코난테크놀로지의 대규모언어모델 (LLM) 개발 전략

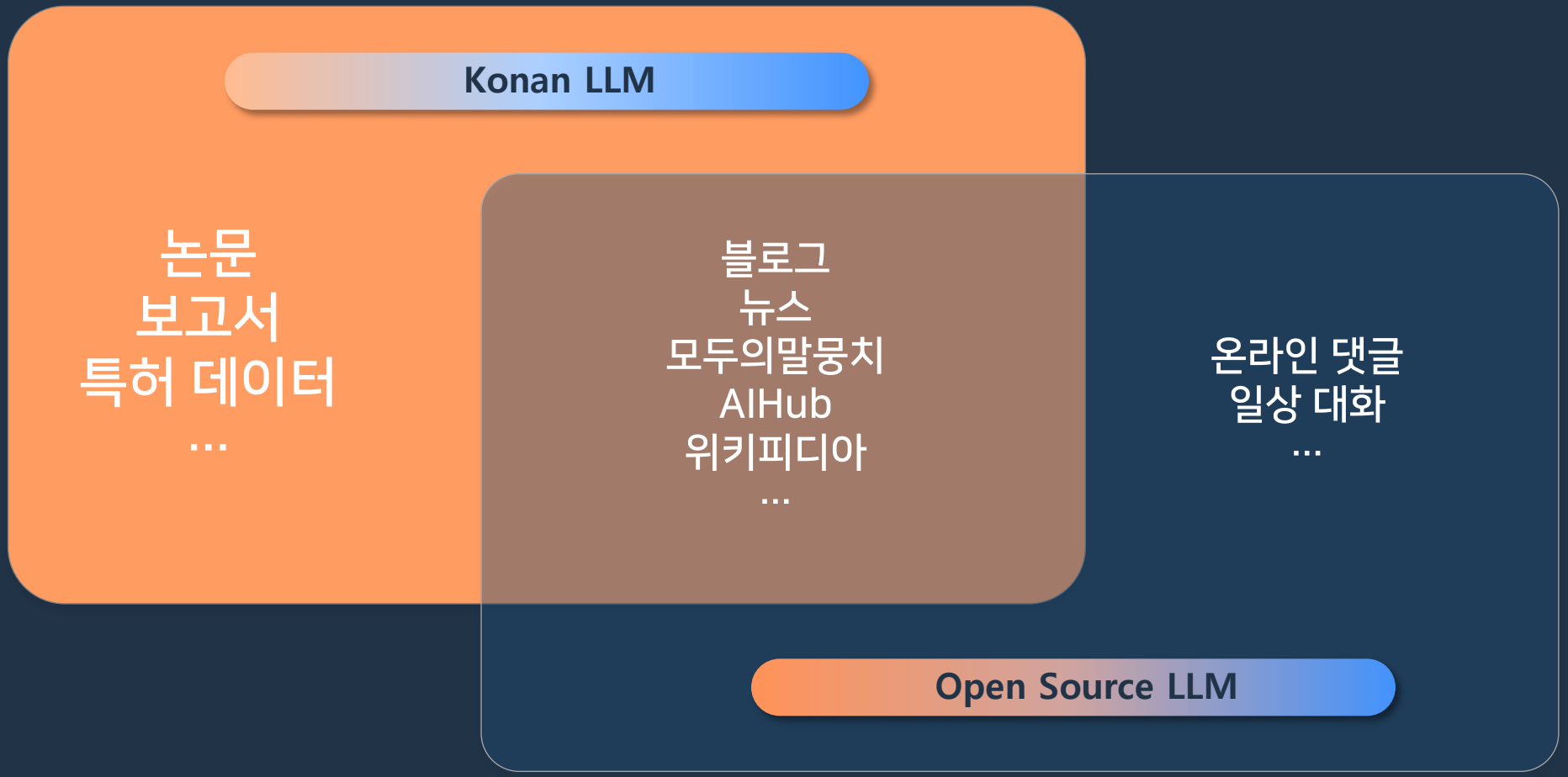
도원철, 코난테크놀로지 연구소, 상무

KONAN TECHNOLOGY

LLM

KONAN TECHNOLOGY

사전학습 데이터셋



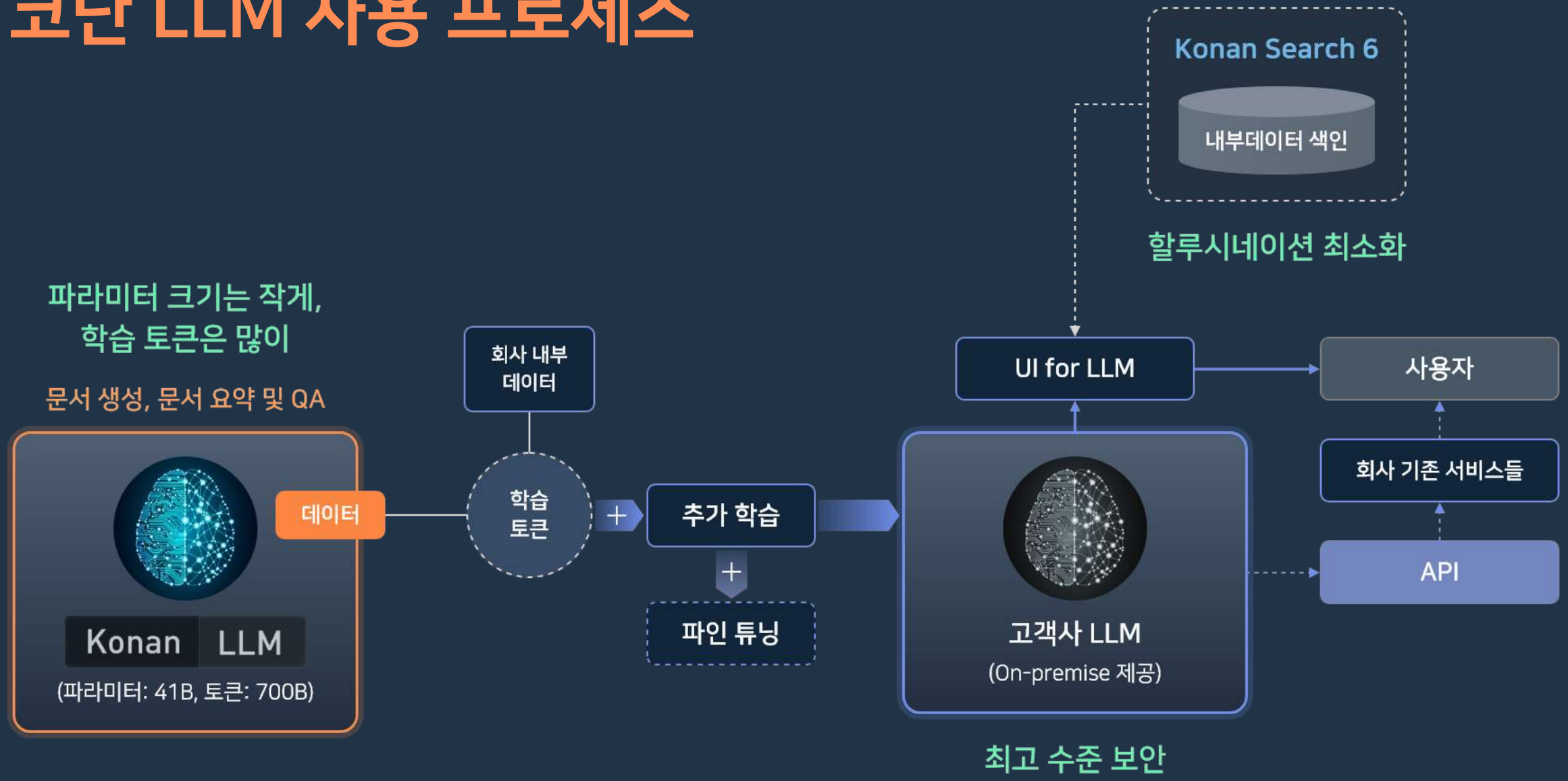
KONAN TECHNOLOGY

파라미터

13B	41B
500B	700B
NVIDIA GeForce RTX 3090	-
2K(4K)	4K

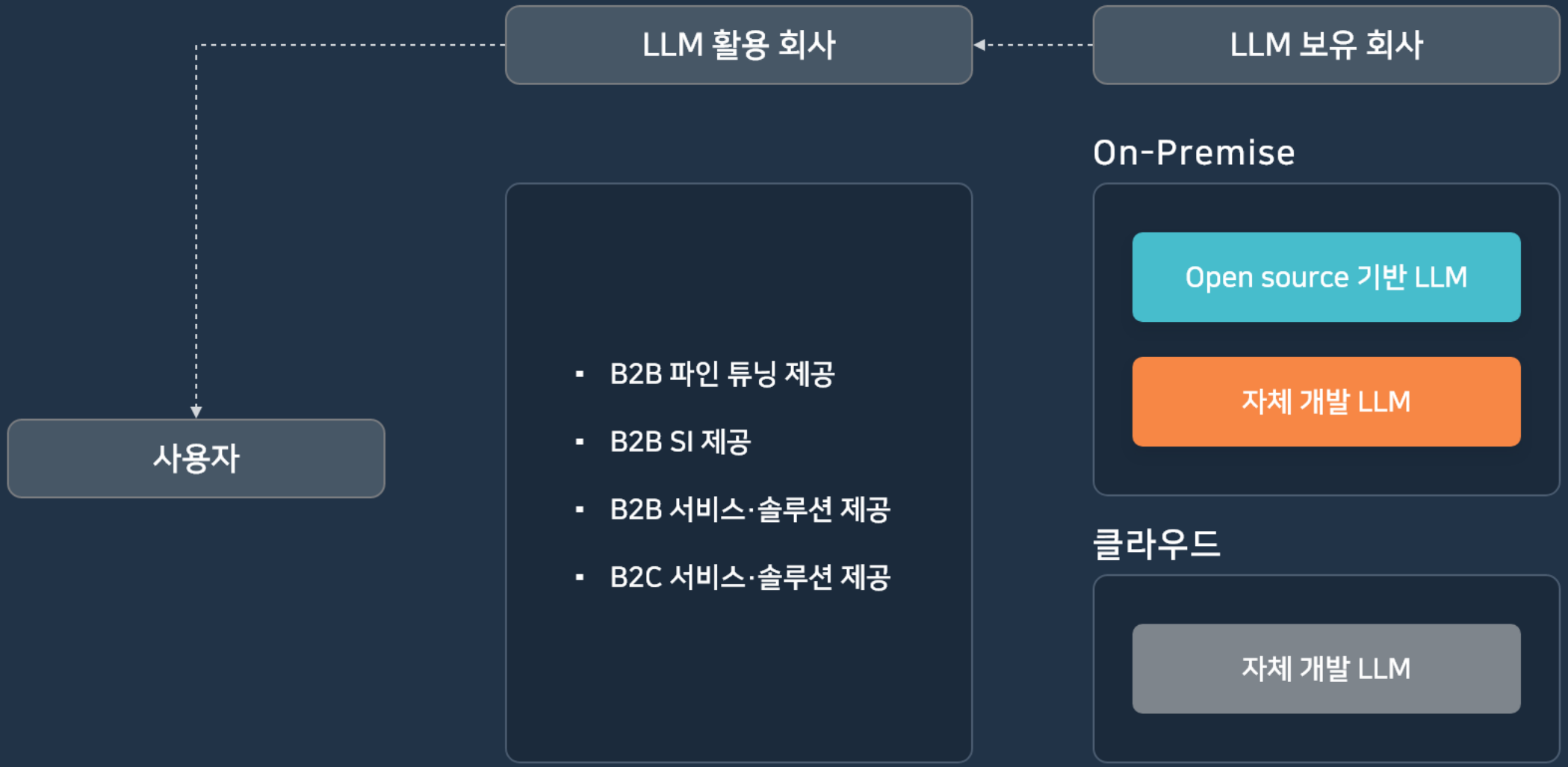
KONAN TECHNOLOGY

코난 LLM 사용 프로세스



KONAN TECHNOLOGY

생성형 AI 시장 참여자 분류



KONAN TECHNOLOGY

코난 LLM 접근 전략 - I

	문제점	해결 방안
Cloud LLM	▪ 회사 내부 데이터 외부 유출 우려 ▪ LLM 학습에 회사 내부 데이터가 사용될 수 있는 우려	▪ On-premise로 LLM을 제공해서 데이터 유출 우려 등을 원천 방지하자.
	▪ 학습 비용이 매우 많이 들고, 추론에 다수의 GPU 서버가 필요하므로 비용이 많이 든다. ▪ 따라서 공공이나 민간 기업에서 사용하기에는 매우 비싸다.	▪ 적절한 비용으로 최고의 성능 제공하자. ▪ 이를 위해 파라미터 크기는 작게, 학습 토큰은 많이 사용하자.
	▪ 모델의 한계로 인해 답변에 할루시네이션 발생 가능성 상존	▪ 벡터 검색 엔진을 활용하여 답변의 근거 등을 제시하여 할루시네이션을 최소화하자.

KONAN TECHNOLOGY

코난 LLM 접근 전략 - II

Konan LLM

On-premise

파라미터 크기는 작게, 학습 토큰은 많이

- 파라미터: 13.1B, 토큰: 492B (한국어: 284B)
- 파라미터: 41B, 토큰: 700B (한국어: 322B)

Konan Search 6 (벡터 검색 엔진) 활용

KONAN TECHNOLOGY

코난 LLM은?

제품 특성

B2B·B2G vs B2C

On-Premise(보안) vs Cloud

모든 도메인 vs 특화 도메인

문서 초안 생성, 문서 요약, 질의 응답

고객

SI회사

파인-튜닝 회사

B2B·B2G 고객사

KONAN TECHNOLOGY

코난 LLM 비즈니스 모델

KONAN TECHNOLOGY

Thank You

감사합니다.