

한국 민사 판결문 구조화·요약 태스크를 위한 최적 LLM 추천 심층 리서치

TL;DR

- 1순위 추천은 **Claude Sonnet 5**다. 한국 법률 추론 벤치마크에서 프론티어 상위권(KCL 기준 Claude 계열이 Gemini·GPT 바로 다음), GA 상태의 constrained-decoding 기반 structured outputs(스키마 100% 강제), effort 파라미터·prompt caching으로 정확도-비용을 동시에 잡을 수 있고, 사용자의 기존 Claude API 파이프라인에 그대로 얹힌다. 1회 작업(입력 ~4만·출력 ~5천 토큰) 실질 비용은 도입가 기준 약 \$0.13, 배치 시 그 절반 수준이다.
- 최고 정확도 우선이라면 **Gemini 3.1 Pro**를 진지하게 병행 평가하라. 한국어 법률 추론 벤치마크(KCL)에서 Gemini 2.5 Pro가 MCQA·Essay 모두 단독 1위였고, 네이티브 JSON 스키마 강제가 가장 신뢰도 높으며, 1M 컨텍스트에 가격도 \$2/\$12로 오히려 저렴하다(약 \$0.14/작업).
- 최저 비용·한국어 특화로는 **Upstage Solar Pro 3**(\$0.15/\$0.60, 약 \$0.009/작업, 100% 스키마 준수 표방, 법률 도메인 타깃)를 1차 처리 계층으로, 셀프호스팅은 **EXAONE 4.0 32B / Qwen3** 계열을 Mac Studio에서 돌리는 방안이 가능하나 근거기반 추론 정확도·롱컨텍스트 프리필 속도에서 손해를 본다.

Key Findings

1. 판결문 구조화는 "회상"이 아니라 "공급된 텍스트에서의 근거기반 추론"이며, 이 영역은 프론티어 추론 모델이 지배한다. 사용자의 태스크는 RAG로 검색된 판결문 2-3건을 LLM에 넣고 요건사실 단위로 분해·구조화하는 것이다. 이는 KCL 벤치마크의 "w/ supporting precedents" 설정(관련 판례를 함께 공급)과 동형이다. 이 설정에서 프론티어 추론 모델들은 한국 변호사시험 MCQA에서 80%대를 넘겨 인간 합격선(68.7%)을 상회한다. 즉 모델의 한국 법 사전지식 자체보다, 공급된 텍스트를 정확히 읽고 구조화하는 추론·지시이행·롱컨텍스트 이해가 결정적이다.

2. 한국어 법률 추론 벤치마크 성적 (최우선 평가기준):

- KCL (Oh, Hwang & On, LBOX·University of Seoul, EACL 2026 Short Papers pp.232-243; arXiv:2512.24572): "KCL-MCQA, multiple-choice problems of 283 questions with 1,103 aligned precedents, and (2) KCL-Essay, open-ended generation problems of 169 questions with 550 aligned precedents and 2,739 instance-level rubrics"로 30+ 모델을 평가. 논문 본문은 "Among Large Reasoning models, **Gemini 2.5 Pro performs highest overall**, followed by Claude and GPT-5"라고 명시한다. 구체적으로 vanilla→판례공급 점수는 다음과 같다:

- KCL-MCOA: Gemini 2.5 Pro 66.1→**89.3%**(1위). Claude Opus 4.1 57.2→83.3.

Claude Sonnet 4.5 54.9→85.2, GPT-5 53.9→84.2, o3 50.2→83.4. Small Reasoning에서는 Gemini 2.5 Flash 51.8→83.7이 최상.

- **KCL-Essay(오픈형 IRAC):** Gemini 2.5 Pro 60.5→**74.9%**(1위), GPT-5 57.5→72.8, o3 51.2→69.4, Claude Sonnet 4.5 47.6→68.5, Claude Opus 4.1 45.5→60.3. 오픈형 에세이에서는 **GPT-5·o3가 Claude** 계열 전체를 상회했다.
- **추론특화 모델이 일반 모델을 일관되게 상회.** 판례 공급 시 프론티어 추론 모델은 MCQA에서 인간 합격선(68.7%)을 넘지만, Essay는 최고가 74.9%로 ~25%p 격차가 남았다. (arXiv)

- **한국어 특화 모델의 위치:** A.X 4.0(SKT), EXAONE 4.0 32B(LG), HCX SEED Think 14B(HyperCLOVA X)는 판례를 공급하지 않은 vanilla(사전지식) 설정에서 강했다 — 특히 A.X 4.0는 MCQA vanilla 49.1로 동급 일반모델 중 GPT-4.1·Claude 3.5·프론티어 오픈모델을 앞섰다. 그러나 판례가 공급되는 근거기반 추론 설정에서는 프론티어 모델에 뒤처졌다(A.X 4.0 71.3, EXAONE 4.0 32B 72.3). 즉 한국 법률 지식은 우수하나 근거기반 추론은 약하다 — 사용자 태스크 (공급된 판결문 구조화)에서는 이 약점이 그대로 드러날 수 있다. (arxiv)
- **KBL (Kim et al., Findings of EMNLP 2024 pp.5573–5595; aclanthology 2024.findings-emnlp.319):** "7 legal knowledge tasks(510), 4 reasoning tasks(288), Korean bar exam(4 domains, 53 tasks, 2,510)"으로 구성. **GPT-4가 종합 1위**(2024 변호사 시험 48.1%, 지식 AVGK 72.0), Claude-3.5-sonnet 근소 2위(AVGK 70.0, 순수 추론평균 89.3으로 오히려 최고). 당시 어떤 모델도 한국 변호사시험 합격선에 미달. (구세대 모델 기준이나 추세 확인용.)

3. 영어권 법률 벤치마크에서는 Claude가 최상위권: Vals AI Legal Research Bench(미국 법률 에이전트형)에서 **Claude Opus 4.8이 43.75%로 1위**, Claude Sonnet 5(41.83%), GPT-5.5(40.38%) 순. LegalBench(정적 지식·추론)에서는 Claude Fable 5(88.56%), Gemini 3.1 Pro(87.40%), Gemini 3 Pro(87.02%), Gemini 3 Flash(86.86%), GPT-5.5(86.52%) 순. 즉 Claude·Gemini·GPT 상위 티어의 법률 역량은 몇 점 차 안에 밀집. (Vals AI) (Vals AI)

4. Schema adherence(2순위 기준): OpenAI·Gemini의 네이티브 structured output이 constrained decoding으로 스키마를 생성 시점에 강제해 가장 신뢰도가 높다(복잡 스키마 1,000샘플 테스트에서 GPT ~99.9%, Gemini ~99.5–99.8% 준수, 3–4단 중첩 시 실패율 1–2%로 상승). **Claude도 2026년 2월 structured outputs가 GA 전환**(Sonnet 4.5/Opus 4.5/Haiku 4.5, constrained decoding으로 스키마 보장, (output_config.format) + (strict) tool use)되어 이제 동급 신뢰도를 확보했다. Solar Pro 3은 "claims 100% schema compliance for structured output generation"을 표방한다(전작 Solar Pro 84.76%→100%로 개선). (Claude)

5. 경제성(1회 작업당 실질 비용, 입력 ~4만·출력 ~5천 토큰 가정):

모델	입력/출력 (\$/1M)	1회 작업 실질비용 (표준)	배치 50%↓	비고
Claude Haiku 4.5	\$1 / \$5	~\$0.065	~\$0.033	저비용 1차 처리
Claude Sonnet 5 (도입가)	\$2 / \$10	~\$0.13	~\$0.065	표준가 \$3/\$15는 9/1부터
Claude Opus 4.8	\$5 / \$25	~\$0.325	~\$0.16	최고 품질 검증용

Gemini 3.1 Pro	\$2 / \$12	~\$0.14	~\$0.07	1M 컨텍스트, 200K 초과 시 \$4/\$18
Gemini 3.5 Flash	\$1.50 / \$9	~\$0.105	~\$0.053	중위 가성비
Gemini 3 Flash	\$0.50 / \$3	~\$0.035	~\$0.018	저비용
GPT-5.6 Terra	\$2.50 / \$15	~\$0.175	~\$0.088	GPT-5.4와 동급
GPT-5.6 Luna	\$1 / \$6	~\$0.07	~\$0.035	저비용 프론티어
Solar Pro 3 (Upstage)	\$0.15 / \$0.60	~\$0.009	—	한국어 특화, 128K, 최저가
EXAONE 4.0 32B / Qwen3 (셀프호스팅)	전력비만	~\$0 마진	—	Mac Studio, 느린 프리필

프롬프트 캐싱은 3사 모두 캐시 읽기를 기본가의 ~10%로 할인(90%↓)한다. 다만 **판결문 본문**은 쿼리마다 달라 캐시 불가하고, 반복되는 것은 system prompt·규칙 문서(수천~1만 토큰)뿐이므로 캐싱 절감폭은 제한적이다. 오프라인 대량 처리라면 배치 **API 50% 할인**이 캐싱보다 절감효과가 크다. (EvoLink) (BenchLM)

6. 한국어 토큰나이제이션 효율: 대부분의 BPE 토큰나이저에서 한국어 fertility(단어당 토큰)는 최적 BPE·GPT-4o가 ~1.5, Gemma ~1.9, Llama ~2.3 수준이다. Gemini의 256K SentencePiece 어휘가 한국어에 유리한 편. 주의: **Claude Sonnet 5/Opus 4.7**의 신형 토큰나이저는 동일 텍스트에 대해 **1.0~1.35배 더 많은 토큰을 소비**하므로, Sonnet 4.6 대비 실질 비용이 최대 35% 상승할 수 있다(도입가는 이를 상쇄하도록 설계됨). 한국어 특화 모델(Solar, HyperCLOVA)은 한국어 토큰 효율이 상대적으로 높다. (arxiv + 2)

7. 국내 법률 AI 업계 동향: 로앤컴퍼니의 슈퍼로이어(SuperLawyer)는 공식 발표(2024.7.1) 기준 "자체 설계한 아키텍처를 바탕으로 **복수의 상용 거대언어모델(LLM)**로 구현된 클라우드 기반 SaaS"로, 국내 최다 458만 건 판례 데이터+RAG를 활용하고 '팩트체커'로 할루시네이션을 억제한다(관련 특허 3건 출원). 실제로 로앤컴퍼니는 ChatGPT·Claude 등 복수 상용 LLM을 활용한다고 공개적으로 밝혔다. 엔터프라이즈용으로는 업스테이지와의 협약(2024.3.12)으로 개발한 "국내 최초 법률 특화 LLM 파운데이션 모델" **솔라 리걸**(국내 최다 443만 건 판례+법령·결정례·유권해석 등 총 16만 건 학습, 온프레미스 우선)을 활용한다. LBox는 자체 법률 데이터+RAG로 범용 LLM의 환각을 억제하는 접근을 강조한다. **업계 공통 교훈: 범용 LLM을 그대로 쓰면 판례·사건번호 환각이 필연**이므로, 사용자처럼 판결문 원문을 RAG로 공급하고 그 위에서만 생성하게 하는 설계가 정답이다.

Details

왜 Claude Sonnet 5를 1순위로 두는가 (Pareto 최적)

사용자 태스크의 성격 — 창의성보다 정확성·일관성, 긴 입력, 엄격한 출력 포맷 — 은 "추론력 있는 지시이행형 모델 + 스키마 강제 + 저비용 튜닝"을 요구한다. Sonnet 5는 (a) 한국어 법률 추론 상위권(KCL-MCQA 판례공급 85.2%로 Opus 4.1 상회), (b) GA structured outputs로 스키마 보장, (c) (effort)(medium/high/xhigh) 파라미터로 요건사실 분해 같은 난도 구간만 추론량을 올리고 나머지는 아껴 비용을 통제, (d) 1M 컨텍스트로 판결문 2~3건 여유 수용, (e) 사용자의 기존 (cache_control:

ephemeral]·effort 최적화 자산을 그대로 재사용, 이라는 다섯 요소를 모두 만족한다. 정확도 절대 1위는 아니지만 통합비용·엔지니어링 리스크를 합산한 총소유비용에서 최적이다. 한 가지 주의: Sonnet 4.6에서 Sonnet 5로 옮길 때 temperature/top_p 등 샘플링 파라미터, 수동 extended thinking 설정이 400 에러를 유발하는 breaking change가 있으므로 마이그레이션 코드 점검이 필요하다.

왜 Gemini 3.1 Pro를 강력한 도전자로 두는가

객관적 한국어 법률 추론 성적(KCL)에서 Gemini 2.5 Pro가 MCQA·Essay 모두 단독 1위였고 후속인 3.x가 이를 계승한다. 네이티브 스키마 강제 신뢰도 최상위, \$2/\$12로 Claude Opus·GPT-5.5보다 저렴, 1M 컨텍스트. 정확도가 최우선이라는 사용자 기준을 문자 그대로 따르면 오히려 1순위 후보다. 다만 **200K 토큰 초과 시 효율이 2배(\$4/\$18)로 뛰므로** 판결문 다건 입력 시 200K 이내 유지가 비용 관건이다. 통합 측면에서는 새 SDK 도입이 필요해 Claude 파이프라인 재사용 이점은 없다. (BenchLM)

한국어 특화·오픈웨이트의 실제 위치

Solar Pro 3(102B MoE/12B active, 128K, \$0.15/\$0.60, 2026.1.27 출시)은 한국어·법률 엔터프라이즈를 명시적으로 겨냥하고 Tau2-all 72.3(전작 36.0의 2배)·Ko-Arena-hard-v2 78.2의 에이전트/한국어 성과와 100% 스키마 준수를 강조한다. 비용은 압도적으로 낮아(작업당 ~\$0.009) 1차/대량 처리에 이상적이나, 독립된 한국 법률 추론 벤치마크 공개 성적이 프론티어만큼 검증되지 않았다. EXAONE 4.0 32B·Qwen3 32B는 오픈웨이트로 Mac Studio(M3 Ultra급)에서 MLX로 구동 가능하나, 4만 토큰 롱 컨텍스트 프리필이 수 분 소요(32B Q8 기준 4K 컨텍스트에서도 TTFT가 길고 32K에서는 분 단위 대기)되고 근거기반 추론에서 프론티어에 밀린다 — 프라이버시·제로 마진비용이 절대 우선일 때만 권장. MoE형(예: Qwen3-35B-A3B)은 활성 파라미터가 작아 Mac에서 상대적으로 빠르다.

요건사실론 관점의 태스크 적합성

요건사실(청구원인 사실) 단위 분해는 본질적으로 IRAC형 구조화 추론이다. KCL-Essay가 바로 이 IRAC/근거기반 생성을 평가하는데, 여기서 추론특화 프론티어 모델(Gemini 2.5 Pro, GPT-5, o3)이 일반 모델을 크게 앞섰다. 사해행위취소·채권자취소권·소유권이전등기말소·근저당권설정등기말소·보증채무금 같은 한자어 법률 용어의 이해는 프론티어 모델과 한국어 특화 모델 모두 vanilla 지식에서 준수하나, 판결문 원문이 공급된 상태에서의 정확한 매핑·구조화는 추론력이 좌우한다. 따라서 "한국어 특화=무조건 유리"가 아니라, 근거기반 추론력이 높은 모델을 고르되 출력 검증으로 용어 오류를 잡는 설계가 실무적으로 우월하다.

Recommendations

1단계 (즉시): 기존 Claude 파이프라인 위에서 **Claude Sonnet 5 + structured outputs**(`output_config.format`, **strict tool use**) + **effort=medium** 기본, 요건사실 분해 프롬프트에서만 **high**로 프로덕션 기준선을 만든다. 반복 system prompt·규칙 문서는 prompt caching, 야간 대량 처리는 배치 API로 돌린다. 마이그레이션 시 샘플링 파라미터/수동 thinking 설정 제거.

2단계 (병행 A/B): 동일 판결문 세트로 **Gemini 3.1 Pro**와 **Solar Pro 3**을 각각 20-50건 벤치마크한다. 평가지표는 요건사실 추출 정확도(전문가 채점)·스키마 위반율·1건당 비용. Gemini가 정확도에서 유의미하게 앞서면(예: 요건사실 정확도 +5%p 이상) 정확도 우선 라인으로 채택. Gemini는 입력을 200K 토큰 이내로 유지해 효율 2배 구간을 피할 것.

3단계 (계층적 라우팅 — 권장 최종형): **Solar Pro 3** 또는 **Gemini 3 Flash**로 1차 초안 → **Sonnet 5**로 검증·교정하는 2단계 라우팅. 저비용 모델이 자신 있게 스키마 준수 결과를 내면 그대로 채택, 스키마 위반·저

신뢰 케이스만 상위 모델로 에스컬레이션하면 블렌디드 비용이 단일 Sonnet 대비 약 30-40% 절감된다.
요건사실처럼 오류비용이 큰 필드는 항상 상위 모델이 최종 확정하도록 분리.

임계값(전략 변경 트리거):

- ① 1차 모델 스키마 위반율 >2% → 상위 모델 비중 확대 또는 네이티브 스키마 강제 모델 (GPT/Gemini)로 전환.
- ② 요건사실 오분해율이 전문가 검수에서 >10% → Opus 4.8/Gemini 3.1 Pro로 상향.
- ③ 월 처리량이 커져 API 비용이 셀프호스팅 손익분기를 넘고 프라이버시·데이터반출 제약이 강하면 EXAONE 4.0 32B 셀프호스팅 검토(단, 롱컨텍스트 처리량 실측 후 결정).
- ④ Gemini A/B에서 요건사실 정확도가 Sonnet 5를 유의미하게 앞서면 정확도 우선 라인을 Gemini로 스위치.

Caveats

- **벤치마크 세대차:** KCL은 Gemini 2.5 Pro/GPT-5/Claude Sonnet 4.5 세대 기준이며, 본문에 인용한 Sonnet 5·Opus 4.8·Gemini 3.x·GPT-5.6은 후속 세대다. 최신 모델의 한국 법률 추론 성적은 공개 벤치마크가 아직 부족하므로 자체 A/B가 필수다.
- **LLM-as-judge 한계:** KCL-Essay는 641개 triplet을 한국 변호사 2개 그룹(각 3명)이 채점해 Gemini 2.5 Flash 심판과 "overall Pearson correlation of about 0.8"(최난도 20%에서 ~0.7)을 보였다. 절대 점수보다 상대 순위로 해석해야 한다.
- **가격 변동성:** 가격은 2026년 7월 웹 출처 기준이며 변동성이 크다(Sonnet 5 도입가는 8/31까지, 이후 \$3/\$15). 프로덕션 전 각 공급사 공식 페이지 재확인 필요. Claude 신형 토크나이즈의 한국어 토큰 증가분(최대 +35%)도 실측 반영할 것. (Anthropic)
- **검증 부족 영역:** Solar Pro 3의 한국 법률 추론 정확도, 셀프호스팅 EXAONE/Qwen3의 실측 롱컨텍스트 성능은 공개 근거가 제한적이므로 파일럿으로 직접 검증할 것.
- **환각 방지 필수:** 어떤 모델도 판례·사건번호를 자체 생성하게 두면 환각 위험이 있으므로(로앤컴퍼니도 범용 LLM의 가짜 판례 생성을 공개 지적), 반드시 RAG로 공급된 판결문 원문에만 근거해 구조화하도록 프롬프트·검증 레이어를 유지해야 한다. structured output의 스키마 준수는 형식 보장일 뿐 내용 정확성을 보장하지 않으므로, 요건사실 필드에 대한 규칙 기반/전문가 검수 레이어를 별도로 둘 것.