GREENNODE

ONE STOP SOLUTION FOR YOUR AI JOURNEY

Join us as we elevate your AI experience and propel your business into the future.

Tung Vu – Head of Product
tung.vu@greennode.ai

Agenda

1. Introduction

- About GreenNode – Nvidia's CSP
- GreenNode's AI Cloud products

2. GreenNode LLM's and Applications

VNG DB - Product Portfolio

Cloud and SaaS

VNG CLOUD

SAAS - Industry Solutions

A4B App for Business

Veka.ai

PAAS

Data platform

AI Driven

IAAS

Compute

Storage

CDN

Managed Services

Infrastructure

Bandwidth

Bare-Metal

AI/ML

GREENNODE

NVIDIA NGC Catalog

LLM

Speech AI

Recommenders

Cybersecurity

Medical Imaging

Video Analytics

Route Optimization

100+ AI frameworks and pre-trained models

Model-as-a-service

Stable Diffusion

OCR

GenAI Chatbot

AI/ML Platform

Training

Fine tuning

Inferencing

Baremetal as a Service

H100

L40S/A40

RTX 4090

Security

verichains

Professional Services

Application Security

Blockchain Audit

Red Team / Pen-test

Managed Detection & Response

Security Compliance

Security Products

BShield
Mobile App Protection

Secure-ID
Biometrics MFA

TrueID eKYC

TrueID AML

VNG Digital Business

WHO WE ARE

The Regional Cloud Services Partner of NVIDIA

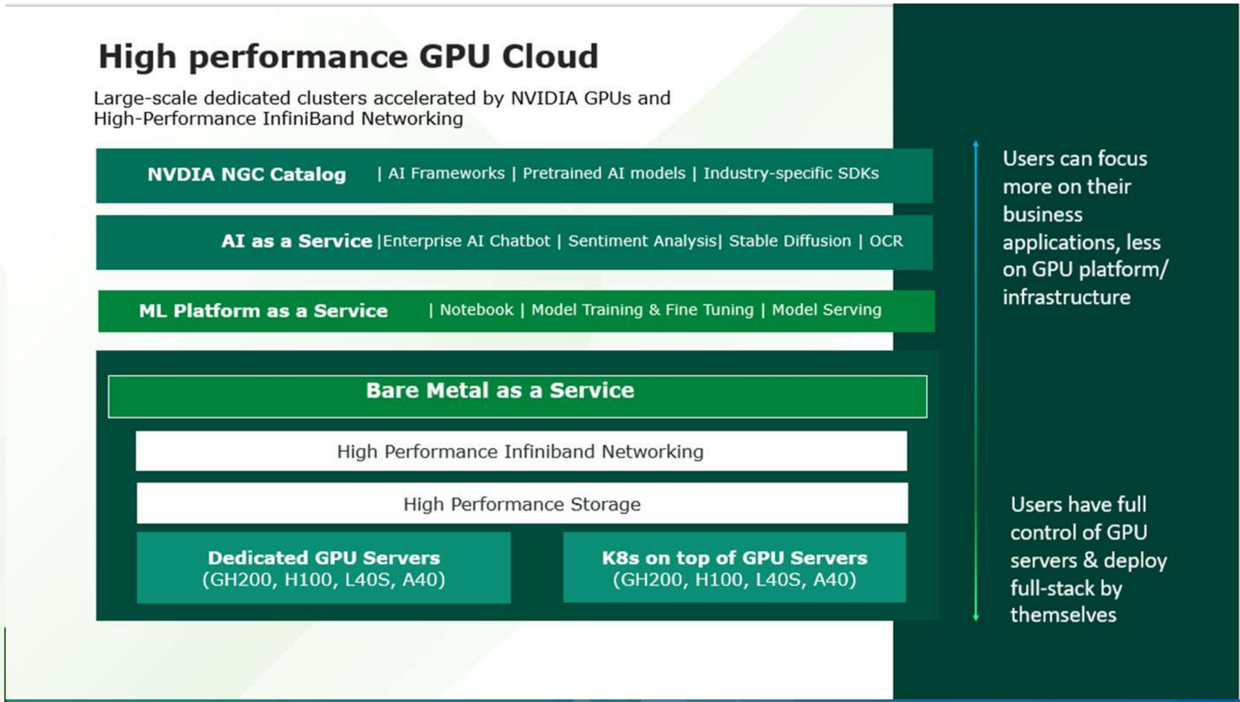
- Leading cloud service provider for hundreds of businesses
- Data centers for GPU Cloud operate in Vietnam, Thailand, and more.
- Certified with LEED Gold, TIA 942 Rating-3 DCDV, Uptime Institute Tier III (TCDD and TCCF).
- Compliant with ISO 27000, PCI DSS, and TVRA standards.

STT Bangkok 1 is a carrier-neutral data center constructed according to the highest industry standards. Strategically located in Hua Mak and forming a part of the STT Bangkok mega-scale data center campus in Thailand, its construction aims to support the rapidly growing digital infrastructure needs driven by the surge in rich media consumption and the ongoing digital transformation of enterprises.

GREENNODE

Enterprise-Grade AI capabilities and scale serving US/regional AI enterprises

- **AI as a Service: Stable Diffusion** (50k images generated over 3 months); **OCR** processing ~14k documents per day for Asia Commercial Bank (~$800M FY23 before tax profit)
- **Model Fine Tuning:** GreenNode-14B Vietnamese LLM achieved first place in VLSP 2023; surpasses ChatGPT (GPT3.5) in VMLU benchmark
- **AI Infrastructure:** initially 2000 H100 GPUs



High performance GPU Cloud

Large-scale dedicated clusters accelerated by NVIDIA GPUs and High-Performance InfiniBand Networking

NVDIA NGC Catalog | AI Frameworks | Pretrained AI models | Industry-specific SDKs

AI as a Service |Enterprise AI Chatbot | Sentiment Analysis| Stable Diffusion | OCR

ML Platform as a Service | Notebook | Model Training & Fine Tuning | Model Serving

Bare Metal as a Service

High Performance Infiniband Networking

High Performance Storage

Dedicated GPU Servers (GH200, H100, L40S, A40)

K8s on top of GPU Servers (GH200, H100, L40S, A40)

Users can focus more on their business applications, less on GPU platform/ infrastructure

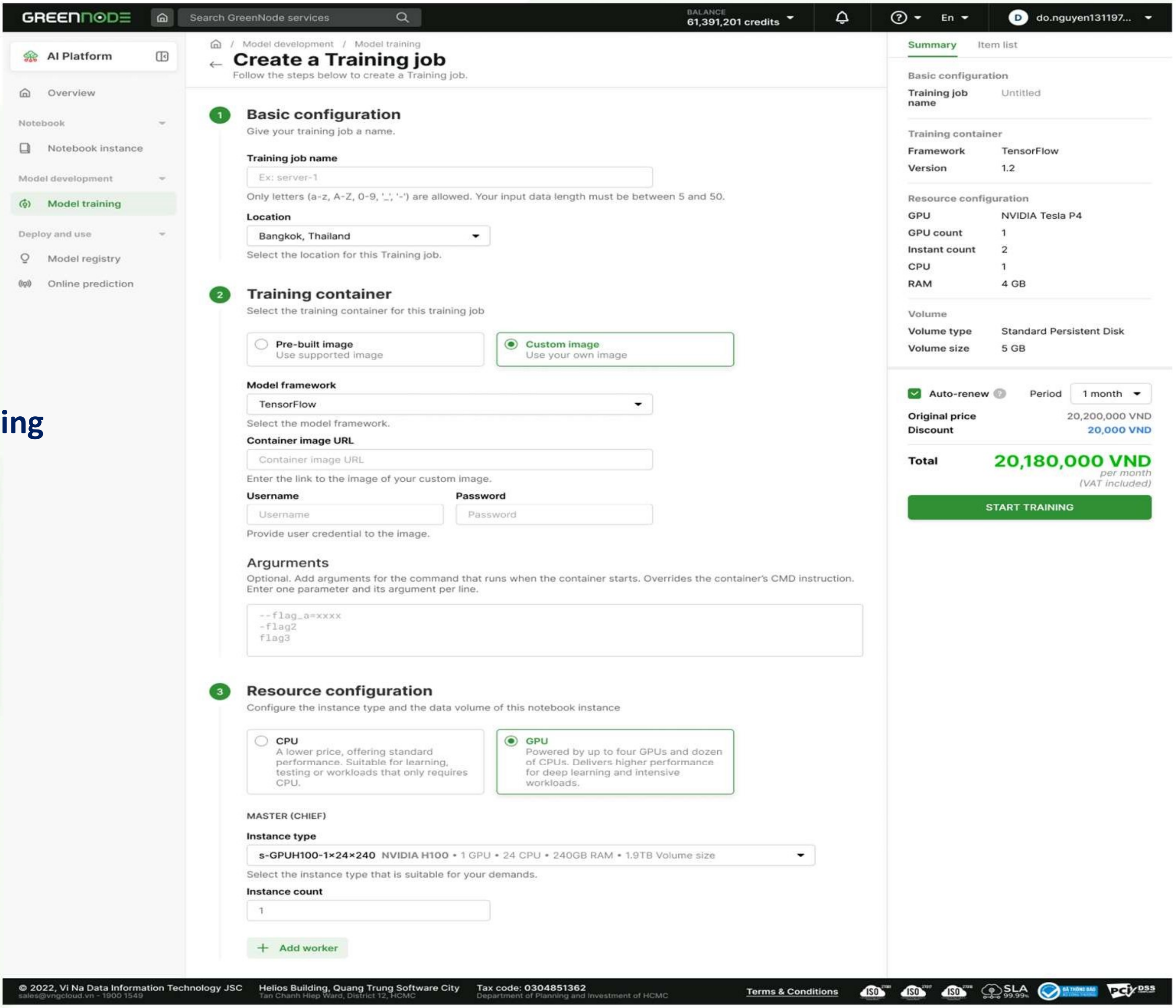
Users have full control of GPU servers & deploy full-stack by themselves

Digital Footprint across Southeast Asia

Our AI-oriented data centers are located across SEA countries.

LLM model training with our ML platform backed by thousands of Nvidia H100 GPUs & IB network

- LLM Model training/finetuning/inferencing
- GenAI chatbot for customer services
- OCR for invoice contract, etc & RPA integration
- Sentiment Analysis



Create a Training job

Follow the steps below to create a Training job.

1 Basic configuration

Give your training job a name.

Training job name

Ex: server-1

Only letters (a-z, A-Z, 0-9, '_', '-') are allowed. Your input data length must be between 5 and 50.

Location

Bangkok, Thailand

Select the location for this Training job.

2 Training container

Select the training container for this training job

○ **Pre-built image** Use supported image

● **Custom image** Use your own image

Model framework

TensorFlow

Select the model framework.

Container image URL

Container image URL

Enter the link to the image of your custom image.

Username Username

Password Password

Provide user credential to the image.

Arguments

Optional. Add arguments for the command that runs when the container starts. Overrides the container's CMD instruction. Enter one parameter and its argument per line.

```
--flag_a=xxxx
-flag2
flag3
```

3 Resource configuration

Configure the instance type and the data volume of this notebook instance

○ **CPU** A lower price, offering standard performance. Suitable for learning, testing or workloads that only requires CPU.

● **GPU** Powered by up to four GPUs and dozen of CPUs. Delivers higher performance for deep learning and intensive workloads.

MASTER (CHIEF)

Instance type

s-GPUH100-1×24×240 NVIDIA H100 • 1 GPU • 24 CPU • 240GB RAM • 1.9TB Volume size

Select the instance type that is suitable for your demands.

Instance count

1

+ Add worker

Summary Item list

Basic configuration	
Training job name	Untitled
Training container	
Framework	TensorFlow
Version	1.2
Resource configuration	
GPU	NVIDIA Tesla P4
GPU count	1
Instant count	2
CPU	1
RAM	4 GB
Volume	
Volume type	Standard Persistent Disk
Volume size	5 GB

☑ Auto-renew Period 1 month

Original price	20,200,000 VND
Discount	20,000 VND
Total	20,180,000 VND per month (VAT included)

START TRAINING

© 2022, Vi Na Data Information Technology JSC
sales@vngcloud.vn - 1900 1549

Helios Building, Quang Trung Software City
Tan Chanh Hiep Ward, District 12, HCMC

Tax code: 0304851362
Department of Planning and Investment of HCMC

Terms & Conditions

AI products applications under VNG Digital Business

AI Technologies

- Face Recognition
- ID Card Information Extraction
- Liveness Detection
- ID Forensics
- OCR
- Speech Recognition
- Edge Processing
- Deepfake
- LLM
- Object Detection/Tracking
- Person Re-Identification
- Stable Diffusion

Features/Applications

- eKYC
- Face Access Control
- AI Camera
- VideoCall Identity
- Document OCR
- Smart Surveillance
- Biometric Authentication/ Identification
- Vehicle Tracking
- Chatbot
- License Recognition
- Receipt OCR
- KMP
- Sentiment Analysis

Industries

- Banking
- Lending/Fintech
- Insurance
- IOT
- Surveillance

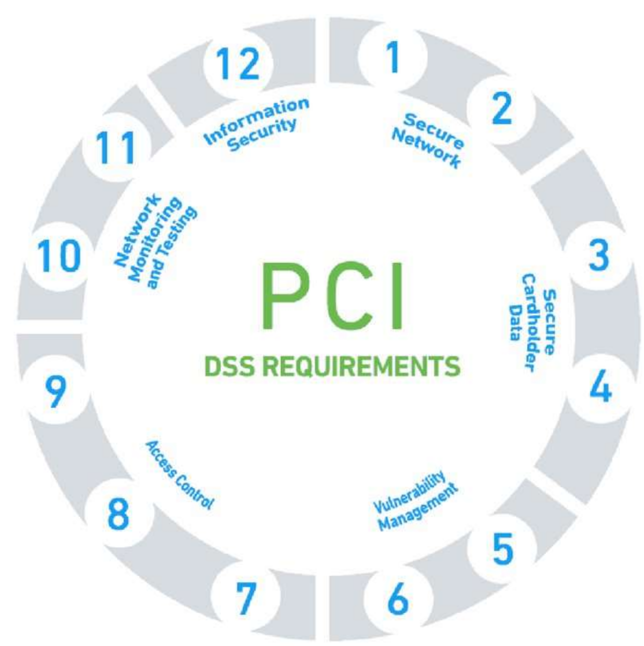
Digital Business aims to deliver Enterprise –Grade AI Services

Quality: continuously improve AI models to enhance accuracy

Reliability and Scalability: ~0% failed rate on API requests

Data Security and Compliance: PCI-DSS, ISO 27001, Data are encrypted at rest, and in transit according to the above standards

Flexible Integration: APIs, Mobile SDKs, Web SDK, web-view SDKs. Support multiple devices and browsers



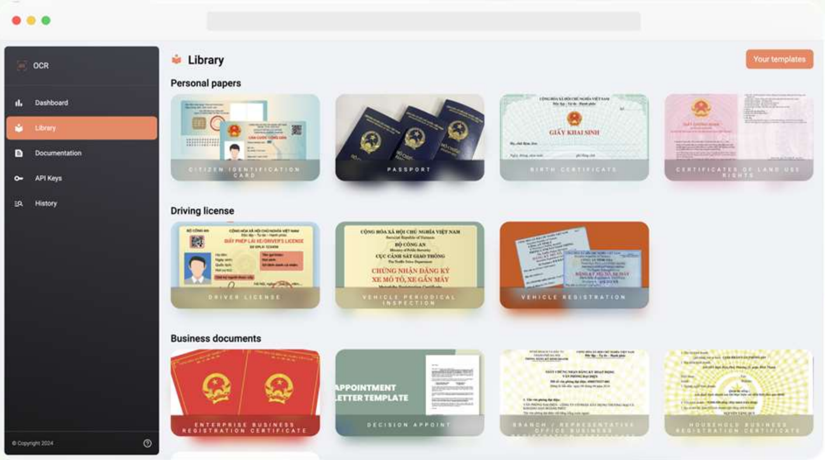
Limitations of traditional OCR

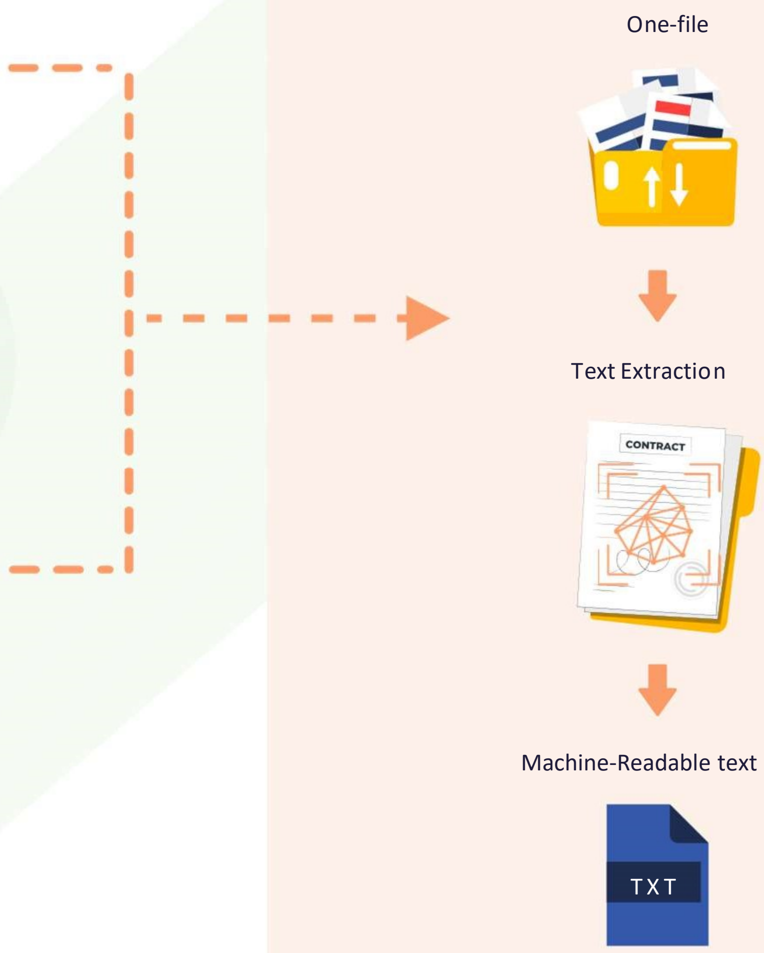
1 Required manual customization and human evaluation

2 Rely on pre-defined patterns, models or templates

3 Lower accuracy rate

4 Limited language support and contextual understanding


OCR
Dashboard
Library
Documentation
API Keys
History
© Copyright 2024
Library
Your templates
Personal papers
CITIZEN IDENTIFICATION CARD
PASSPORT
GIẤY KHAI SINH
BIRTH CERTIFICATE
CERTIFICATES OF LAND USE RIGHTS
Driving license
DRIVER LICENSE
CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
BỘ CÔNG AN
CỤC CẢNH SÁT GIAO THÔNG
CHỨNG NHẬN ĐĂNG KÝ XE MÔ TÔ, XE GẮN MÁY
VEHICLE PERIODICAL INSPECTION
VEHICLE REGISTRATION
Business documents
ENTERPRISE BUSINESS REGISTRATION CERTIFICATE
APPOINTMENT LETTER TEMPLATE
DECISION APPOINT
BRANCH / REPRESENTATIVE OFFICE BUSINESS
HOUSEHOLD BUSINESS REGISTRATION CERTIFICATE

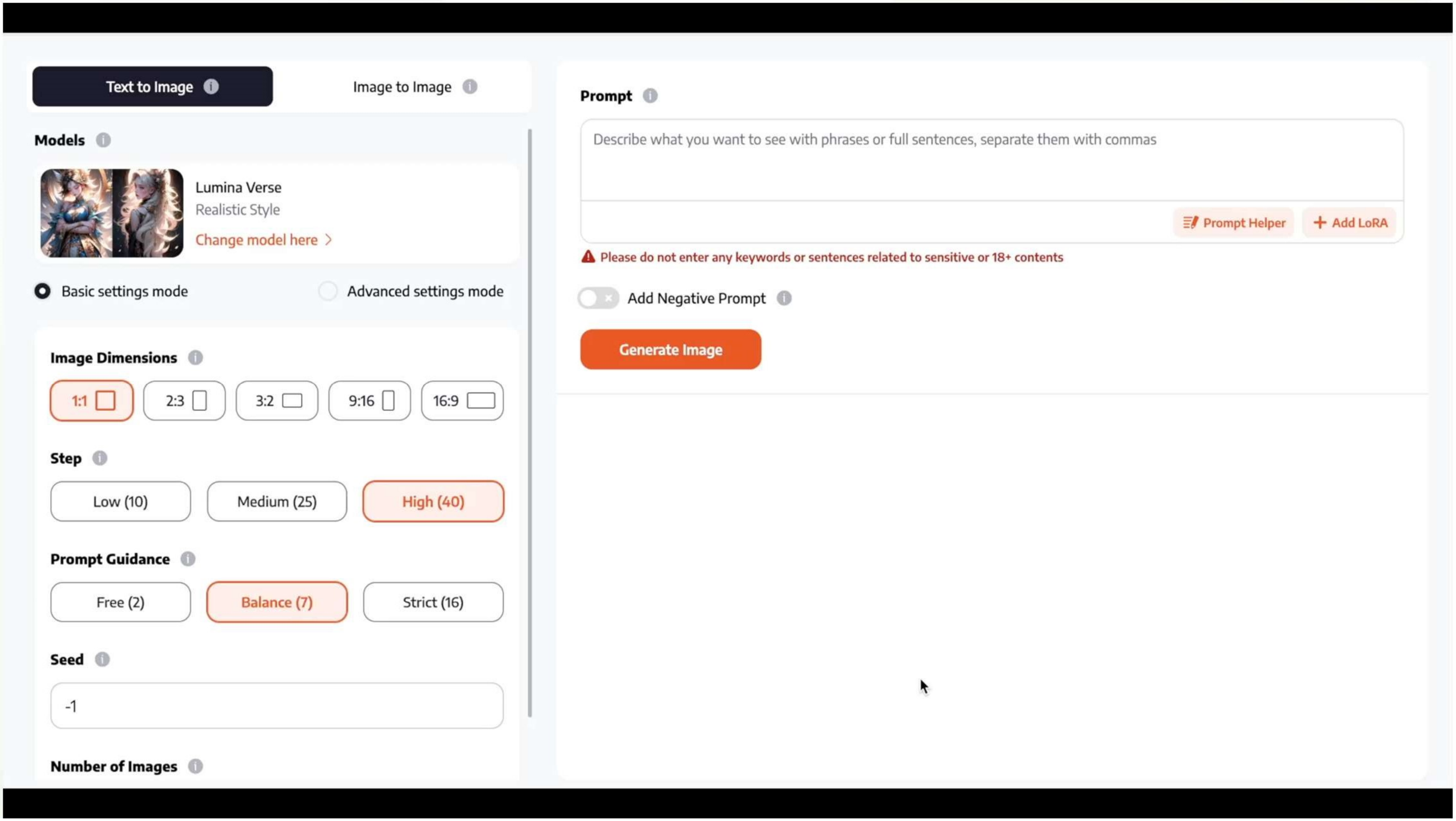


One-file

Text Extraction

Machine-Readable text

Gaming Characters and Marketing Banners Design by Stable Diffusion



[✓ Text to Image] Image to Image

Models

Lumina Verse
Realistic Style
Change model here >

● Basic settings mode ○ Advanced settings mode

Image Dimensions

[✓ 1:1] 2:3 3:2 9:16 16:9

Step

Low (10) Medium (25) [✓ High (40)]

Prompt Guidance

Free (2) [✓ Balance (7)] Strict (16)

Seed

-1

Number of Images

Prompt

Describe what you want to see with phrases or full sentences, separate them with commas

Prompt Helper + Add LoRA

⚠ Please do not enter any keywords or sentences related to sensitive or 18+ contents

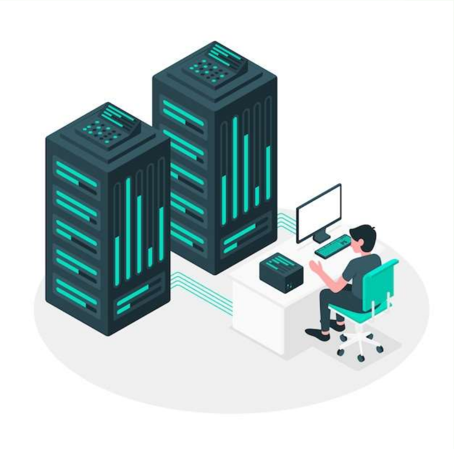
Add Negative Prompt

Generate Image

GREENNODE

Digital Footprint across Southeast Asia

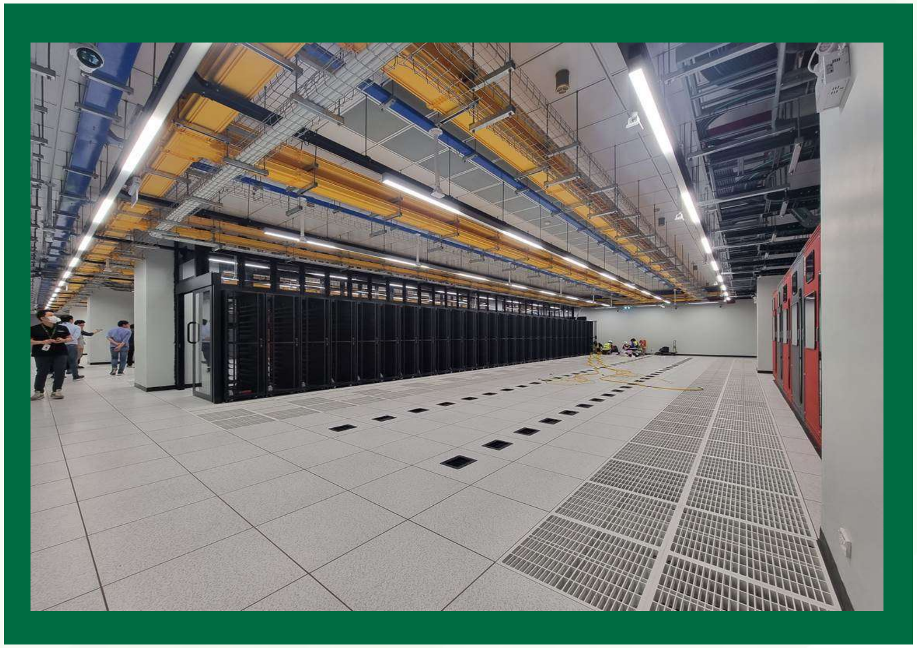
Our top-tier data centers are located across SEA countries, such as **Vietnam, Thailand** and **Malaysia.**

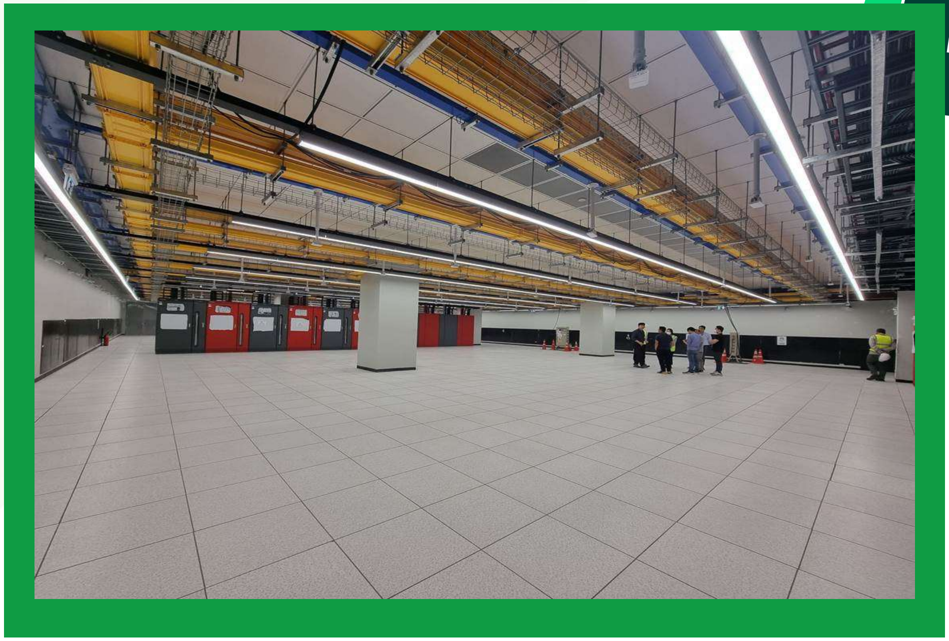


SouthEast ASIA
Vietnam
Thailand
Malaysia
Edit

Top-Tier Data Center: STT, Bangkok, Thailand

Starting March 2024: Equipped with a 3MW power reserve, fully meeting operational needs for 256 HGX H100 servers (including a matched IB network and high-performance storage systems).





> 200 Engineers & AI Experts

- **TOE:** 65 staff
- **Engineer:** 70 staff
- **Data Center Operations:** 50 staff
- **AI Lab:** 30 staff

GREENNODE

Pricing

STT	Instance Flavor	vCPU	Memory (GB)	vRAM (GB)	# of H100 SXM5 card	Local NVMe	Price per GPU instance per hour
1	g5-standard-16x250-1h100	16	250	80	1	3.75TB	$3.89
2	g5-standard-32x500-2h100	32	500	160	2	7.5TB	$7.78
3	g5-standard-64x1000-4h100	64	1000	320	4	15TB	$15.56
4	g5-standard-128x2000-8h100	128	2000	640	8	30TB	$31.12

- 10Mbps international bandwidth: $60
- 10Mbps domestic (within Thailand): $6
- Public IP Address: $6/IP/month
- High Performance File System Storage: 0.11$ per GB/month

Agenda

1. Introduction

2. GreenNode’s LLMs & Applications

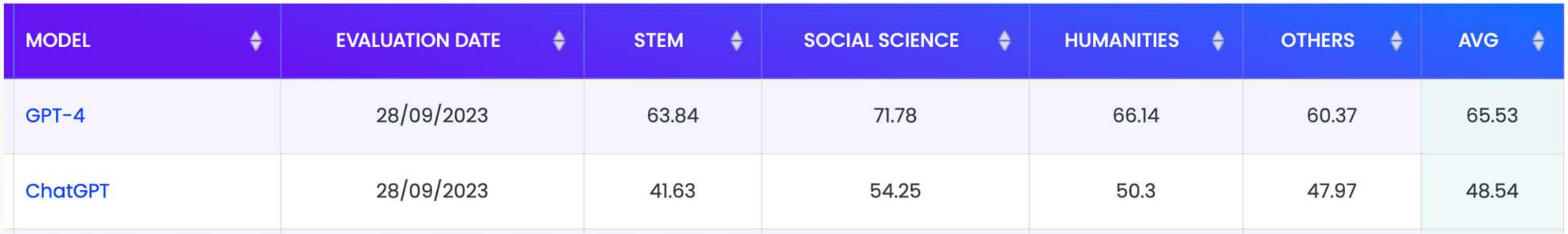
2023 Recap

In progress for developing GreenNode Large Language Model (Focusing on Vietnamese language), and Knowledge Management Platform (KMP). Key achievements:

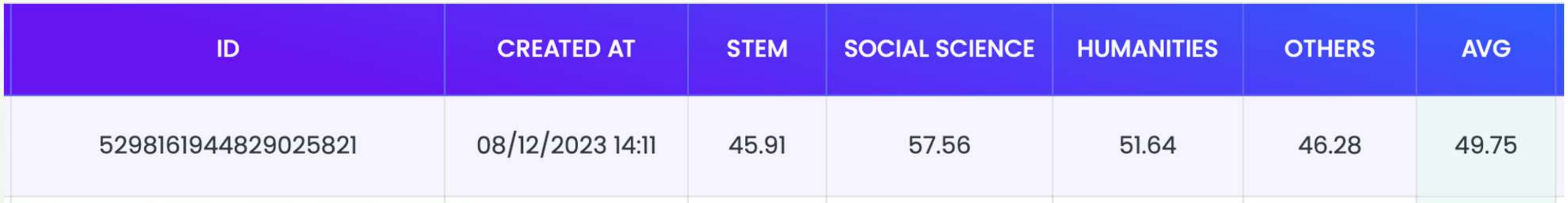
GreenNode-14B achieved first place in VLSP 2023.

GreenNode-14B surpasses ChatGPT (GPT3.5) in VMLU benchmark

Released first version of KMP: integrate to MyVNG for knowledge search, in talk with Customers (BFSI), and internal teams



MODEL	EVALUATION DATE	STEM	SOCIAL SCIENCE	HUMANITIES	OTHERS	AVG
GPT-4	28/09/2023	63.84	71.78	66.14	60.37	65.53
ChatGPT	28/09/2023	41.63	54.25	50.3	47.97	48.54

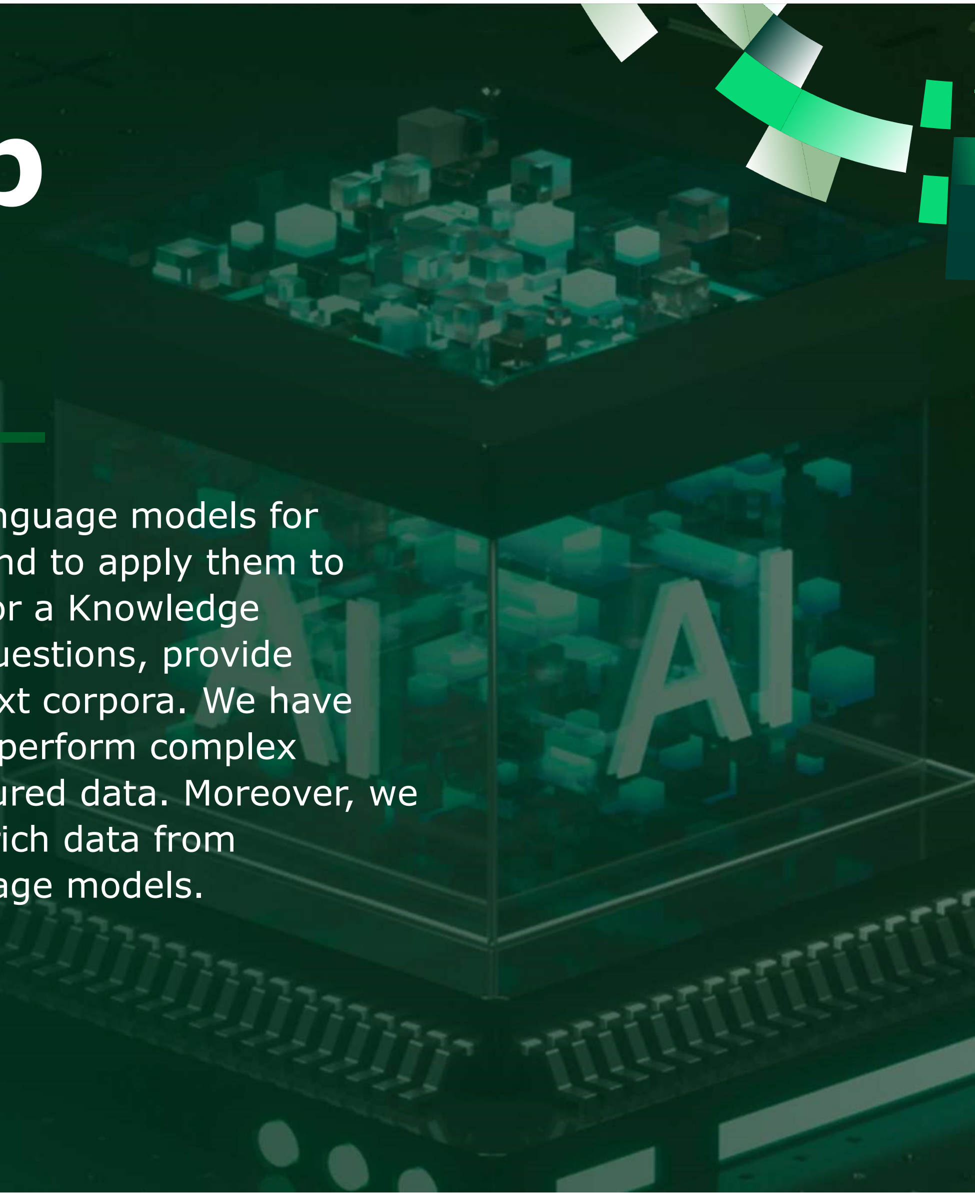


ID	CREATED AT	STEM	SOCIAL SCIENCE	HUMANITIES	OTHERS	AVG
5298161944829025821	08/12/2023 14:11	45.91	57.56	51.64	46.28	49.75

2024 Roadmap

GreenNode
LLMs and Applications

The mission for 2024 is to develop state-of-the-art language models for Vietnamese and other languages in Southeast Asia, and to apply them to various domains and tasks. The team is in progress for a Knowledge Management platform and chatbot that can answer questions, provide insights, and generate summaries from large-scale text corpora. We have also planned to develop a data analytics bot that can perform complex queries and visualizations on structured and unstructured data. Moreover, we have created a data platform that can update and enrich data from CommonCrawl, a public web archive, using our language models.



GreenNode LLM’s

Data Platform

- build data pipeline to get update data from CommonCrawl and process

- Create instruct data for LLM and NLP tasks

- Research: publish a paper (grade from B+)

Models

- LLLM for Vietnamese: GreenNode-14B, GreenNode-34B. Objective: 80% of GPT4. by Q1/2024, 90% of GPT4 by Q2/2024

- Machine Translation: Translation model for languages in SEA (Vietnamese, English, Chinese, Thai, Indonesian, Mandarin) – GreenNode-MT-SEA. Timeline: Q3/2024

- LLM for SEA: First version GreenNode-SEA-34B. Timeline: Q4/2024

Applications

- KMP AI Chatbot: complete features for Vietnamese, combine finetuned models. Timeline: Q1/2024

- Analytics Bot: initial version by Q1/2024. complete version by Q2/Q3 2024

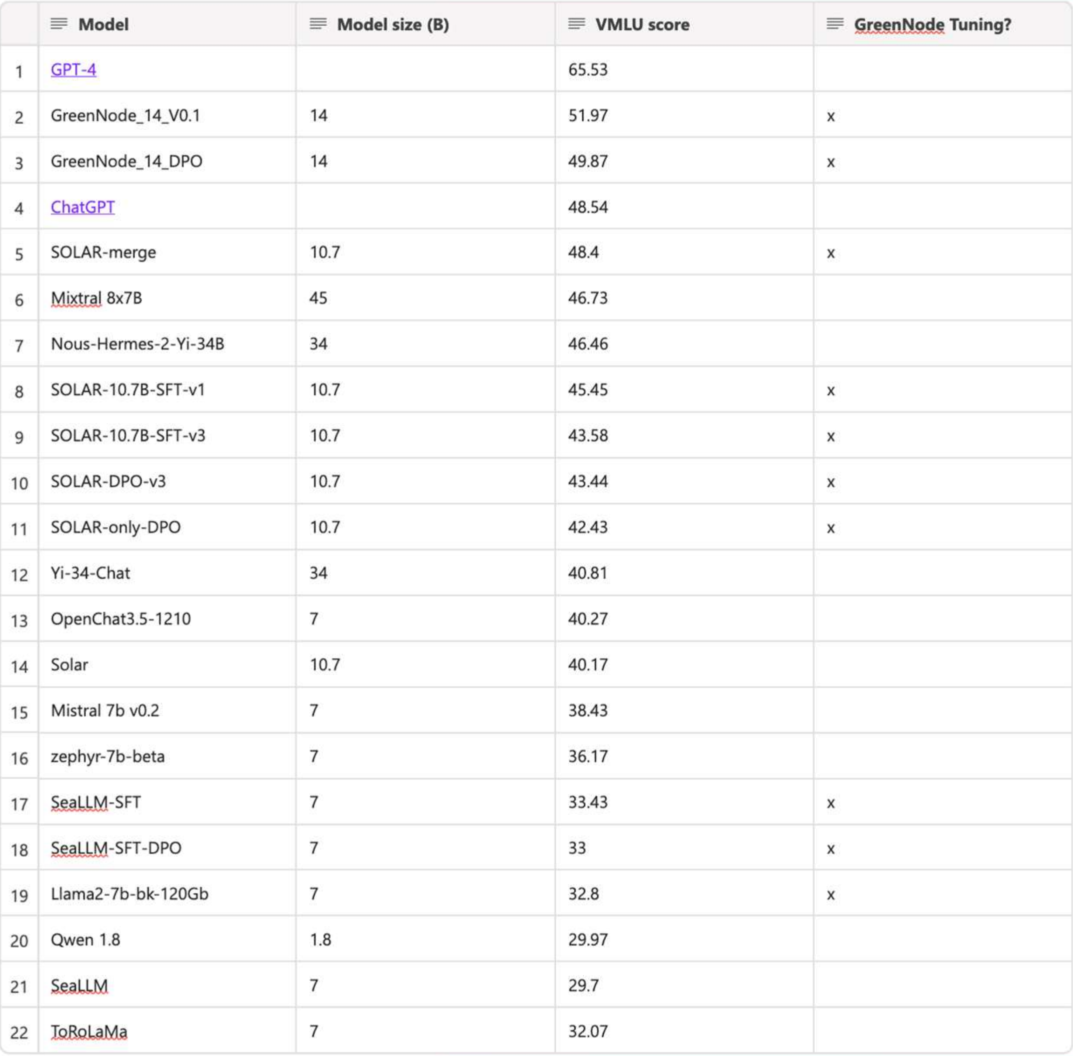
- NLP tasks: Sentiment Analysis, Machine Translation (Q1/Q2 2024)

- Automation task -> Agent (Q3-4/2024)

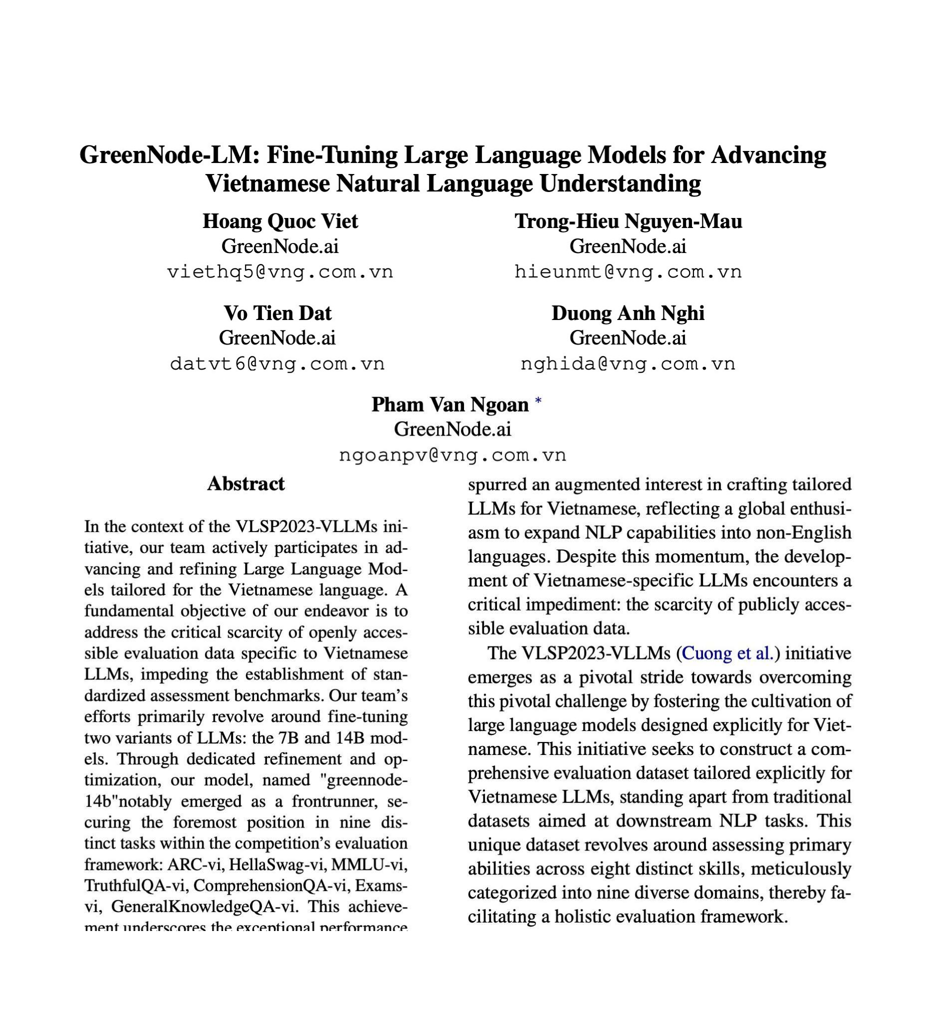
GreenNode on A Vietnamese Multitask Language Understanding (VMLU)

by Jan 2024

More about GreenNode: LLM Diary: GreenNode Makes Striking Debut with Exceptional Results at VLSP 2023 (Part 1)



	Model	Model size (B)	VMLU score	GreenNode Tuning?
1	GPT-4		65.53	
2	GreenNode_14_V0.1	14	51.97	x
3	GreenNode_14_DPO	14	49.87	x
4	ChatGPT		48.54	
5	SOLAR-merge	10.7	48.4	x
6	Mixtral 8x7B	45	46.73	
7	Nous-Hermes-2-Yi-34B	34	46.46	
8	SOLAR-10.7B-SFT-v1	10.7	45.45	x
9	SOLAR-10.7B-SFT-v3	10.7	43.58	x
10	SOLAR-DPO-v3	10.7	43.44	x
11	SOLAR-only-DPO	10.7	42.43	x
12	Yi-34-Chat	34	40.81	
13	OpenChat3.5-1210	7	40.27	
14	Solar	10.7	40.17	
15	Mistral 7b v0.2	7	38.43	
16	zephyr-7b-beta	7	36.17	
17	SeaLLM-SFT	7	33.43	x
18	SeaLLM-SFT-DPO	7	33	x
19	Llama2-7b-bk-120Gb	7	32.8	x
20	Qwen 1.8	1.8	29.97	
21	SeaLLM	7	29.7	
22	ToRoLaMa	7	32.07	



GreenNode-LM: Fine-Tuning Large Language Models for Advancing Vietnamese Natural Language Understanding

Hoang Quoc Viet
GreenNode.ai
viethq5@vng.com.vn

Trong-Hieu Nguyen-Mau
GreenNode.ai
hieunmt@vng.com.vn

Vo Tien Dat
GreenNode.ai
datvt6@vng.com.vn

Duong Anh Nghi
GreenNode.ai
nghida@vng.com.vn

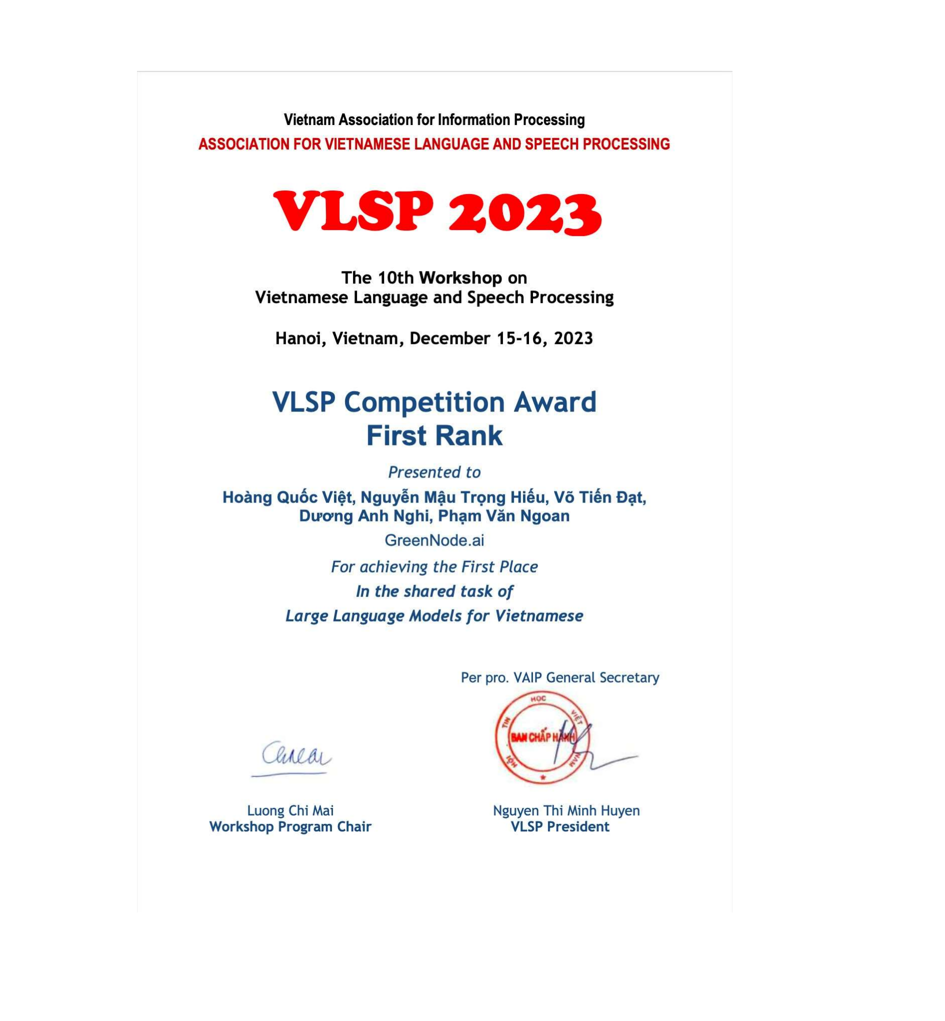
Pham Van Ngoan *
GreenNode.ai
ngoanpv@vng.com.vn

Abstract

In the context of the VLSP2023-VLLMs initiative, our team actively participates in advancing and refining Large Language Models tailored for the Vietnamese language. A fundamental objective of our endeavor is to address the critical scarcity of openly accessible evaluation data specific to Vietnamese LLMs, impeding the establishment of standardized assessment benchmarks. Our team's efforts primarily revolve around fine-tuning two variants of LLMs: the 7B and 14B models. Through dedicated refinement and optimization, our model, named "greennode-14b"notably emerged as a frontrunner, securing the foremost position in nine distinct tasks within the competition's evaluation framework: ARC-vi, HellaSwag-vi, MMLU-vi, TruthfulQA-vi, ComprehensionQA-vi, Exams-vi, GeneralKnowledgeQA-vi. This achievement underscores the exceptional performance

spurred an augmented interest in crafting tailored LLMs for Vietnamese, reflecting a global enthusiasm to expand NLP capabilities into non-English languages. Despite this momentum, the development of Vietnamese-specific LLMs encounters a critical impediment: the scarcity of publicly accessible evaluation data.

The VLSP2023-VLLMs (Cuong et al.) initiative emerges as a pivotal stride towards overcoming this pivotal challenge by fostering the cultivation of large language models designed explicitly for Vietnamese. This initiative seeks to construct a comprehensive evaluation dataset tailored explicitly for Vietnamese LLMs, standing apart from traditional datasets aimed at downstream NLP tasks. This unique dataset revolves around assessing primary abilities across eight distinct skills, meticulously categorized into nine diverse domains, thereby facilitating a holistic evaluation framework.



Vietnam Association for Information Processing
ASSOCIATION FOR VIETNAMESE LANGUAGE AND SPEECH PROCESSING

VLSP 2023

The 10th Workshop on
Vietnamese Language and Speech Processing

Hanoi, Vietnam, December 15-16, 2023

VLSP Competition Award
First Rank

Presented to

Hoàng Quốc Việt, Nguyễn Mậu Trọng Hiếu, Võ Tiến Đạt,
Dương Anh Nghi, Phạm Văn Ngoan
GreenNode.ai

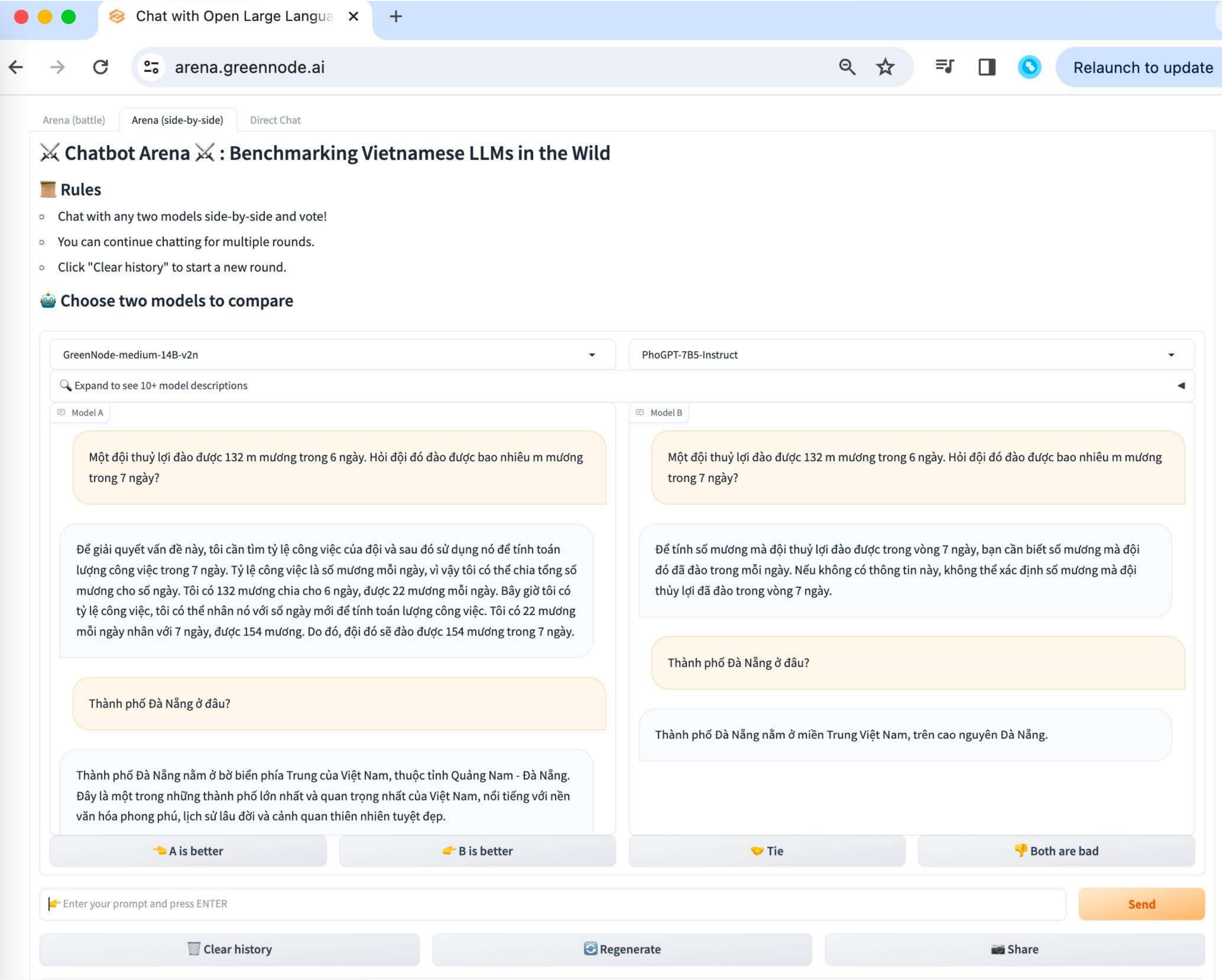
For achieving the First Place
In the shared task of
Large Language Models for Vietnamese

Per pro. VAIP General Secretary

BAN CHẤP HÀNH

Luong Chi Mai
Workshop Program Chair

Nguyen Thi Minh Huyen
VLSP President



⚔️ Chatbot Arena ⚔️ : Benchmarking Vietnamese LLMs in the Wild

📜 Rules

- Chat with any two models side-by-side and vote!
- You can continue chatting for multiple rounds.
- Click "Clear history" to start a new round.

🤖 Choose two models to compare

GreenNode-medium-14B-v2n

PhoGPT-7B5-Instruct

🔍 Expand to see 10+ model descriptions

Model A

Một đội thuỷ lợi đào được 132 m mương trong 6 ngày. Hỏi đội đó đào được bao nhiêu m mương trong 7 ngày?

Để giải quyết vấn đề này, tôi cần tìm tỷ lệ công việc của đội và sau đó sử dụng nó để tính toán lượng công việc trong 7 ngày. Tỷ lệ công việc là số mương mỗi ngày, vì vậy tôi có thể chia tổng số mương cho số ngày. Tôi có 132 mương chia cho 6 ngày, được 22 mương mỗi ngày. Bây giờ tôi có tỷ lệ công việc, tôi có thể nhân nó với số ngày mới để tính toán lượng công việc. Tôi có 22 mương mỗi ngày nhân với 7 ngày, được 154 mương. Do đó, đội đó sẽ đào được 154 mương trong 7 ngày.

Thành phố Đà Nẵng ở đâu?

Thành phố Đà Nẵng nằm ở bờ biển phía Trung của Việt Nam, thuộc tỉnh Quảng Nam - Đà Nẵng. Đây là một trong những thành phố lớn nhất và quan trọng nhất của Việt Nam, nổi tiếng với nền văn hóa phong phú, lịch sử lâu đời và cảnh quan thiên nhiên tuyệt đẹp.

Model B

Một đội thuỷ lợi đào được 132 m mương trong 6 ngày. Hỏi đội đó đào được bao nhiêu m mương trong 7 ngày?

Để tính số mương mà đội thuỷ lợi đào được trong vòng 7 ngày, bạn cần biết số mương mà đội đó đã đào trong mỗi ngày. Nếu không có thông tin này, không thể xác định số mương mà đội thủy lợi đã đào trong vòng 7 ngày.

Thành phố Đà Nẵng ở đâu?

Thành phố Đà Nẵng nằm ở miền Trung Việt Nam, trên cao nguyên Đà Nẵng.

GREENNODE

Address: 18 Cross Street #14-01
Cross Street Exchange Singapore
(048423)

Email: tung.vu@greennode.ai

Follow us on

f X in 🌐