파라미터 수가 많을수록 연산량과 메모리 사용량이 급증해 학습·추론 비용도 높아집니다.
따라서 고사양 모델은 높은 성능을 주지만, 실사용 시 운영비용과 응답 속도 문제도 함께 고려해야 합니다.

- **대표적인 LLM 모델들의 파라미터 수와 컨텍스트 윈도우 길이**

모델명	파라미터 수	컨텍스트 윈도우	특징	적합한 용도
Llama 4 (Maverick)	400B	1M	초대형 다국어 대응, 멀티모달 처리	글로벌 대용량 문서 분석, RAG
DeepSeek R1	671B (MoE)	128K	추론 특화, 비용 효율	고난도 질문 응답, 분석 도구
Claude 3.7 Sonnet	(비공개)	350K	안전성과 판단력 강화	챗봇, 문서 정리, 비서형 AI
Mistral Small 3.1	24B	128K	가볍고 빠름, 오픈소스	내부 시스템 연동, 엣지 디바이스
Phi-4	14.7B	16K	저사양 환경 대응	임베디드 AI, 비용 민감 환경
GPT-4o	1.8T (추정)	2M	텍스트+이미지 동시 처리	최고 성능, 복합 AI 서비스
XGen-7B	7B	8K	중소규모 조직용	문서 요약, 간단 질의응답

3. 권한 관리는 어떤 식으로 이루어지는가? (관리자, 사용자)

권한 관리는 관리자(Admin)와 일반 사용자(User)에게 서로 다른 접근 범위를 부여하는 방식으로 구현된다.