폐쇄망(Air-gap) 내에서 완성되는 LLM Ops: 기업 전용 올인원 온프레미스 플랫폼, 'Puteron AI'

Beyond Cloud Limits: The All-in-One Solution for Enterprise AI Sovereignty

Puteron AI Team, SLEXN

www.slexn.com | gate@slexn.com

SLEXN
IT Solution & Service Provider

클라우드 기반 AI 구독 모델의 구조적 한계

현재 대부분의 기업은 미국 빅테크가 주도하는 클라우드 API에 의존하고 있습니다. 그러나 이는 데이터 주권 상실, 비용 폭증 가능성, 그리고 범용 모델의 낮은 정확도와 커스텀 모델 불가라는 세 가지 구조적 제약이 있습니다.

🔐 보안 및 데이터 주권

PLM, CAD, 금융 기록, 고객 정보 등 민감한 기업 데이터가 퍼블릭 클라우드 API로 전송됩니다. 파인튜닝 학습 데이터 전송에 대한 규제적·보안적 제약이 날로 강화되고 있습니다.

💸 예측 불가능한 비용 구조

Token 기반 과금 모델은 추론 건수가 늘어날수록 OPEX(Operating Expense)가 기하급수적으로 증가합니다. Long-Context 처리 시 비용이 폭발적으로 상승하여 대규모 임직원 활용이 사실상 불가능합니다.

🤖 범용 모델의 한계

GPT-4, Claude 등 범용 모델은 기업 고유의 약어, 자체 비즈니스 로직, 특수 산업군에 대한 지식(방산 전술, 반도체 공정 등)을 내재화하지 못해 환각(Hallucination) 발생 가능성이 높습니다.

새로운 표준: Open Model 기반의 자율적 온프레미스 AI 인프라

트렌드가 바뀌고 있습니다. GLM과 같은 오픈 모델은 특정 영역에서 GPT-4 이상의 성능을 보이며, 기업 서버에 직접 설치·소유·통제가 가능한 시대가 열렸습니다.

🔄 From Closed to Open

GLM, Kimi, Mistral, Qwen 등 고성능 Open Weight 모델의 급격한 발전으로 폐쇄형 모델과의 성능 격차가 실질적으로 해소되었습니다.

🏛️ From API to Ownership

모델 가중치(Weights)를 직접 소유하고, 배포하며, 수정할 수 있는 완전한 권한을 기업이 확보합니다.

현대 AI의 3대 기술 요구사항

1 Fine-tuning (SFT)

기업 특화 데이터셋(Domain-Specific Data)으로 모델의 성격을 완전히 맞춤화합니다.

2 RAG (Retrieval-Augmented Generation)

내부 문서 베이스(Vector DB)와 실시간 연동하여 정확하고 신뢰할 수 있는 답변을 생성합니다.

3 Agentic Workflows

MCP(Model Context Protocol) 등을 활용하여 기존 ERP/CRM 시스템과 AI가 직접 상호작용하는 자동화 환경을 구현합니다.

온프레미스 LLM 구축의 높은 진입 장벽

왜 모든 기업이 당장 온프레미스로 넘어가지 않을까요? **너무 복잡하고 비싸기 때문**입니다. 기업은 기술 자체가 목적이 아니라 비즈니스 문제 해결이 목적입니다. 이 모든 복잡함을 제거해야 합니다.

🧱 하드웨어 병목 (Memory Wall)

LLM 추론 속도는 GPU의 VRAM(HBM) 용량에 좌우됩니다. 대규모 파라미터 모델 운영을 위해 H100/A100 GPU 군이 필수적이어서 초기 투자비가 천문학적입니다. GPU·NVLink·Storage 간 병목 최적화 설계 난이도 또한 매우 높습니다.

🧩 소프트웨어 파편화 (Toolchain Chaos)

모델 서빙(vLLM, TGI), 학습(DeepSpeed, FSDP), 벡터 DB(Milvus, Weaviate), 프롬프트 관리(LangChain) 등 수십 가지 도구를 직접 통합해야 합니다. 유지보수 또한 어려운 도전 입니다.

👩‍💻 인재 부족 (Talent Shortage)

모델 배포 후 성능 모니터링, 재학습 트리거, 버전 관리 등 LLM Ops 프로세스를 구축·운영할 전문 인력 채용이 극히 어렵습니다. 이 모든 스택을 이해하는 인재는 구하기 어렵습니다.

Puteron AI: 기업 AI 자립을 위한 완벽한 올인원 솔루션

SLEXN의 Puteron AI는 고객이 직접 복잡한 퍼즐을 맞추게 하지 않습니다. **이미 완성된 그림**을 제공합니다. Plug & Play 방식으로 전원 연결 후 즉시 사용 가능한 사전 검증(Pre-validated) AI 어플라이언스입니다.

① Puteron AI Server

SSD 기반 KV Cache Offloading으로 GPU 메모리 한계를 극복하고 비용 효율을 극대화합니다.

② Puteron AI Model Hub

통합 LLM Ops 플랫폼. 모델 관리, 파인튜닝, RAG 구축, Agent 개발 및 마켓플레이스 기능을 **One-Stop**으로 제공합니다.

🔒 Security

Air-gap/폐쇄망 100% 지원

⚡ Efficiency

GPU 투자 비용 대비 성능 극대화 (High ROI)

🏛️ Sovereignty

기업 데이터·모델에 대한
완전한 소유권 및 통제권

SLEXN
IT Solution & Service Provider

Puteron AI Server: GPU 메모리 벽을 허무는 차세대 인프라 솔루션

LLM을 운영할 때 가장 큰 병목은 GPU의 HBM 용량입니다. Puteron AI Server는 **SSD 기반 KV Cache Offloading** 기술로 이 문제를 근본적으로 해결합니다. SSD 비용으로 추가적으로 필요한 GPU 성능을 뽑아내는 혁신입니다.

작동 원리 (Mechanism)

- → GPU VRAM 내 KV Cache 데이터를 **실시간 모니터링**
- → Inactive/Low-priority Page 데이터를 **고속 NVMe SSD로 페이징 아웃 (Paging Out)**
- → 필요 시 NAND 플래시에서 마치 DRAM처럼 **Paging In 수행**
- → AI 트래픽 패턴 최적화 커스텀 알고리즘으로 **지연 시간(Latency) 최소화**

핵심 혜택 (Key Benefits)

💰 비용 효율성

1TB SSD 비용으로 수천만 원 상당의 GPU VRAM 효과. TCO 획기적 절감.

📦 용량 확장

70B+ 파라미터 모델도 소수의 GPU로 구동 가능. 물리적 한계를 넘은 메모리 풀 생성.

🔄 Hot-Swap

서버 재부팅 없이 다양한 모델 가중치를 GPU 메모리로 즉시 로드. 멀티 모델 운영 지원.

GPU 메모리 한계를 극복하는 SSD기반의 KV Cache 아키텍처

LLM의 가장 큰 적은 '메모리'입니다. Puteron AI Server 는 **소프트웨어 정의 메모리(SDM)** 개념을 도입하여 GPU·DRAM·SSD를 하나의 계층적 메모리 풀로 통합합니다.

문제 진단: KV Cache 폭발

Transformer 기반 LLM은 추론 시 이전 토큰의 Key, Value 쌍을 메모리에 보관해야 합니다. 시퀀스 길이 **N**이 증가하면 메모리 요구량은 **N^2에 비례**하여 급증합니다. 80GB HBM GPU도 대용량 모델과 긴 컨텍스트 처리에는 여전히 부족합니다.

HBM 비용 효율성 저하로 인해 기존 GPU-only 방식은 대규모 배포에 구조적 한계를 지닙니다.

Puteron AI Server 솔루션 아키텍처

01

소프트웨어 정의 메모리(SDM)

GPU VRAM, System DRAM, NVMe SSD를 계층적인 하나의 메모리 풀로 추상화하여 운영합니다.

02

유닛 오프로딩(Unit Offloading)

Attention Head 단위 혹은 Tensor Block 단위로 활성화 데이터와 KV Cache를 실시간 판별하여 SSD로 스왑합니다.

03

PCIe Gen 5 지원

최대 32GT/s의 대역폭을 제공하는 PCIe Gen 5 인터페이스로 CPU-GPU-SSD 간 데이터 병목을 최소화합니다.

기존 OS Swap과는 다른 AI 최적화 지능형 페이징

단순히 SSD를 메모리처럼 쓴다고 해서 다 되는 것이 아닙니다. Puteron AI Server의 차이점은 **'지능'**에 있습니다. AI가 다음에 필요한 데이터를 미리 예측하고 비동기 처리로 GPU 연산이 멈추지 않도록 설계되었습니다.

Non-Blocking I/O

비동기 I/O(Asynchronous I/O)와 Pre-fetching 기술로 GPU 연산 파이프라인이 데이터 이동 중에도 멈추지 않도록 설계되었습니다. 기존 OS Swap 방식의 Latency Spike를 완전히 제거합니다.

Hot/Cold 데이터 분류

모델의 Attention Map을 분석하여 즉시 필요한 데이터(Hot)는 GPU VRAM에, 당장 불필요한 데이터(Cold)는 SSD Tier로 자동 이동시키는 지능형 알고리즘을 탑재했습니다.

Compression 기술

SSD로 이동하는 KV Cache 데이터를 압축(Compression)하여 저장 공간 효율을 높이고 전송 시간을 단축합니다. 배치 크기를 최대 2~4배 확장하여 처리 처리량(Throughput)이 향상됩니다.

4x

배치 크기 확장

동일 GPU 메모리에서 최대 4배 큰 배치 처리 가능

↑99%

Cache Hit Ratio

스마트 예측 알고리즘으로 SSD 접근 지연을 효과적으로 은닉

TCO 혁신: GPU 투자 비용 절감과 스토리지 혁신

기술의 최종 목적은 비용 절감입니다. HBM과 NVMe SSD의 가격 차이를 전략적으로 활용하여, Puteron AI Server 도입 시 **총 소유 비용(TCO)을 40% 이상 절감**하면서도 더 큰 모델을 더 많은 사용자에게 서비스할 수 있습니다.

경제적 분석: GPU vs NAND 가격 비교

고성능 HBM(HBM3)은 GB당 약 **$20~$30**인 반면, Enterprise NVMe SSD는 GB당 **$0.10~$0.20** 수준으로 약 **100배 이상의 가격 경쟁력**을 제공합니다.

> 예: 8장의 GPU가 필요한 작업을 2장의 GPU + 대용량SSD로 구현 가능하여 CapEx(Capital expenditures)를 획기적으로 절감합니다.

Puteron AI Server 사양

Storage

PCIe Gen 5 x4, 최대 8TB 단일 드라이브, 다중 드라이브 RAID 구성 지원

Compatibility

NVIDIA H100/A100, AMD MI 시리즈 등 주요 데이터센터 GPU 완벽 지원

최적 워크로드

대규모 문서 요약, 긴 컨텍스트 채팅, RAG 검색 증강 등 메모리 집약 워크로드

Puteron AI Model Hub: 기업 내부의 완벽한 허깅페이스 클론

온프레미스 환경에서 구동되는 통합 LLM 관리·운영 플랫폼입니다. 직관적인 GUI(Web Dashboard)를 통해 복잡한 CLI 설정 없이 LLM 전 생애 주기(Lifecycle)를 관리합니다.

Model Repository

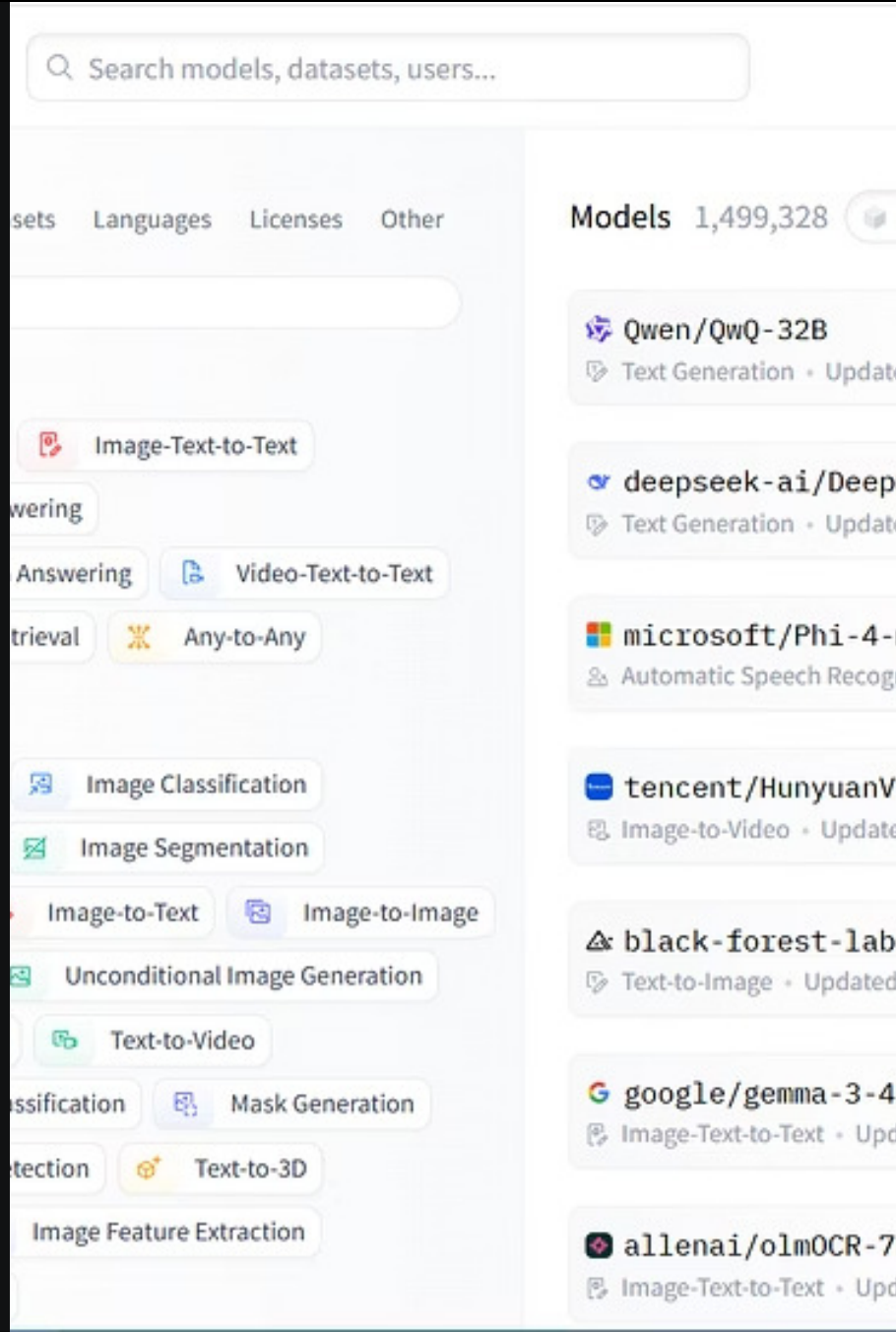
GLM, Kimi, Mistral, Qwen 등 주요 Open Weight 모델을 내장 카탈로그에서 **원클릭 다운로드**. Fine-tuning된 커스텀 모델 체크포인트 자동 관리 및 A/B 테스트 지원.

Visual Serving

모델 배포 시 인스턴스 수, GPU 할당 전략(Tensor/Pipeline Parallelism)을 GUI에서 설정. 코드 한 줄 없이 버튼 클릭만으로 모델이 GPU에 로드되고 API 서버가 즉시 시작됩니다.

OpenAI API 호환

OpenAI API 포맷과 100% 호환되는 엔드포인트를 자동 생성하여, 기존 클라이언트 애플리케이션 **수정 없이 즉시 연결** 가능합니다.



SLEXN
IT Solution & Service Provider

Model Lifecycle Management: 배포부터 버전 관리까지

Model Hub는 단순한 모델 창고가 아닙니다. 완벽한 엔터프라이즈급 운영 시스템입니다. 개발자가 코드를 짜지 않아도 배포부터 안전한 버전 전환까지 모두 GUI에서 처리됩니다.

통합 모델 레지스트리

HuggingFace, PyTorch, ONNX, GGUF 등 다양한 포맷의 모델을 등록하고 메타데이터(Parameters, License, 생성일)를 자동 수집합니다.

단계별 모델 승격(Promotion) 및 격리 관리로 안전한 릴리즈 파이프라인을 구현합니다.

A/B 테스트 및 Canary 배포

트래픽의 일부(예: 10%)를 새로운 모델 버전으로 우회시켜 성능을 검증한 후, 전체 트래픽을 전환하는 **안전한 배포 전략**을 지원합니다.

고성능 추론 서빙

Dynamic Batching으로 들어오는 요청을 실시간으로 모아서 GPU 활용률을 극대화합니다. Tensor Parallelism & Pipeline Parallelism으로 여러 GPU에 거대 모델을 분산 배치하여 추론 속도를 획기적으로 단축합니다.

Visual Fine-Tuning Studio: Code Free - 모델 튜닝 자동화

기존에는 모델을 미세 조정하려면 Python 코드를 짜고 터미널에서 훈련시켜야 했습니다. Puteron AI Model Hub는 이를 **No-Code/Low-Code** 환경으로 완전히 전환했습니다.

PEFT 지원 (LoRA / QLoRA)

전체 모델 재학습 대신, 적은 파라미터만 업데이트하는 **LoRA(Low-Rank Adaptation)**와 QLoRA 방식을 GUI에서 지원합니다. 단일 GPU에서도 70B 규모의 대형 모델 튜닝이 가능하여 GPU 비용을 획기적으로 절감합니다.

데이터셋 관리 및 Validation

JSONL, CSV 형식의 학습 데이터를 업로드하면 시스템이 자동으로 형식을 검증합니다. 학습 전 데이터 분포를 시각화하여 불균형을 사전에 점검하고, 데이터 품질 문제로 인한 학습 실패를 방지합니다.

실시간 학습 모니터링

Real-time Loss Curve, Learning Rate Scheduler, GPU Utilization을 대시보드에서 실시간 확인합니다. Checkpoint 중단과 Resume 기능으로 실패 비용을 최소화하여 안전한
학습 환경을 제공합니다.

LLM Ops의 완성: 학습, RAG, 그리고 Agent 생태계 구축

Puteron AI Model Hub의 진정한 힘은 모델 학습부터 RAG, 그리고 Agent 제작까지 **하나의 플랫폼**에서 모두 해결된다는 점입니다. 내부 AI 생태계를 완성하는 세 가지 핵심 기능입니다.

1️⃣ Full Life Cycle 관리

- 이미지·텍스트·비정형 데이터 **전처리 및 라벨링** 도구 내장
- SFT(Supervised Fine-Tuning): Learning Rate, Batch Size, Epoch 등 하이퍼파라미터 GUI 조절
- LoRA/QLoRA PEFT 지원으로 적은 리소스에 고품질 튜닝
- 정답률·유창성·안전성(Safety) **자동 평가 지표** 제공

2️⃣ Enterprise RAG

- PDF, Word, Excel, 웹페이지 업로드 및 **자동 청킹(Chunking)**
- 벡터 DB 연동, 임베딩(Embedding) 저장
- **Hybrid Search**: 키워드 + 시맨틱 검색 결합으로 Hallucination 방지

3️⃣ Agent Studio

- 시스템 프롬프트 설계 및 Few-shot 예제 관리
- 회사 내부 API(ERP, CRM 등) 등록 → **AI 자율 도구 호출**
- 완성된 Agent를 **사내 마켓플레이스**에 등록하여 전 부서 공유·구독

Advanced Enterprise RAG: Hallucination 방지와 정확도 극대화

🎯 Re-ranking으로 최종 정확도 제고

검색된 문서 후보군(상위 20개)에 **Reranker 모델(Cross-Encoder)**을 적용하여 질문과 가장 관련성 높은 상위 5개를 최종 선정합니다. 이 과정을 통해 Hallucination 확률을 획기적으로 줄입니다.

🔍 지능형 데이터 전처리

고정 크기 분할이 아닌 **의미 단위(Semantic Chunking)**로 문서를 분할하여 문맥 파괴를 최소화합니다. PDF의 표(Table)와 이미지 내 텍스트를 정확히 추출하는 OCR 기능이 통합되어 있습니다.

단순히 벡터 DB에 문서를 넣는 것만으로는 부족합니다. Puteron AI Model Hub는 **3단계 고급 RAG 파이프라인**으로 AI 응답의 신뢰성을 획기적으로 높입니다.

Agent Studio: 자율적인 AI 워커 생성 및 생태계 구축

이제 챗봇은 질문만 받는 게 아닙니다. **실제로 일을 해야 합니다.** Agent Studio는 Function Calling과 Multi-Agent 협업을 통해 회사의 Engineering(PLM,CAD,ALM) / Business Platform (ERP, CRM) 등과 직접 연결되는 자율 AI 워커를 GUI로 설계합니다.

Function Calling (도구 호출)

OpenAPI(Swagger) 스펙을 업로드하면 자동으로 API 스키마를 파싱합니다. LLM이 자연어 요청을 분석하여 SAP ERP 조회, 이메일 발송 등 **적절한 API를 직접 호출**하고 결과를 종합합니다.

멀티 에이전트 협업

Planner-Agent가 복잡한 작업을 하위 작업으로 분해하면, **Worker-Agent**가 각각 수행하고, **Evaluator-Agent**가 결과를 검증합니다. 이 협업 과정을 시각적 워크플로우로 설계합니다.

내부 마켓플레이스

생성된 Agent를 '앱'처럼 패키징하여 사내 카탈로그에 등록합니다. 다른 부서 사용자가 **구독(Subscribe)**하여 개발 없이 자신의 대시보드에서 즉시 활용할 수 있습니다.

Puteron AI-최적화된 통합 스택(Hardware + Software)

시중에서 GPU 서버를 사고 LLM 오픈소스를 따로 설치하면 스토리지 I/O와 GPU 메모리 간 튜닝이 최적화되지 않아 성능이 100% 발휘되지 않습니다. **Puteron AI는 다릅니다.**

Native Integration (네이티브 통합)

Software-Aware Storage: Model Hub(소프트웨어)는 SSD 캐싱 레이어(하드웨어)를 완벽히 인식하여 최적화된 데이터 배치 전략을 수행합니다. 소프트웨어가 하드웨어에 직접 명령을 내리는 수준의 최적화입니다.

Zero Compatibility Issues: OEM 단계에서 결합된 사전 검증(Pre-validated) 솔루션으로 드라이버 충돌, 버전 불일치 문제를 완전히 제거합니다.

Air-gap Support: 모든 컴포넌트(모델 다운로더, 벡터 DB, 라이브러리)가 오프라인 환경 설치를 완벽히 지원합니다.

성능 지표 (Performance Metrics)

40~60%

TCO 절감

기존 GPU-only 대비 총 소유 비용 절감 (SSD 확장 활용 시)

80%

구축 기간 단축

기존 DIY 구축 대비 인프라 구축 기간 단축 (All-in-One Appliance)

전원 켜기만으로 최고 성능의 AI 인프라를 즉시 확보할 수 있습니다. 이것이 SLEXN이 제안하는 진정한 올인원 솔루션입니다.

SLEXN
IT Solution & Service Provider

방산/군사

완벽한 폐쇄망(Air-gap)에서 실시간 전략을 수행 — 외부 인터넷이 물리적으로 차단된 지하 벙커형 C4ISR 작전 통제실

⚠️ 임계 임무 요구 사항

데이터 처리량

초당 15GB의 고해상도 영상 스트림 + 5,000건/분 암호 전문 실시간 분석

제약 조건

100% Air-gap. 외부 클라우드 연결 불가. 보안 레벨 '극비(Top Secret)'

성능 목표

레이턴시(Latency) 200ms 이하의 실시간 탐지 및 전술 분석

✅ Puteron AI 활용

Puteron AI Server: SSD KV 캐싱으로 20TB 규모 영상 데이터를 가상 메모리 풀에 로드. **4대의 GPU만으로 기존 32대 GPU 서버의 처리량 달성.**

Puteron AI Model Hub: 과거 30년 작전 보고서로 RAG 학습된 '전술 분석 LLM' 구동. "북동쪽 구릉지대 이상 열원 감지"라는 음성 명령 입력 시, AI가 즉시 유사 패턴을 검색하여 전략을 제시합니다.

🏆 결과: 적의 은밀한 이동을 발견하여 30분 만에 대응 전술 완료. 데이터 한 줄도 외부로 유출 없음.

자동차/R&D

차세대 전기차의 '열 폭주(Thermal Runaway)' 원인을 500TB 데이터에서 24시간 내 발견

⚠️ 프로젝트 위기 상황

분석 데이터량

500TB 시계열 로그 데이터, BMS 센서 데이터 10억 개 라인 이상

보안 이슈

BMS 소스 코드 및 셀 화학 조성 비율 — 유출 시 경쟁사 기술 모방 위험

시간 제약

원인 규명에 최소 2주 예상 → **24시간 내 해결 필요**

✅ Puteron AI 기술적 해결

Puteron AI Server: 고속 SSD에 벡터 인덱스를 구축, 필요 시 GPU로 즉시 로드하여 **검색 속도 100ms 달성.** 기존 방식 대비 수십억 원의 추가 비용 절감.

Puteron AI Model Hub: 연구원이 "영하 20도 충전 시 전압 강하가 급격한 구간을 찾아줘"라고 자연어 질의. 이 분석 로직을 **'배터리 안전 진단 Agent'**로 저장하여 개발팀·품질팀·생산팀 모두 내부 AI 마켓플레이스에서 즉시 활용.

🏆 결과: 특정 배치(Batch)의 센서 캘리브레이션 오류를 단 4시간 만에 특정. 신차 출시 일정 준수.

AI가 만드는 AI: 확장 가능한 자율 생태계

Puteron AI의 진정한 가치는 이 모든 시스템이 **'스스로 성장'**한다는 점입니다. 첫 번째 사용자가 만든 뛰어난 에이전트가 회사의 자산이 되어 전체 조직으로 전파되는 복리(Compound) 성장 구조를 만듭니다.

Phase 1: Infrastructure

Puteron AI Server 도입으로 안전한 HW 기반 마련

Phase 2: Model & Knowledge

Puteron AI Model Hub를 통해 각 부서의 노하우를 LLM으로 커스터마이징

Phase 3: Agent Marketplace

영업팀의 '고객 응대 Agent', 인사팀의 '채용 screening Agent', 법무팀의 '계약서 검토 Agent'가 '회사 내부 앱 스토어'에 등록되어 전 직원이 구독·활용

AI의 민주화 (Democratization of AI)

모든 직원이 AI 사용자가 아닌 **AI 창작자(Creator)**로 변신합니다. 기술 부서만의 전유물이 아닌, 전사적 가치 창출 도구가 됩니다.

복리 성장 (Compound Growth)

하나의 Agent가 만든 데이터가 다른 Agent의 학습 소스가 되며 기업의 지능이 **기하급수적으로 성장**하는 복리 효과가 창출됩니다.

클라우드를 넘어, 여러분의 주권으로

Beyond Cloud, To Your Sovereignty

🔒 Unrivaled Security

Air-gap/폐쇄망 환경에서 완벽하게 작동하는 **국내 유일의 올인원 솔루션**. 데이터가 회사 밖으로 한 줄도 나가지 않습니다.

⚡ Revolutionary Efficiency

SSD 기반 KV Cache 기술로 **GPU 비용 40~60% 절감**과 AI 메모리 확장성을 동시에 달성합니다.

🌐 Complete Ecosystem

Model Hub로 모델 튜닝, RAG, Agent 개발 및 배포까지 **One-Stop 해결**. 개발 역량 없이도 즉시 운영 가능합니다.

SLEXN의 제안

더 이상 클라우드 구독료에 지출하고, 데이터 유출을 걱정하지 마십시오. 여러분의 데이터와 노하우가 머무는 곳에, 여러분의 AI 인프라를 구축하십시오. **SLEXN이 여러분의 'AI 자립'을 위한 든든한 파트너가 되겠습니다.**

SLEXN
IT Solution & Service Provider