용 목록(단일 IP·CIDR 대역, IPv4/IPv6)	허용 외 출처 즉시 거부

조직·구성원 관리

owner와 member 역할 분리. 보안 설정·감사 로그·키 삭제는 owner 전용입니다. 이메일 초대(유효기간 만료)와 기존 계정 이관을 지원합니다.

API 키 수명주기

키는 해시로만 저장되고 발급 시 한 번만 노출됩니다. 삭제된 키도 감사 이력은 보존되어 과거 사용 기록의 추적성이 유지됩니다.

부서·프로젝트별 비용 배분

요청에 사용 주체·그룹 태그를 강제할 수 있어, 하나의 키를 쓰더라도 부서·프로젝트·최종 사용자 단위로 사용량이 집계됩니다.

조회 전용 Usage API

과금·사용량·로그 조회는 발급 권한이 분리된 전용 읽기 키로 제공합니다. 정산 시스템 연동에 호출 권한을 넘겨줄 필요가 없습니다.

운영자 행위도 기록

관리자가 고객 화면을 대신 확인하는 순간까지 사유·시각·접속 정보가 남습니다. 통제하는 쪽도 감사 대상입니다.

재배포 없는 정책 변경

모델 차단·허용, 한도, 보안 정책 변경은 콘솔에서 즉시 반영됩니다. 개발팀 배포 일정을 기다릴 필요가 없습니다.

모든 요청이 지나는 거버넌스 파이프라인

01
요청 출처 식별
키·구성원·IP·사용 주체 태그로 요청의 주인을 확정합니다

›

02
정책 적용
한도·모델 통제·보안 정책을 요청마다 강제합니다

›

03
모델 라우팅
허용 범위 안에서 지정 모델 또는 스마트 라우팅으로 호출합니다

›

04
감사 로그 기록
누가·무엇을·얼마에 — 확정 원가와 함께 남깁니다

12 — SECURITY & COMPLIANCE

프롬프트가 회사 밖으로 나가기 전, 마지막 검문소.

외부 모델로 향하는 모든 요청은 게이트웨이의 보안 파이프라인을 통과합니다. 개인정보는 가리고, 금지된 내용은 세우고, 지나간 것은 전부 기록합니다 — **개인정보보호법(PIPA)을 염두에 둔 설계**입니다.

개인정보 자동 마스킹

주민등록번호·여권번호를 요청 단계에서 자동 마스킹합니다. 예: `901231-1234567 → 901231-1******`
원문은 저장되지 않습니다 — 감사 로그에는 마스킹된 텍스트와 탐지 유형 코드만 남습니다.

사내 금칙어 차단

조직별 금칙어 사전을 등록하면 대소문자 구분 없이 요청을 차단하고 차단 이력을 남깁니다. 프로젝트 코드네임, 미공개 제품명 같은 사내 기밀 유출을 게이트웨이에서 막습니다.

Corepin 콘텐츠 필터 (선택)

한국어 특화 AI 콘텐츠 필터 엔진을 연동하면 이름·연락처 등 확장 개인정보 비식별화, 유해 발화 차단, 정신건강 위험 신호 모니터링(차단 없이 기록)까지 확장됩니다. 필터별로 켜고 끌 수 있습니다.

BYOK — 자사 계약 키 반입

이미 보유한 프로바이더 계약이 있다면 자체 키를 암호화 보관 방식으로 등록해 그대로 쓸 수 있습니다. 통제와 감사는 BizRouter 체계를 유지한 채로요.

감사 로그 — 요청 하나에 남는 것들

조직·사용자·API 키 | 모델·프로바이더 | 입력·출력·캐시 토큰 | 확정 원가(KRW)·원가 구성 | 마스킹·차단 이벤트 | 사용 주체·그룹 태그 | 지연 시간·처리 속도 | 라우팅 결정·제외 사유 | 실패 원인·재시도 이력

감사 로그는 owner 권한으로 기간을 지정해 파일로 내보낼 수 있고, 내보낸 파일에는 만료 기한이 적용됩니다. 요청·응답 본문을 아예 저장하지 않는 **콘텐츠 무저장 모드**, 파일 업로드 차단 옵션도 조직 단위로 제공합니다.

> "어떤 프롬프트가, 누구 이름으로, 어느 모델에, 얼마에 나갔는가" — 보안팀과 재무팀이 같은 질문을 합니다. BizRouter는 이 질문에 **같은 로그**로 답합니다.

달러 영수증 12장 대신,
원화 청구서 1장.

프로바이더마다 흩어진 달러 결제를 원화 기준 단일 체계로 통합합니다. 모든 요청의 원가는 응답과 동시에 원화로 확정되고, 재무팀이 쓰는 리포트까지 한 번에 이어집니다.

KRW
요청별 확정 원가를 응답의 usage.cost로 즉시 회신

1장
전 프로바이더 사용액을 묶은 단일 청구

B2B
후불 결제·세금계산서·거래 증빙 등 국내 정산 지원

선불 · 후불, 조직에 맞게

선불 크레딧 충전(카드·계좌이체)과 후불 월 정산을 모두 지원합니다. 후불 조직은 월 사용 한도로 상한을 통제하고, 등록 카드 자동 결제로 정산을 자동화할 수 있습니다.

임계값 알림

크레딧 잔액·월 사용액·키별 한도 각각에 알림 임계값을 걸 수 있습니다. 예산이 소진되고 나서가 아니라, 소진되기 전에 압니다.

월별 사용 리포트

조직·모델별 요청 수, 토큰, 원화 비용 구성을 월 단위로 집계합니다. 부서·프로젝트 태그와 결합하면 사내 배부 기준으로 바로 쓸 수 있습니다.

변조 불가능한 정산 원장

모든 과금 항목은 추가만 가능한 원장에 기록되고 정정도 새 항목으로 남습니다. "이 금액이 왜 나왔는가"를 언제든 역추적할 수 있습니다.

response.usage — 모든 응답에 원가가 함께 옵니다

```
"usage": {
  "prompt_tokens": 1204,
  "completion_tokens": 512,
  "prompt_tokens_details": { "cached_tokens": 896 },   // 캐시 절감분까지 반영
  "cost": 12.84  // 확정 원가, KRW — 월말 추정이 아닌 요청 시점 확정
}
```

사용한 만큼의 원가가 **요청 시점에 원화로 확정**되기 때문에, "이번 달 예상 요금"이 아니라 "지금까지의 확정 요금"으로 예산을 운영할 수 있습니다.

B BizRouter

GET STARTED

오늘 붙이고,
내일부터 통제하세요.

가입 후 API 키를 발급받고 base_url 한 줄을 바꾸는 것으로 시작됩니다. 코딩 에이전트는 `bizrouter claude` 한 줄이면 됩니다. 스마트 라우팅은 응답을 바꾸지 않는 검증 모드부터 켤 수 있어, 도입 리스크가 없습니다.

도입 3단계

1 bizrouter.ai에서 조직 계정을 만듭니다
2 콘솔에서 API 키를 발급하고 한도·보안 정책을 설정합니다
3 기존 코드의 base_url 한 줄, 또는 명령 앞에 `bizrouter` 한 단어 — 끝입니다

WEB
bizrouter.ai

CONTACT
support@bizrouter.ai

© BizRouter