



SAI-671B

6,710억 Parameter



SAI-70B

700억 Parameter



SAI-32B

320억 Parameter

Video Memory **768GB**
Computing Power **FP8 10560 TFLOPS**
Memory **1TB**
Throughput (Total tokens/s) **1400**
Concurrent Users (Simultaneous Access) **100**

Video Memory **192GB**
Computing Power **BF16 1320 TFLOPS**
Memory **512GB**
Throughput (Total tokens/s) **2100**
Concurrent Users (Simultaneous Access) **260**

Video Memory **96GB**
Computing Power **BF16 660 TFLOPS**
Memory **256GB**
Throughput (Total tokens/s) **2100**
Concurrent Users (Simultaneous Access) **260**

SYSTEM TO Ai On-Premises Server 기능별 활용 제안

고성능 챗봇

GPT-4 수준의
자연어 대화

문서 요약 및 생성

보고서, 이메일,
회의록 자동화

전문 Q&A (RAG)

사내 문서 기반
질의응답 시스템

코드 생성 및 분석

개발자 도우미,
코드 설명

음성 챗봇

Whisper + TTS 연동,
AI 스피커 구현

이미지 생성

Stable Diffusion
프롬프트 생성

LoRA 파인튜닝

특정 업무에 특화된
사내 모델 구축

완전 오프라인 지원

인터넷 없이
자체 AI 운용 가능

SYSTEM TO Ai On-Premises Server 업종 별 활용 예시



금융/보험

오프라인 챗봇,
문서 요약,
내부 규정 Q&A



공공/국방

내부망 AI 비서,
민원 자동화,
음성 보조



의료/제약

EMR 요약,
의료 Q&A,
환자 상담 자동화



제조/산업

매뉴얼 요약,
고장 진단,
음성 지시 시스템



통역/번역

실시간 다국어
통역 / 번역



SAI A.I Server 도입 과정



세계최대 GPU DePIN Cloud 에이서 Aethir 와
탈중앙화 DePIN 전용 데이터센터 NANODC의 제타큐브 공동 프로모션으로
'AI Ready' 된 강력한 서버를 DePIN 네트워크를 통해 저렴하고
편리하게 업무에 바로 적용 해 보세요!



deepseek

vLLM



Stable Diffusion



LangChain

Llama 3
From Meta

OpenAI
Whisper



Hugging Face



deepspeed