



클라우드 배포 및 산업화 기술

성명 박종현

소속 래블업 주식회사

SUBJECT

인공지능 기술의 대중화 (AI Democratization)를 위한
TANGO 커뮤니티 3회 컨퍼런스

주관



주최

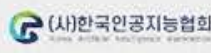


과학기술정보통신부



정보통신기획평가원

후원



목 차

1

산업 현황과 과제

03

1. LLM 모델
2. 인프라 및 전력량
3. 엔터프라이즈 시장
4. 과제 - S/W 플랫폼 대응 영역

2

TANGO 클라우드 배포 기술 개발 성과

07

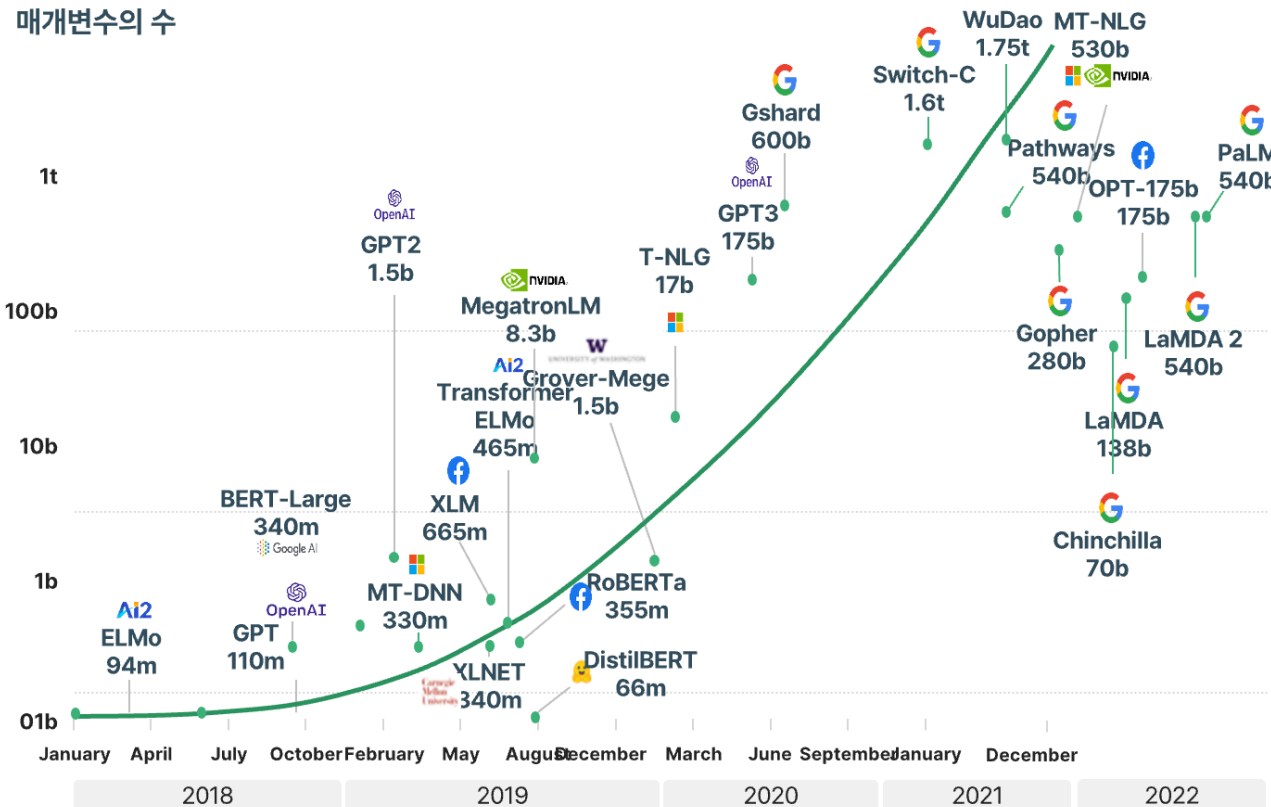
1. TANGO 클라우드 배포 기술 개요
2. AWS 클라우드 배포 모듈
3. kt cloud 배포 모듈
4. TANGO 서비스의 AWS 배포
5. LLM 모델 구동 및 응용 사례

3

향후 개발 계획

16

LLM 모델



- 파라미터 규모의 급속한 증가
 - GPT: 2 (1.5B) / 3 (175B) / 4 (~1.8T?)
 - Claude 3.5 Sonnet (>175B?)
 - LLaMA: 2 (70B) / 3 (405B)
 - Bloom: 176B
- 규모 증가와 다른 방향의 접근법
 - Mixture of Experts (MoE): 맥락에 따라 특화된 전문가 모델만 선택적으로 활성화(Mixtral 7Bx8, GPT4 등)
 - Chain of Thought (CoT): 단계별 추론 능력 강화(OpenAI o1-preview)

Gadi Singer 2021 / Google 2022

인프라 및 전력량

- 증가하는 연산 자원 수요
 - 고성능 모델 학습에 대한 요구
 - 늘어가는 추론 수요
 - 학습용/추론용 GPU가 더 이상 구분되지 않음

- 전력 요구량 문제
 - 연산 요구량 대응을 위한 인프라
 - 전세계 데이터센터 전력 소비량
 - ✓ 460TWh: 2022년
 - ✓ 1,050TWh: 2026년(예상)
 - ✓ 대한민국 `24 총 전력 사용량 650TWh

Revolutionary Performance Backed by Evolutionary Innovation

NVIDIA DGX™ B200 is an unified AI platform for develop-to-deploy pipelines for businesses of any size at any stage in their AI journey. Equipped with eight NVIDIA Blackwell GPUs interconnected with fifth-generation NVIDIA® NVLink®, DGX B200 delivers leading-edge performance, offering 3X the training performance and 15X the inference performance of previous generations. Leveraging the NVIDIA Blackwell GPU architecture, DGX B200 can handle diverse workloads—

에너지 효율적인 인프라

수냉식 GB200 NVL72 랙은 데이터센터의 탄소 발자국과 에너지 소비를 줄여줍니다. 수냉식 냉각은 컴퓨팅 밀도를 높이고, 상면 사용 공간을 줄이며, 대규모 NVLink 도메인 아키텍처와의 고대역폭, 저지연 GPU 통신을 용이하게 합니다. NVIDIA H100 공랭식 인프라에 비해 GB200은 동일한 전력으로 25배 더 높은 성능을 제공하면서 물 소비는 줄입니다.



Gadi Singer 2021 / Google 2022

엔터프라이즈 시장

- 모델 학습 전략
 - 대부분의 경우, 파운데이션 모델 자체 개발 대신 공개 모델의 활용 중심
 - 모델 미세조정(fine-tuning) 또는 RAG 도입 확대
 - ✓ 비교적 적은 비용과 노력으로 사내 규정이나 절차 등의 학습을 통한 업무 효율화 가능
- 자체 튜닝 모델 및 공개 모델을 쉽게 사내외에 서비스 하길 원하는 수요
- 자체 인프라 확충 패턴
 - 격리망이어서 외부 서비스 이용이 불가능한 경우
 - 외부 AI 모델에 대한 API 비용 절감 목적(B2C 서비스를 위해 대규모로 호출하는 경우 만만치 않은 API 비용)
 - 학습은 온프레미스 환경에서 진행하고 서비스는 클라우드 인프라를 이용하는 하이브리드 형태의 운영
- 가속 하드웨어
 - 여전히 압도적인 NVIDIA GPU 도입
 - 대안 탐색의 움직임: 비용 절감 및 수급 등의 문제로 AI 전용 가속칩 PoC를 진행하는 경우가 종종 있음
 - 엣지 장치 추론 시장 성장이 예상되는 바, 비NVIDIA 가속 장치의 새로운 시장 기회 창출 가능성

과제 - S/W 플랫폼 대응 영역

- 기존에 강조되던 학습 시장보다 AI 모델 서비스 및 운영 지원 시장이 더 확대될 가능성
 - 파인튜닝 및 RAG를 학습을 용이하게 지원하는 기술
 - 연속/반복 학습 지원을 위한 파이프라인 기술
- 안정적인 서빙 시스템
 - 서비스는 학습보다 서비스 안정성에 대한 민감도가 훨씬 높음
 - 모델 서비스 고가용성 또는 replica 지원(클라우드의 경우 이 부분 대응이 훨씬 유리)
- 인프라 유연성 확보
 - 하이브리드 클라우드 사용 패턴에 대응하기 위한 멀티 클라우드 연동 지원
 - 향후 비용 절감 및 수급 문제 대응을 다양한 AI 전용 가속칩에 대한 지원

2. TANGO 클라우드 배포 기술 개발 성과

7

TANGO 클라우드 배포 기술 개요

```
TARGET_CLAS_MAP = {
    "docker": LocalDocker,
    "gcp-cloudrun": CloudRun,
}

class CloudTargetBase(ABC):
    def __init__(self, user_id: str, project_id: str):
        self.user_id = user_id
        self.project_id = project_id

    @abstractmethod
    async def start_service(self):
        pass

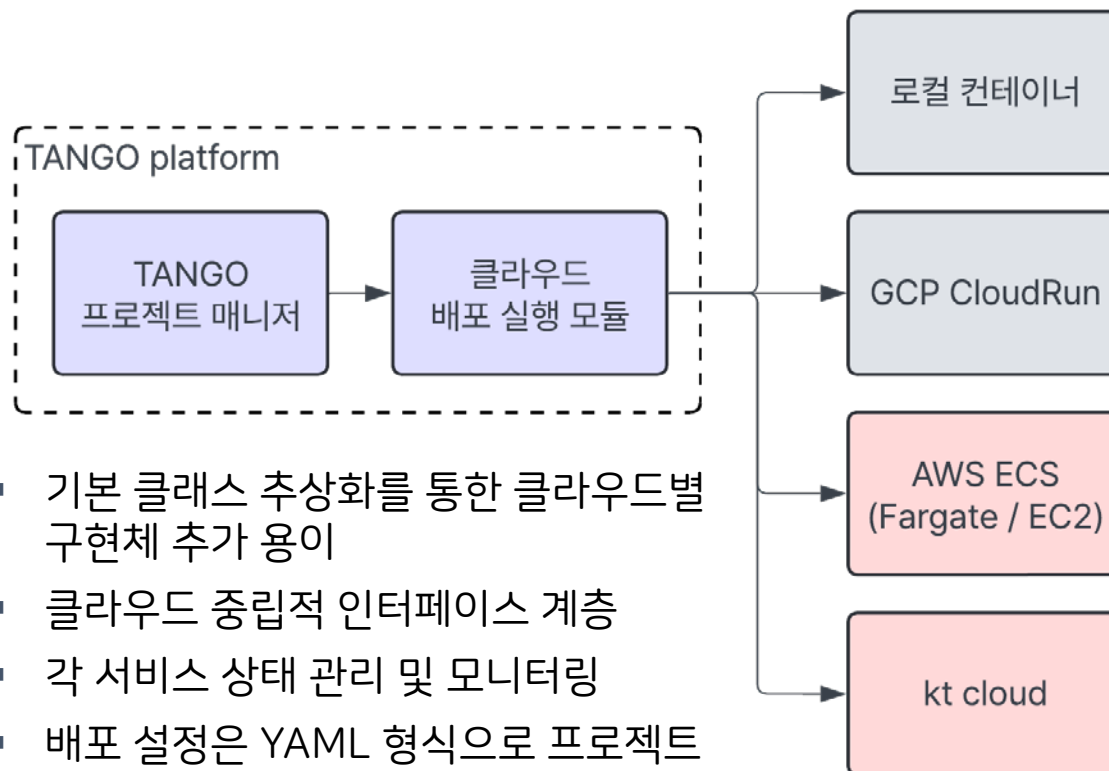
    @abstractmethod
    async def stop_service(self):
        pass

    @abstractmethod
    async def get_service_status(self):
        pass

    @abstractmethod
    async def _build_image(url: URL, data: dict):
        pass

    @abstractmethod
    async def push_image(self):
        pass

class CloudRun(CloudTargetBase):
    # Implements details
```



- 기본 클래스 추상화를 통한 클라우드별 구현체 추가 용이
- 클라우드 중립적 인터페이스 계층
- 각 서비스 상태 관리 및 모니터링
- 배포 설정은 YAML 형식으로 프로젝트 매니저가 전달

AWS 클라우드 배포 모듈

- AWS ECS (Elastic Container Service) 기반 배포
 - <https://aws.amazon.com/ecs/>
 - 컨테이너화된 애플리케이션을 배포 및 관리하는 매니지드 서비스
 - 이미지로 패키징 된 서비스 배포를 편리하고 안정적으로 할 수 있음
- 인프라가 중심이 아닌 서비스 중심의 배포
 - ECS cluster 생성
 - ✓ Task 및 service의 논리적 단위 그룹
 - ✓ 하나 이상의 task 및 service를 포함할 수 있음
 - Task definition 생성
 - ✓ 컨테이너가 어떻게 실행되어야 하는지 지정하는 일종의 템플릿
 - ✓ 이미지 이름, 리소스 할당 등과 같은 단일 앱(컨테이너) 구동 스펙을 정의
 - Service 생성
 - ✓ Task definition을 실제 개체화 하여 구동. 동일 인스턴스를 여러 개 띄워 LB 또는 HA 가능

AWS 클라우드 배포 모듈

Request: /start (param: user_id, project_id)
Deploy spec: deployment.yaml



```

cloud_deploy-1 2024-11-27 01:29:42,627 - root - INFO - service.py launch_service - Let's start start server
cloud_deploy-1 2024-11-27 01:29:42,628 - cloud_manager.targets.aws.ecs - INFO - Starting service for nn-model
cloud_deploy-1 2024-11-27 01:29:42,628 - cloud_manager.targets.aws.ecs - INFO - get cluster list
cloud_deploy-1 2024-11-27 01:29:42,757 - root - INFO - list_cluster: ['TangoCluster', 'nn-model_1_7', 'TangoFullCluster', 'nn-model_1_2', 'yolov7-e6e-1_1']
cloud_deploy-1 2024-11-27 01:29:42,757 - cloud_manager.targets.aws.ecs - INFO - Ensuring cluster nn-model exists
cloud_deploy-1 2024-11-27 01:29:42,799 - cloud_manager.targets.aws.ecs - INFO - Cluster nn-model_1_2 already exists
cloud_deploy-1 2024-11-27 01:29:42,799 - cloud_manager.targets.aws.ecs - INFO - Registering task definition for nn-model
cloud_deploy-1 2024-11-27 01:29:42,799 - cloud_manager.targets.aws.ecs - INFO - Creating task definition
cloud_deploy-1 2024-11-27 01:29:42,862 - cloud_manager.targets.aws.ecs - INFO - Task definition registered successfully
cloud_deploy-1 2024-11-27 01:29:42,862 - cloud_manager.targets.aws.ecs - INFO - Creating ECS service for nn-model
cloud_deploy-1 2024-11-27 01:29:42,862 - root - INFO - deploy yaml: build=Build(architecture='x86', accelerator='cuda', os='ubuntu', image_uri='public.ecr.aws/1318d1d9/tango/nn-model:latest', components=Components(engine='pytorch', libs=['python=3.8', 'torch=1.10'], custom_packages=CustomPackages(ap=[vim, libgl-mesa-glx], pypi=[cython, numpy<17, imutils, flask, opencv-python, opencv-contrib-python, imageio, pyyaml, matplotlib, pandas, tqdm, seaborn, requests, werkzeug, torch, torchvision, python-math, albumentations, pathlib])), deploy=Deploy(type=aws-ecs, work_dir='/workspace', pre_exec=None, entrypoint=['python', '/output.py'], resources=Resources(cpu=8, memory=16384, gpu=None), network=Network(service_host_ip='2.3.4.5', service_host_port=2222, service_container_port=2222), service_name='nn-model', execution_role_arn='arn:aws:iam::502530743065:role/ec2TaskExecutionRole', launch_type='FARGATE', aws_vpc={'assign_publicip': 'ENABLED', 'subnets': ['subnet-0e1c240913e7c488c'], 'security_groups': ['sg-0458daea1d78303fa']})
cloud_deploy-1 2024-11-27 01:29:43,345 - cloud_manager.targets.aws.ecs - INFO - ECS service nn-model created successfully
  
```

암호화

관리형 스토리지

Fargate 임시 스토리지

서비스

태스크

인프라

지표

예약된 태스크

태그

태스크 (1)

원하는 상태 필터링

시작 유형 필터링

속성 또는 값으로 태스크 필터링

모든 원하는 상태

모든 시작 유형

<input type="checkbox"/>	태스크	마지막 상태	원하는 상태	태스크 정의	상태	시작 위치	컨테이너 인스턴스	시작 유형	플랫폼 버전
<input type="checkbox"/>	e396bfc2d...	실행 중	실행 중	nn-model:50	알 수 없음	13시간 전	-	FARGATE	1.4.0

AWS 클라우드 배포 모듈

로그 (13+) 정보

아래 필터 막대를 사용하여 로그 이벤트에서 용어, 구분 또는 값을 검색하고 일치시킬 수 있습니다. [필터링 패턴에 대해 더 알아보기](#)

타임스탬프(UTC+09:00)	메시지
2024년 11월 20일 16:55 (UTC+9:00)	47.115.231.124 - - [20/Nov/2024 07:55:43] code 400, message Bad request syntax ('hi\x00')
2024년 11월 20일 16:55 (UTC+9:00)	47.115.231.124 - - [20/Nov/2024 07:55:43] "[35m [1mhi\x00 [0m" HTTPStatus.BAD_REQUEST -
2024년 11월 20일 15:57 (UTC+9:00)	[31m [1mWARNING: This is a development server. Do not use it in a production deployment. Us
2024년 11월 20일 15:57 (UTC+9:00)	* Running on all addresses (0.0.0.0)
2024년 11월 20일 15:57 (UTC+9:00)	* Running on http://127.0.0.1:2222
2024년 11월 20일 15:57 (UTC+9:00)	* Running on http://10.80.5.137:2222
2024년 11월 20일 15:57 (UTC+9:00)	[33mPress CTRL+C to quit [0m
2024년 11월 20일 15:57 (UTC+9:00)	Detection Service

Detection Service

파일 선택 선택된 파일 없음

제출

Selected File : 150101_2015_6_.mp4

File Start/Stop Camera Start/Stop



kt cloud 배포 모듈

- 국산 클라우드 연동 및 사용성 강화를 위한 kt cloud 배포
 - <https://cloud.kt.com/product/productDetail?prodId=P000000005>
- 원격 인스턴스 Docker 엔진과의 상호작용 모듈 개발을 통한 컨테이너 제어 및 배포

```

2024-11-14 11:24:34,189 - root - INFO - request /kvc/stop
2024-11-14 11:24:34,189 - root - INFO - parameter: service_name=tango/nn-model
2024-11-14 11:24:44,720 - werkzeug - INFO - 52.79.201.29 - - [14/Nov/2024 11:24:44] "POST /kvc/stop HTTP/1.1" 200 -
2024-11-14 11:25:01,511 - root - INFO - request /kvc/start
2024-11-14 11:25:01,512 - root - INFO - request data: {'service_name': 'tango/nn-model', 'port': 2222, 'cpu': 8, 'memory': 16384, 'gpu': 1}
2024-11-14 11:25:01,950 - root - INFO - output: 00a4d6ee29892ea5a4adc4cb23c1727a3ab76a88eda7d14224ed631b355511d
2024-11-14 11:25:01,950 - root - ERROR - error: None
2024-11-14 11:25:01,950 - werkzeug - INFO - 52.79.201.29 - - [14/Nov/2024 11:25:01] "POST /kvc/start HTTP/1.1" 200 -
2024-11-14 11:47:54,574 - root - INFO - request /kvc/stop
2024-11-14 11:47:54,575 - root - INFO - parameter: service_name=tango/nn-model
2024-11-14 11:48:05,109 - werkzeug - INFO - 52.79.201.29 - - [14/Nov/2024 11:48:05] "f
2024-11-14 11:48:52,836 - root - INFO - request /kvc/start
2024-11-14 11:48:52,836 - root - INFO - request data: {'service_name': 'tango/nn-model'
2024-11-14 11:48:53,115 - root - INFO - output: 40162f40af11e650b43abda37b7a9e2fd4c4e

```

NVIDIA-SMI 525.105.17 Driver Version: 525.105.17 CUDA Version: 12.0									
GPU	Name	Persistence-M	Bus-Id	Disp.A	Volatile	Uncorr. ECC			
Fan	Temp	Perf	Pwr:Usage/Cap	Memory-Usage	GPU-Util	Compute M.	MIG	M.	
0	Tesla	V100-PCIE...	Off	00000000:00:05.0 Off					
N/A	43C	P0	104W / 250W	1358MiB / 32768MiB	63%	Default			N/A

Processes:							
GPU	GI	CI	PID	Type	Process name	GPU Memory	
	ID	ID				Usage	

kt cloud 배포 모듈

Detection Service

파일 선택 선택된 파일 없음

제출

Selected File : A00_S01_F_F_03_089_02_WA_MO.mp4

File Start/Stop Camera Start/Stop

FPS : 30.4

795

본문



TANGO 서비스의 AWS 배포

- TANGO의 docker-compose.yml 파일을 ECS의 task definition으로 변환
- TANGO를 구성하는 각 모듈을 개별 컨테이너 기반으로 띄워 연동 후 클라우드를 통한 서비스

구성

운영 체제/아키텍처 Linux/X86_64	용량 공급자 FARGATE	ENI ID eni-092023ec171187c51	퍼블릭 IP 54.18
CPU 메모리 8 vCPU 16 GB	시작 유형 FARGATE	네트워크 모드 awsvpc	프라이빗 IP 172.3
플랫폼 버전 1.4.0	컨테이너 인스턴스 ID -	서브넷 ID subnet-50C7ea1a	MAC 주소 0a:5d
태스크 정의: 가장 TangoFullCluster:2		TANGO	
작업 그룹 family:TangoFullCluster			

컨테이너 (7)

컨테이너 이름	컨테이너 인스턴스 ID	이미지 URI	이미지 디지...	상태	상태	CPU
project_manager	6d8bb4b2f04241...	50253074...	sha256:fb...	Running	정상	0
postgresql	6d8bb4b2f04241...	50253074...	sha256:dc...	Running	정상	0
ondevice_deploy	6d8bb4b2f04241...	50253074...	sha256:11...	Running	정상	0
labelling	6d8bb4b2f04241...	50253074...	sha256:66...	Running	정상	0
code_gen	6d8bb4b2f04241...	50253074...	sha256:7e...	Running	정상	0
cloud_deploy	6d8bb4b2f04241...	50253074...	sha256:f3...	Running	정상	0
autonn	6d8bb4b2f04241...	50253074...	sha256:c0...	Running	정상	0

project_manager의 컨테이너 세부 정보

세부 정보 | 로그 구성 | 재시작 정책 | **네트워크 바인딩** | Docker 레이어 및 호스트 | 환경 변수 및 파일 | 볼륨 구성



TANGO

ID

Password

Log in

Don't have account? [Register now](#)

LLM 모델 구동 및 응용 사례

restructuredText 기반으로 영문 매뉴얼 작성

- 도구를 통해 문장 단위로 분해 후 번역 파일 생성
- msgstr 부분은 빈 문자열

번역 파일의 모든 msgid를 불러와 LLM 모델에 번역 요청하는 스크립트

```
prompt_template = (
    "You are tasked with translating Lablup's Backend.AI Web UI manual. "
    "The text is formatted in reStructuredText. Please translate the following "
    f"text into {target_language} exactly as written, without adding any extra "
    "information. Follow these rules:\n\n"
    "- If the text contains a URL, simply translate it without accessing or modifying the link.\n"
    "- For inline code wrapped in double backticks (`):\n"
    "  + Ensure there is a space before the opening backticks and after the closing backticks.\n"
    "  + Do not add space immediately after the opening backticks or immediately before the closing backticks.\n"
    "  + Example: '``code``' is correct, but '``code``' is not.\n"
    "- For links that start with ` and end with `_: \n"
    "  + Add a space before the opening backtick and after the closing `_. \n"
    "  + Do not add space immediately after the opening backtick or immediately before the closing `_. \n"
    "  + Example: '``link``' is correct, but '``link``' is not.\n"
    "- For references in the format :ref:`reference-name<reference-id>`: \n"
    "  + Add a space before :ref: and after the closing backtick. \n"
    "  + Do not add space immediately after :ref: or immediately before the closing backtick. \n"
    "  + Example: ':ref:`reference-name<reference-id>`' is correct, but ':ref:`reference-name<reference-id>`' is not.\n"
)

def translate_text(text, target_language):
    response = client.chat.completions.create(
        model=llm_model,
        messages=[
            {
                "role": "system",
                "content": prompt_template,
            },
            {
                "role": "user",
                "content": text,
            },
        ],
    )
    return response.choices[0].message.content.strip()
```

#: ../../overview/overview.rst:34

msgid ""

"User: The user is a person who connects to Backend.AI and performs work. "
 "Users are divided into normal users, domain admins, and superadmins "
 "according to their privileges. While ordinary users can only perform tasks "
 "related to their computing sessions, domain admins have the authority to "
 "perform tasks within a domain, and superadmins perform almost all tasks "
 "throughout the system. A user belongs to one domain and can belong to "
 "multiple projects within a domain."

msgstr ""

"ผู้ใช้: ผู้ใช้คือบุคคลที่เชื่อมต่อกับ Backend.AI และทำงาน "
 "ผู้ใช้งานจะถูกแบ่งออกเป็นผู้ใช้ทั่วไป, ผู้ดูแลโดเมน "
 "และผู้ดูแลระบบซูเปอร์ตามสิทธิ์ของพวกเขา "
 "ในขณะที่ผู้ใช้ทั่วไปสามารถทำงานที่เกี่ยวข้องกับเซสชันการคอมพิวเตอร์ของตนได้เท่านั้น "
 "ผู้ดูแลโดเมนมีอำนาจในการทำงานภายในโดเมน "
 "และผู้ดูแลระบบซูเปอร์สามารถทำงานเกือบทั้งหมดทั่วทั้งระบบ "
 "ผู้ใช้งานต้องเป็นสมาชิกของโดเมนหนึ่งและสามารถเป็นสมาชิกของโครงการหลายโครงการภายในโดเมนได้"

태국어 문장 반환

LLM model

영문장 태국어 번역 요청(API)

LLM 모델 구동 및 응용 사례


 A form for updating user information. It contains two input fields: 'Full Name' and 'Original password'. Below the 'Original password' field is a 'New password' field. To the right of the form are two buttons: 'Cancel' and 'Update'.

Each item has the following meaning.

- Full Name: User's name (up to 64 characters)
- Original password: Original password. Click the right view icon to see the input contents.
- New password: New password (8 characters or more containing at least 1 alphabet, number, and symbol). Click the right view icon to see the input contents. Ensure this is the same as the Original password.
- 2FA Enabled: 2FA activation. The user needs to enter the OTP code when logging in if it is checked.

Note: Depending on the plugin settings, the 2FA Enabled column might be invisible. In that case, please contact administrator of your system.

Enter the desired value and click the UPDATE button to update the user information.


 A form for updating user information. It contains two input fields: 'Full Name' and 'Original password'. Below the 'Original password' field is a 'New password' field. To the right of the form are two buttons: 'Cancel' and 'Update'.

แต่ละรายการมีความหมายดังต่อไปนี้




- ชื่อเต็ม: ชื่อผู้ใช้งาน (สูงสุด 64 ตัวอักษร)
- รหัสผ่านเดิม: รหัสผ่านเดิม คลิกที่ไอคอนดูทางขวาเพื่อดูเนื้อหาที่ป้อน
- รหัสผ่านใหม่: รหัสผ่านใหม่ (8 ตัวอักษรขึ้นไปโดยต้องประกอบด้วยตัวอักษร 1 ตัว ตัวเลข และสัญลักษณ์อย่างน้อย 1 ตัว) คลิกที่ไอคอนดูที่ถูกต้องเพื่อดูเนื้อหาที่ป้อน ตรวจสอบให้แน่ใจว่านี่เหมือนกับรหัสผ่านเดิม
- เปิดใช้งาน 2FA: การเปิดใช้งาน 2FA ผู้ใช้จำเป็นต้องป้อนรหัส OTP เมื่อเข้าสู่ระบบหากมีการเลือกใช้งาน

Note: ขึ้นอยู่กับการตั้งค่า plugin คอลัมน์ 2FA Enabled อาจไม่แสดงให้เห็น ในกรณีนั้นกรุณาติดต่อผู้ดูแลระบบของระบบคุณ

ป้อนค่าที่ต้องการและคลิกปุ่ม UPDATE เพื่ออัปเดตข้อมูลผู้ใช้.

- 다양한 하드웨어 환경에서 TANGO 생성 모델 배포 및 구동 지원
 - NVIDIA GPU 외 AI 전용 가속 장치에 대한 컨테이너 기반 모델 배포 지원
 - 클라우드 사업자 또는 국내외 AI 반도체 업체의 가속기 실증 진행
- 비용과 성능을 고려한 최적 배포 프로세스를 지원

감사합니다.

주관 ETRI () 주최  과학기술정보통신부  정보통신기획평가원

후원

