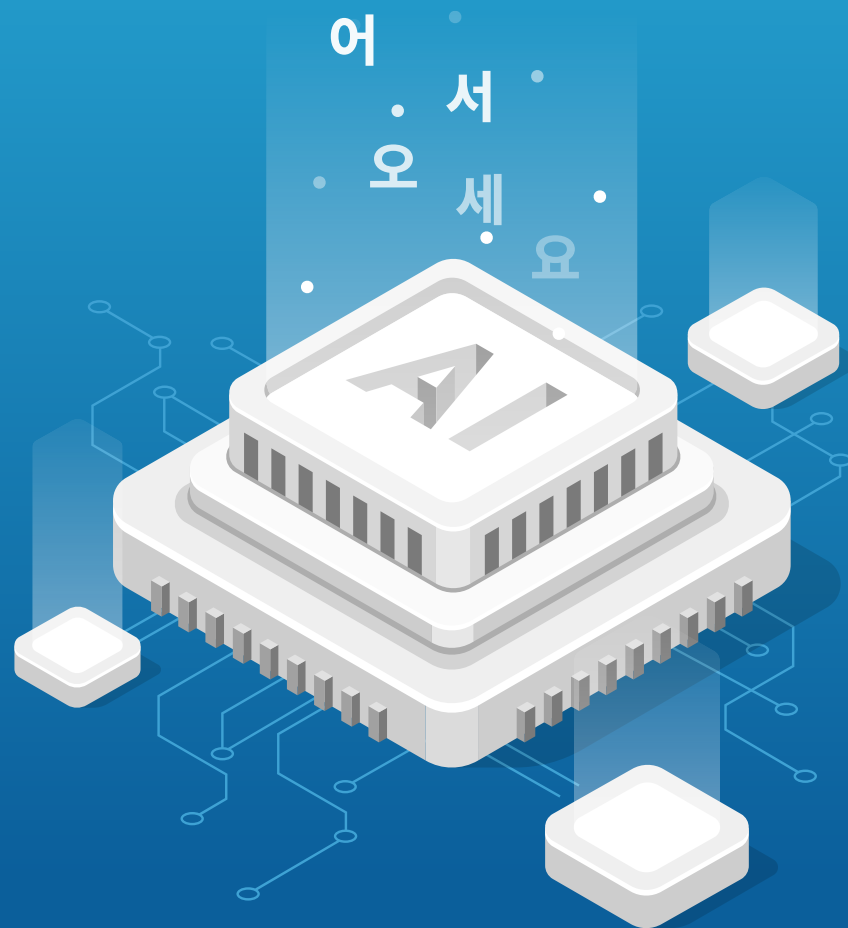


Leap, Leap ahead

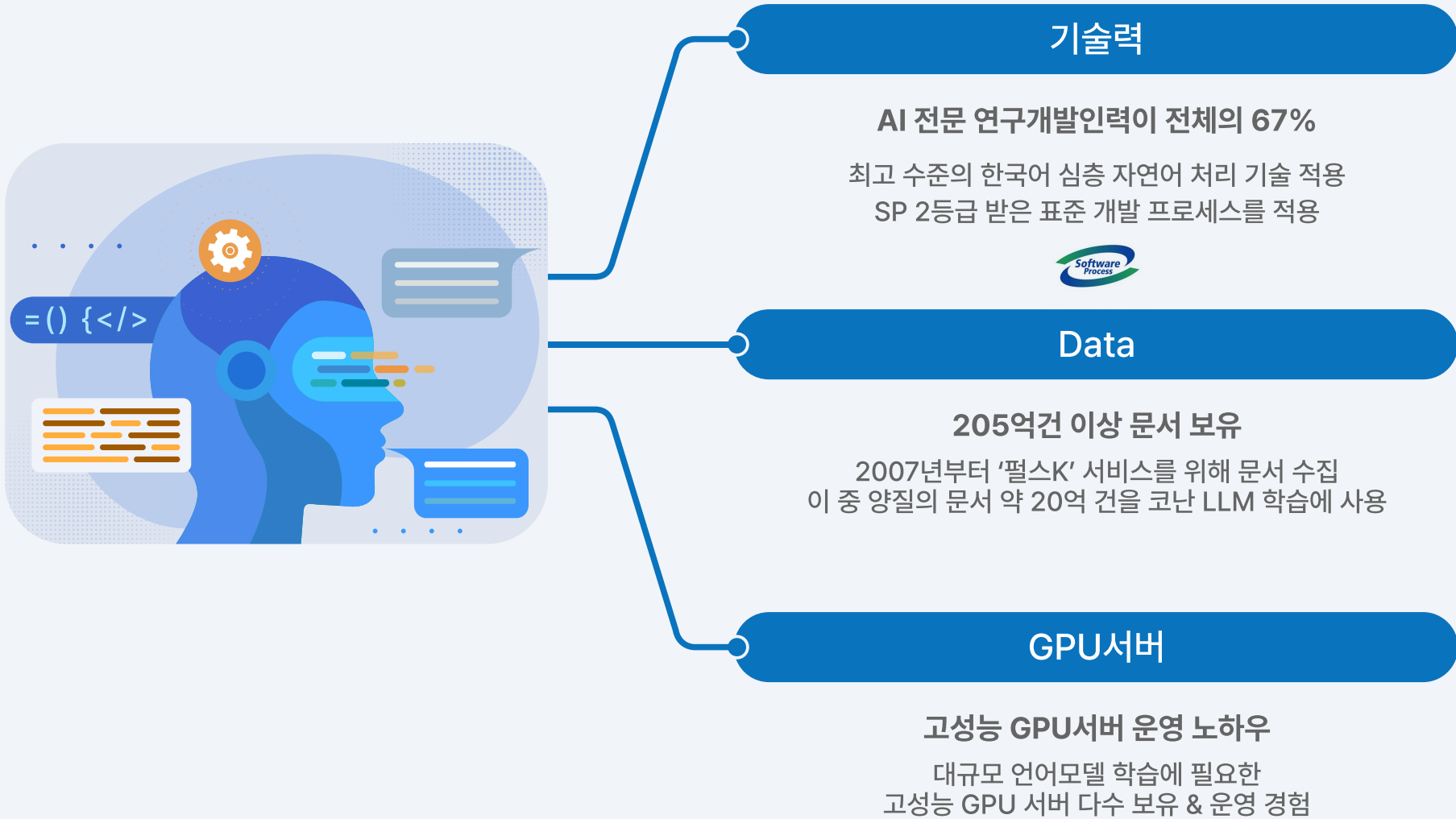
Konan LLM

대규모 언어모델



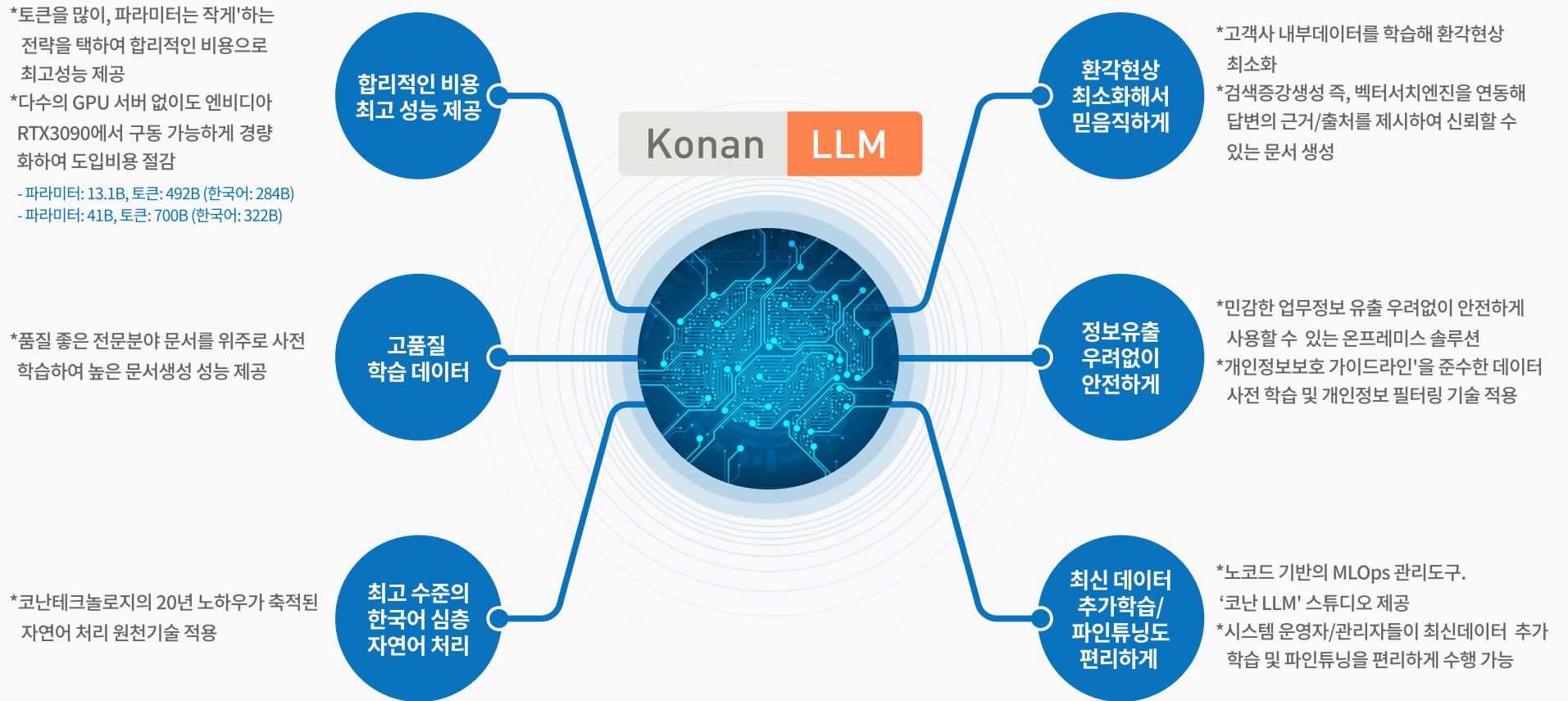
개요

코난 LLM은 코난테크놀로지의 자연어처리 전문 AI연구소에서 개발한 문서생성AI(대규모 언어모델)로 공공기관 및 기업의 업무혁신을 가속화합니다.



특징

코난 LLM은 합리적인 비용으로 최고 성능을 제공하는 온프레미스 솔루션입니다. 고객사 내부데이터를 학습해 환각현상을 최소화할 뿐만 아니라, RAG를 연동해 답변의 신뢰성을 높입니다.



활용분야

코난 LLM은 대표적인 능력인 문서생성 외에도 질문-답변, 요약, 고쳐쓰기, 분류 등 다양한 자연어 처리 능력을 갖고 있습니다.

자연어 인터페이스

물류, 제조, ERP 등
기간제 시스템에 자연어로
데이터 조회/분석/답변 가능

계약서 검토

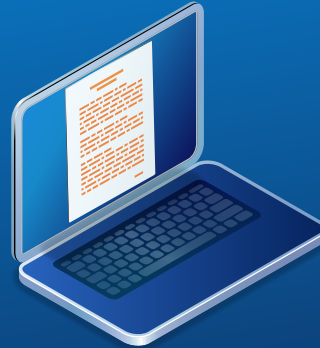
계약서 내 불공정 또는
함정 조항 탐지

타겟 마케팅

상담 내용 기반 고객 관심사
맞춤형 문구 초안 생성

업무문서/보고서 초안 생성

문서 작성 요청 프롬프트에
입력한 내용에 따라
자동으로 초안 생성



고객 맞춤 상담

AICC, 챗봇 등
고객 상담 시스템에서
고객 개인 정보에 기반한
답변 초안 생성

고객 상담 후처리 자동화

상담 내용 자동요약/분류/예약 등
레거시 시스템과 자동 연동

지속적 정보 수집과 실시간 대응 방안 초안 생성

정부·국방 분야에서 지속적으로
정보수집 시 효과적인 요약 제공/
긴급 시 대처방안 초안 생성



생성사례

코난 LLM이 생성한 '행정기관의 정책 보고서 초안' 사례입니다.

문서초안 생성사례

프롬프트 입력

전략 보고서

주제

'디지털플랫폼 기반의 과학적 재난안전관리 시스템 구현' 정책 보고서 초안을 작성해줘

참고문서

1. 2023년 행정안전부 업무보고.pdf X 2. (세정부 행정안전부 정책방향) 대통령 업무보고.pdf X

문서 생성

제목

제목: '디지털플랫폼 기반의 과학적 재난안전관리 시스템 구현' 정책 보고서 초안

- 전략 배경 및 필요성
- 기술 발전과 함께 4차 산업혁명이 본격화되면서 디지털플랫폼을 기반으로 하는 혁신적인 재난안전관리 시스템의 구현이 요구됨
- 급변하는 재난환경과 첨단기술을 적극 활용하여 국가의 재난대응 역량을 강화하기 위해 디지털 기술과 재난안전 관리기술을 융합한 새로운 차원의 재난안전관리 시스템 구축 필요
- 목적 및 범위
- 목적: '디지털플랫폼 기반의 과학적 재난안전관리 시스템 구현을 위한 정책방안 제시
- 범위: 전자정부법 제2조 정의에 따른 행정기관, 공공기관 및 민간부분의 디지털시스템 전체 대상

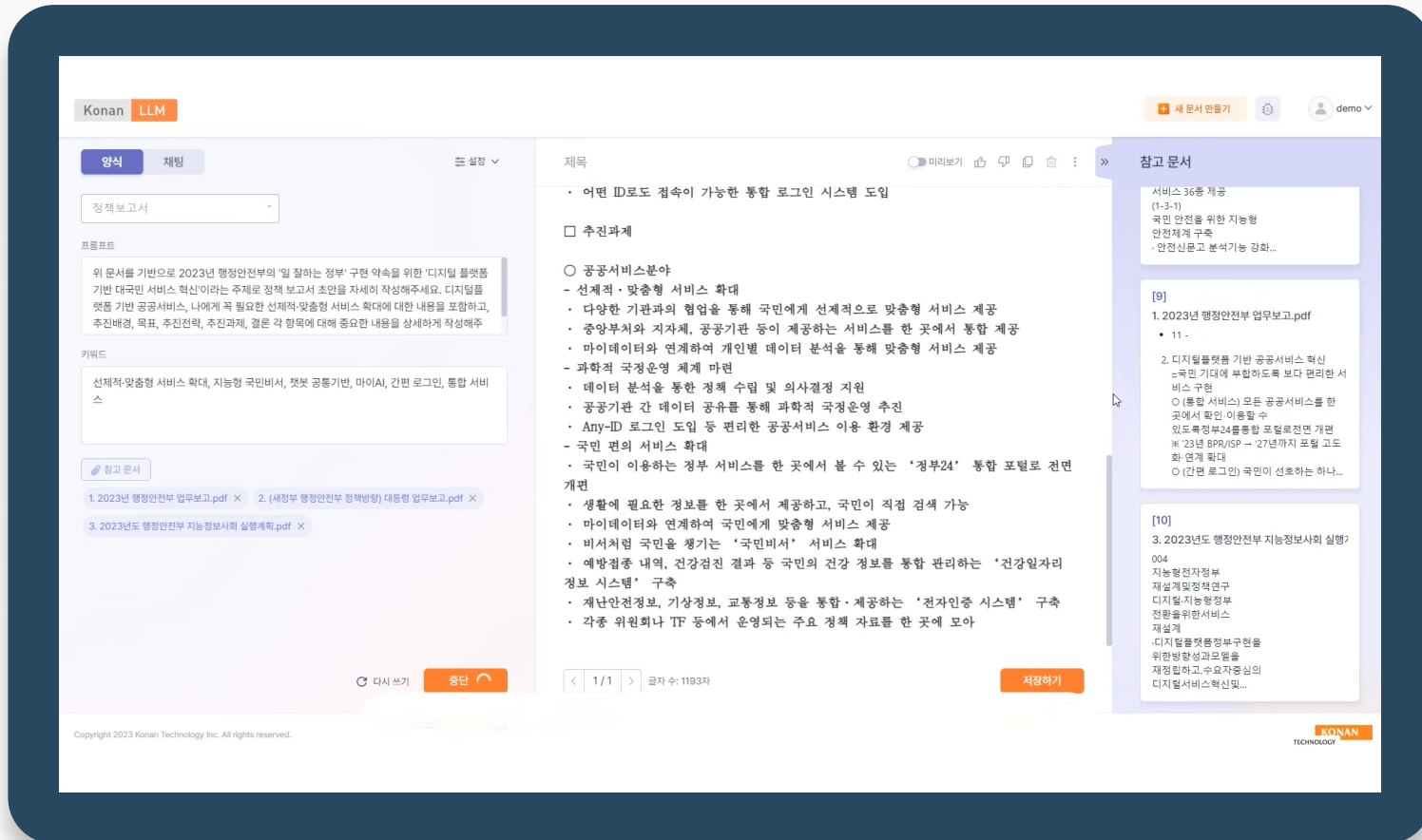
프롬프트 | 다시쓰기 | 줄여쓰기 | 늘려쓰기

보고서 초안 작성 방법



질문답변 사례

관련 법령 등 참고 문서에 근거하여 적절한 답변을 제시합니다.



자주 묻는 질문

기본 제공 템플릿 외 맞춤보고서 생성도 가능한가요? 

네. 파인튜닝을 통해 맞춤 형식의 보고서 출력이 가능합니다.

표나 그래프도 작성 가능한가요? 

표는 마크다운과 html 형식으로 작성 가능합니다. 향후 그래프 작성도 가능해지면, 제품 업데이트를 통해 알려드리겠습니다.

한국어 외 어떤 외국어를 지원하나요? 

현재 영어를 지원하고 있으며, 다양한 언어로 지원 확대할 예정입니다.

프롬프트에 입력 가능한 데이터는 어떤 것들이 있나요?



현재는 텍스트만 입력이 가능합니다.

프롬프트에 참고자료로 추가되는 첨부파일의 개수 제한이 있나요?



아니요. 제한 없습니다.

프롬프트에 입력할 수 있는 글자 수나 단어 수 제한이 있나요?



4k 토큰까지 입력 가능합니다. 4k는 A4 용지 6페이지 분량입니다.

문서저장 시 파일포맷은 어떤 것들을 지원하나요?



docx, txt, hwp 등 다양한 형식으로 저장 가능 합니다.

개인정보 보호조치는 어떻게 하고 있나요?



고객 개인정보 유출 피해가 발생하지 않도록 안전하게 보호하고 있습니다. 데이터 사전학습 과정에서부터 개인정보 보호 가이드라인을 준수하여 '개인정보 필터링'을 지원합니다.

파인튜닝을 쉽게 할 수 있는 방법을 제공하나요?



네. '코난 LLM 스튜디오'를 제공하여 최신데이터 추가학습 및 파인튜닝을 쉽고 편리하게 수행할 수 있습니다.

기반모델은 무엇인가요? 자체개발했나요?



Transformer 모델의 디코더에 기반하여 자체개발한 모델입니다.

클라우드로도 제공하나요?



아니요. 클라우드 상에서 SaaS형태로 제공하지 않지만, 고객사의 프라이빗 클라우드 위에서는 제공 가능합니다.

상품라인업은 어떻게 되나요?



2023년 8월에 출시한 13B모델과 12월에 출시한 41B모델이 있습니다.

GPU 하나로 돌아가나요?



네. 13B 추론서버는 GPU 하나로 운영 가능합니다. 다만, 학습 데이터양에 따라 필요 GPU개수는 더 추가될 수 있습니다. 자세한 내용은 아래 표를 참조하세요.

▣ GPU 장비-추론

모델	최소사양	권장사양
13B	RTX 3090(24GB) : 1장	RTX 3090(24GB) : 2장 A6000(48GB) : 1장 H100(80GB) : 1장
41B	RTX 3090(24GB) : 4장	RTX 3090(24GB) : 8장 A6000(48GB) : 4장 H100(80GB) : 2장

▣ GPU 장비-학습

모델	최소사양	권장사양
13B	RTX 3090(24GB) : 2장	A6000(48GB) : 8장 H100(80GB) : 4장
41B	A6000(48GB) : 8장	A6000(48GB) : 32장 H100(80GB) : 16장

*참고 : CPU의 경우, 모델 추론 이외에도 서버, 토큰나이저, 검증 등의 작업들을 위하여 "Intel(R) Xeon(R) Silver 4214R CPU @ 2.40GHz" 사양이 기본적으로 필요함.

필요한 하드웨어는 무엇 무엇이 있나요?



학습 및 추론용 GPU서버가 필요합니다. RAG 적용 시 추가적으로 벡터검색용 GPU 및 CPU 서버가 필요합니다.

더 궁금한 점이 있으신가요?
지금 바로 문의하세요!

