Training a Vietnamese LLM for Complex Reasoning ↘

Chain-of-Thought Data and Self-Correction Cycles

- Overview
- LLM – GreenMind NIM
- Embedding - VN-MTEB
- GreenMind NIM for Enterprise Data flywheel

Presented by

Vo Trong Thu
Head of AI Lab

thu@greennode.ai

Pham Bao Loc
Senior AI Engineer

locpb@greennode.ai

Overview ↘

Presented by

Vo Trong Thu
Head of AI
thu@greennode.ai

Pham Bao Loc
Senior AI Engineer
locpb@greennode.ai

GreenNode AI Presentation

Training a Vietnamese LLM for Complex Reasoning
Chain-of-Thought Data and Self-Correction Cycles

AI APPLICATION FOR BUSINESS

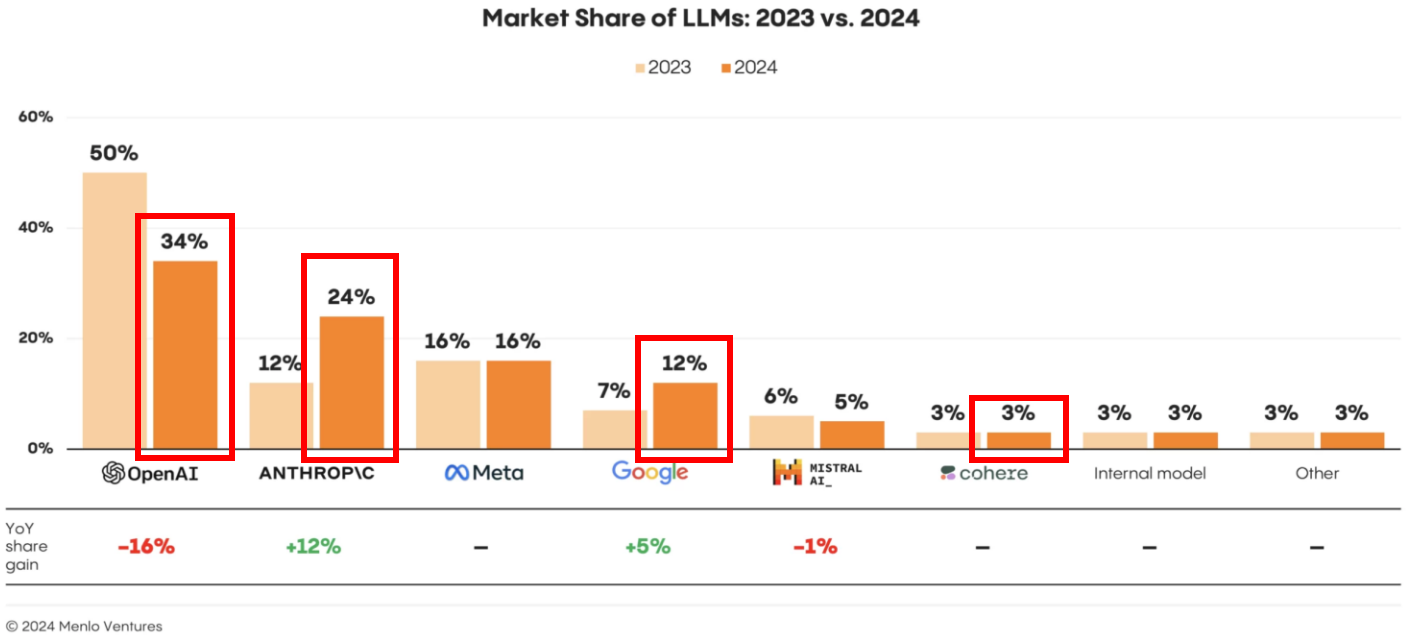
Closed-source models dominated with **73% market share**

while open-source held just 27% in enterprise use for various workload

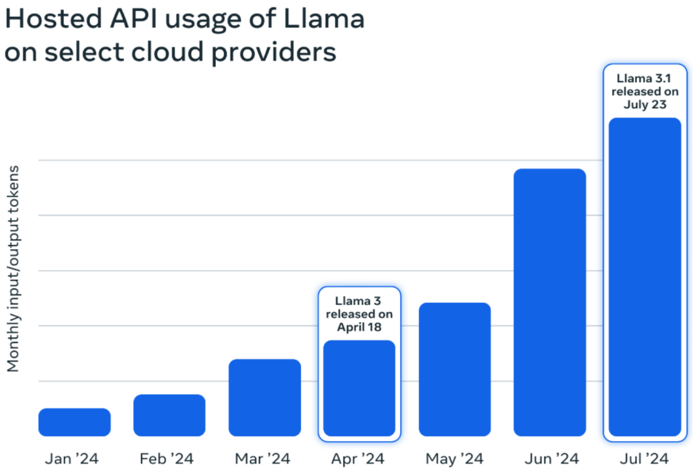
However

77% of enterprises either don't use or plan to **move away from commercial (closed-source) LLMs** beyond prototypes due to **privacy, cost, and customization** concerns.

(Source: Inside AI news)

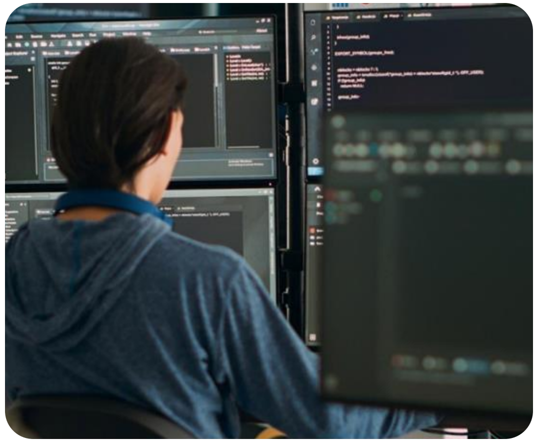


Source: *2024: The State of Generative AI in the Enterprise*



Meta's openly available models have now been downloaded more than 400 million times, at a rate 10 times higher than last year, with usage doubling from May through July 2024

Source: *With 10x growth since 2023, Llama is the leading engine of AI innovation*



WHY BUILDING SPECIFIC MODEL IS A MUST? ↘

Domain-specific optimization

Fine-tuned models understand enterprise-specific terminology and workflows

Enterprise data & privacy risk

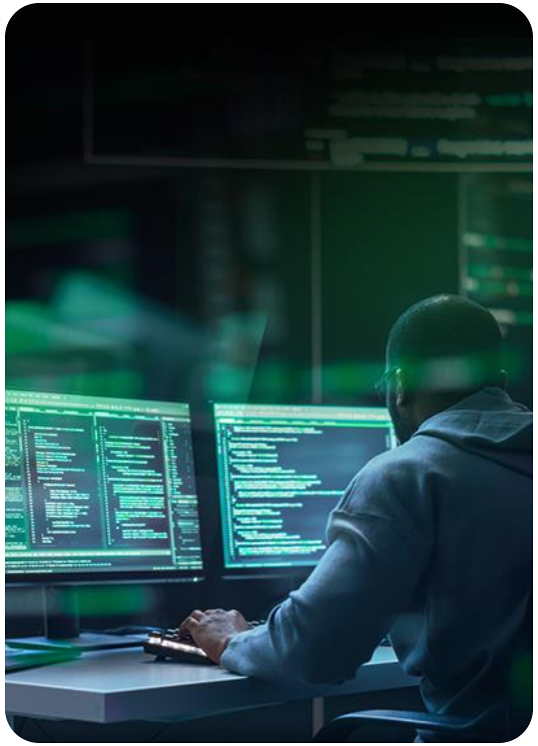
Enterprise data may leak when using closed-source models

Latency & performance control

Self-hosted models offer lower latency and better system integration

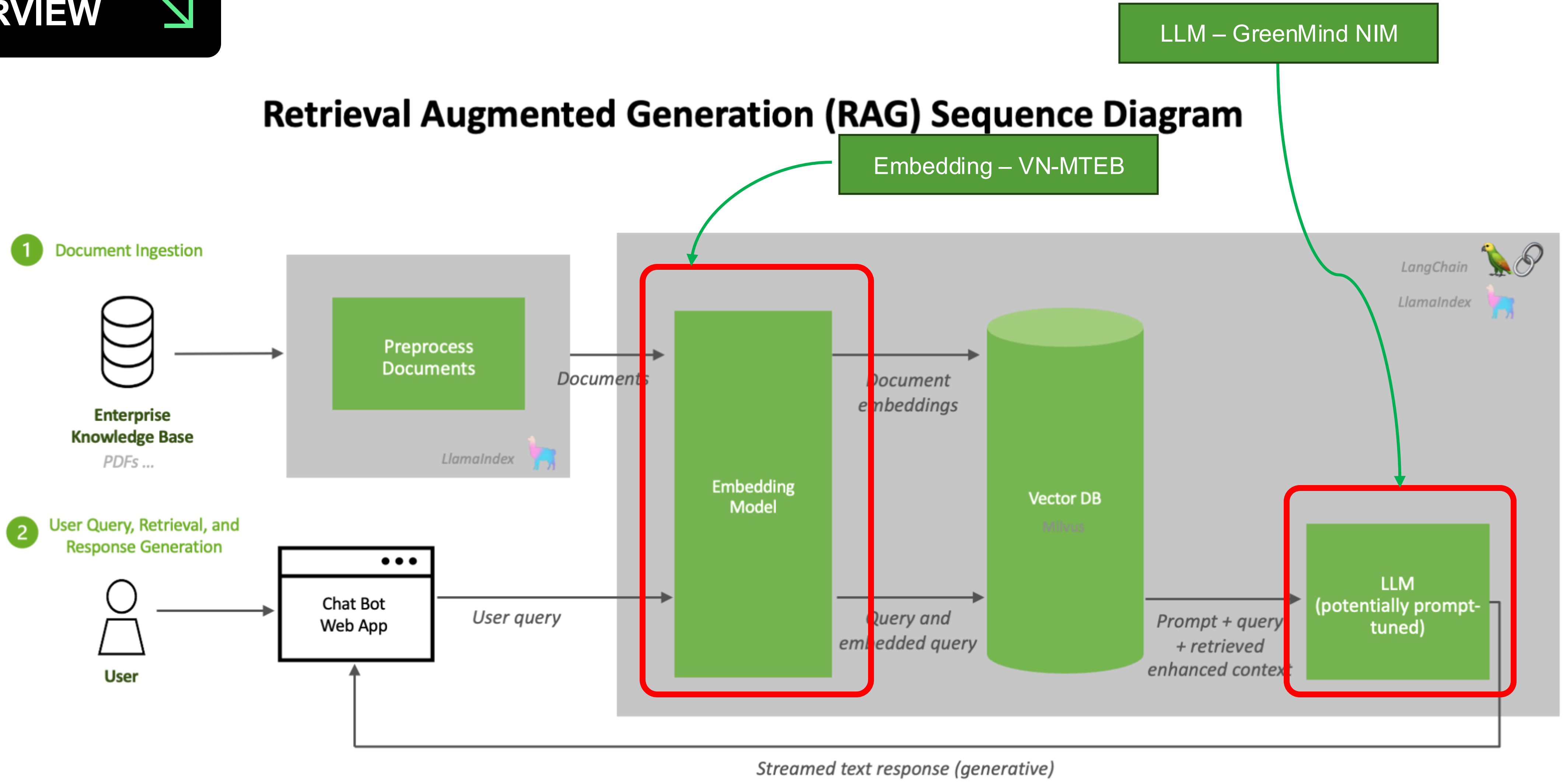
Long-term independence

Avoid vendor lock-in and maintain flexibility in infrastructure and deployment



OVERVIEW

Retrieval Augmented Generation (RAG) Sequence Diagram



Source: RAG 101: Demystifying Retrieval-Augmented Generation Pipelines

LLM ↘
GreenMind NIM

Presented by

Vo Trong Thu
Head of AI

thu@greennode.ai

Pham Bao Loc
Senior AI Engineer

locpb@greennode.ai

GreenNode AI Presentation

Training a Vietnamese LLM for Complex Reasoning
Chain-of-Thought Data and Self-Correction Cycles

GREENMIND ↘

GreenMind: A Next-Generation Vietnamese Large Language Model for Structured and Logical Reasoning

Luu Quy Tung[1] **Hoang Quoc Viet**[1*] **Pham Bao Loc**[1] **Vo Trong Thu**[2]
[1]GreenNode.ai [2]John Von Neumann Institute
`{tunglq,viethq5,locpb}@greennode.ai,thuvt@jvn.edu.vn`

Abstract

Chain-of-Thought (CoT) is a robust approach for tackling LLM tasks that require intermediate reasoning steps prior to generating a final answer. In this paper, we present **GreenMind-Medium-14B-R1** [1], the Vietnamese reasoning model inspired by the finetuning strategy based on Group Relative Policy Optimization. We also leverage a high-quality Vietnamese synthesized reasoning dataset and design two reward functions to tackle the main limitations of this technique: i) Language mixing, where we explicitly detect the presence of biased language characters during the process of sampling tokens, and ii) We leverage Sentence Transformer-based models to ensure that the generated reasoning content maintain factual correctness and do not distort the final output. Experimental results on the Vietnamese dataset from the VLSP 2023 Challenge demonstrate that our model outperforms prior works and enhances linguistic consistency in its responses. Furthermore, we extend our evaluation to SeaExam — a multilingual mutiple-choices dataset, showing the effectiveness of our reasoning method compared to few-shot prompting techniques.

reasoning, the act of users prompting models to produce direct answers only often fails to ensure accuracy. Meanwhile, at each generation step, models rely on the probability distribution over a list of candidate tokens to select the potential one by greedy or random sampling algorithms. Consequently, producing only a short sequence of tokens as the final output does not guarantee correctness, as these distributions are conditioned solely on the preceding input tokens. This implies that the models often lack the contextual understanding necessary for reasoning toward a correct answer. To address this issue, the CoT (Wei et al., 2022b) technique remains an effective approach to fully leverage the power of next token prediction. CoT encourages the model to articulate a sequence of intermediate reasoning steps, which facilitates the resolution of tasks that require multi-step logical thinking. To further enhance the reasoning capabilities of language models, a series of reinforcement learning-based methods have been proposed. Reinforcement Learning with Human Feedback (RLHF) (Ouyang et al., 2022) leveraged human-provided feedback to refine LLM outputs, ensuring that the reasoning steps generated by CoT align more closely with

---

reasoning paths and selects the most consistent answer, yielding substantial gains on GSM8K +17.9%, SVAMP (+11.0%), AQUA (+12.2%), STRATEGYQA (+6.4%), and ARC-CHALLENGE (+3.9%) (Wang et al., 2022). These studies reveal that structured intermediate reasoning can be an emergent capability in sufficiently large models.

2.2 Vietnamese Large Language Models

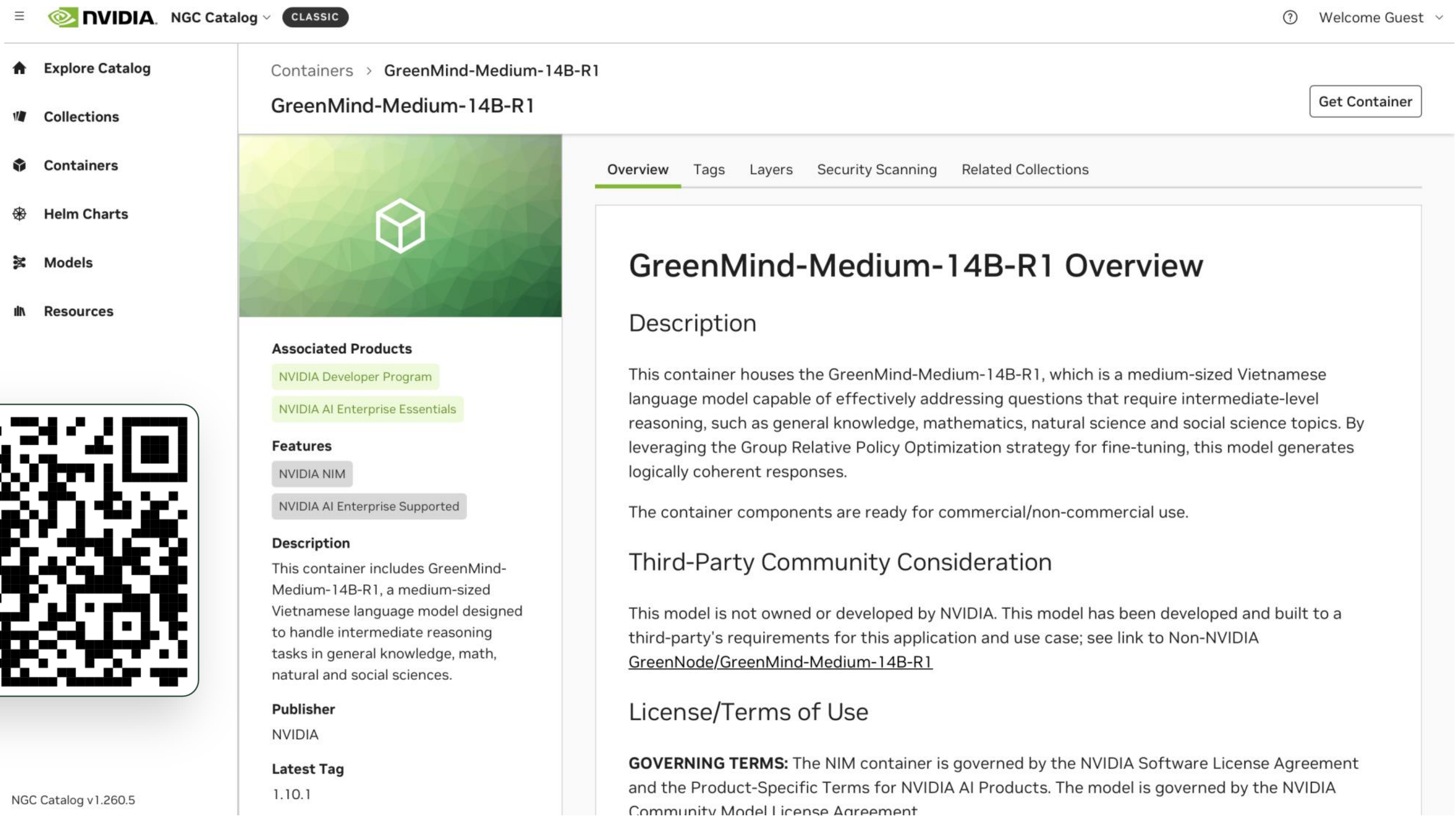
While the domain of open-source models for the Vietnamese language is relatively nascent, there are already some notable models available. These include Vietcuna 3B[2], Vietcuna-7B-v3[3], URA-LLaMA-7B[4], and URA-LLaMA-13B[5]. Vietcuna-3B and Vietcuna-7B-v3 were developed from the foundational models BLOOMZ-3B[6] and BLOOMZ-7B1[7] (Scao et al., 2022), respectively, and were further trained using 12GB of Vietnamese news texts for causal language modeling[8]. This process included fine-tuning with 200K instructional question and answer pairs, and 400K conversational samples. The URA-LLaMA models, originating from LLaMA-2, were pre-trained on Vietnamese content from Wikipedia and online news sources, with additional fine-tuning for instruction following. Furthermore, PhoGPT (Nguyen et al., 2023) have recently introduced the PhoGPT series, a new addition to the open-source generative models for Vietnamese, which includes a base 7.5B-parameter model and its instruction-following variant.

2.3 Group Relative Policy Optimization

Reinforcement learning (RL) is a subfield of Machine Learning (ML) in which an agent learns to make decisions through interactions with its environment, aiming to maximize cumulative rewards. When applied to LLMs, RL helps fine-tune these models to better align with human preferences and improve their performance on specialized tasks that require complex reasoning processes. A key category of RL algorithms is policy optimiza-

GREENNODE

GREENMIND IS NOW ON NVIDIA NIM



GreenMind-Medium-14B-R1

Associated Products

NVIDIA Developer Program

NVIDIA AI Enterprise Essentials

Features

NVIDIA NIM

NVIDIA AI Enterprise Supported

Description

This container includes GreenMind-Medium-14B-R1, a medium-sized Vietnamese language model designed to handle intermediate reasoning tasks in general knowledge, math, natural and social sciences.

Publisher

NVIDIA

Latest Tag

1.10.1

GreenMind-Medium-14B-R1 Overview

Description

This container houses the GreenMind-Medium-14B-R1, which is a medium-sized Vietnamese language model capable of effectively addressing questions that require intermediate-level reasoning, such as general knowledge, mathematics, natural science and social science topics. By leveraging the Group Relative Policy Optimization strategy for fine-tuning, this model generates logically coherent responses.

The container components are ready for commercial/non-commercial use.

Third-Party Community Consideration

This model is not owned or developed by NVIDIA. This model has been developed and built to a third-party's requirements for this application and use case; see link to Non-NVIDIA GreenNode/GreenMind-Medium-14B-R1

License/Terms of Use

GOVERNING TERMS: The NIM container is governed by the NVIDIA Software License Agreement and the Product-Specific Terms for NVIDIA AI Products. The model is governed by the NVIDIA Community Model License Agreement

REASONING DATA CURATION ↘

Pre-trained models are **too general** and lack **domain-specific knowledge** for industries such as healthcare, finance, or eCommerce. **Fine-tuning** is necessary to align models with specific industry requirements.

Each sample contain pair of Question-Answer Instruction
$i \epsilon I$

The reasoning chain r is a structure sequence of intermediate steps

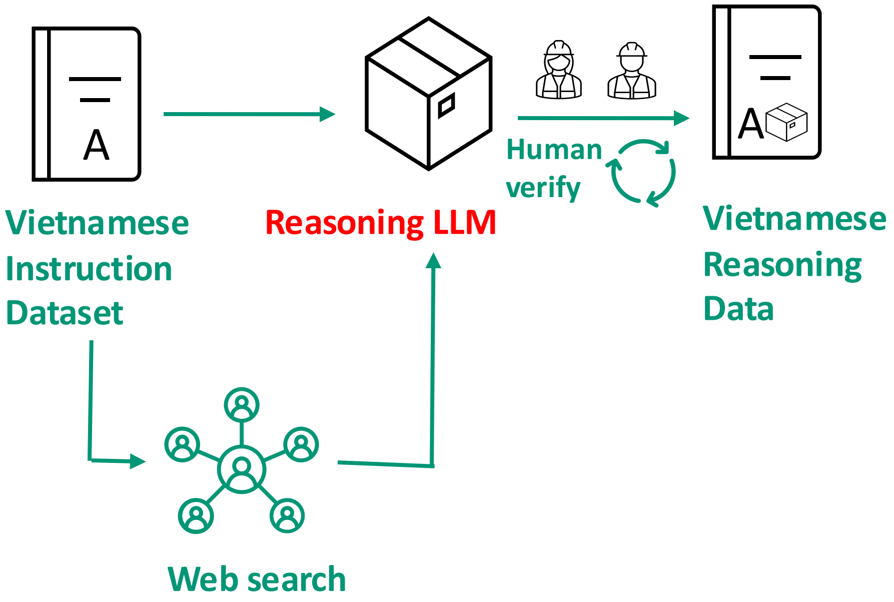
$$\{s_1, s_2, s_3, \dots, s_n\}$$

The objective for model is to generate:

- a final answer $\mathrm{a} \epsilon A$
- Reasoning chain $\mathrm{r} \epsilon R$

To enrich the factual corrects, we retrieve the supplementatry context $\mathrm{c} \epsilon C$ from web

Training data: -> 55,418 samples



Vietnamese Instruction Dataset

Reasoning LLM

Human verify

Vietnamese Reasoning Data

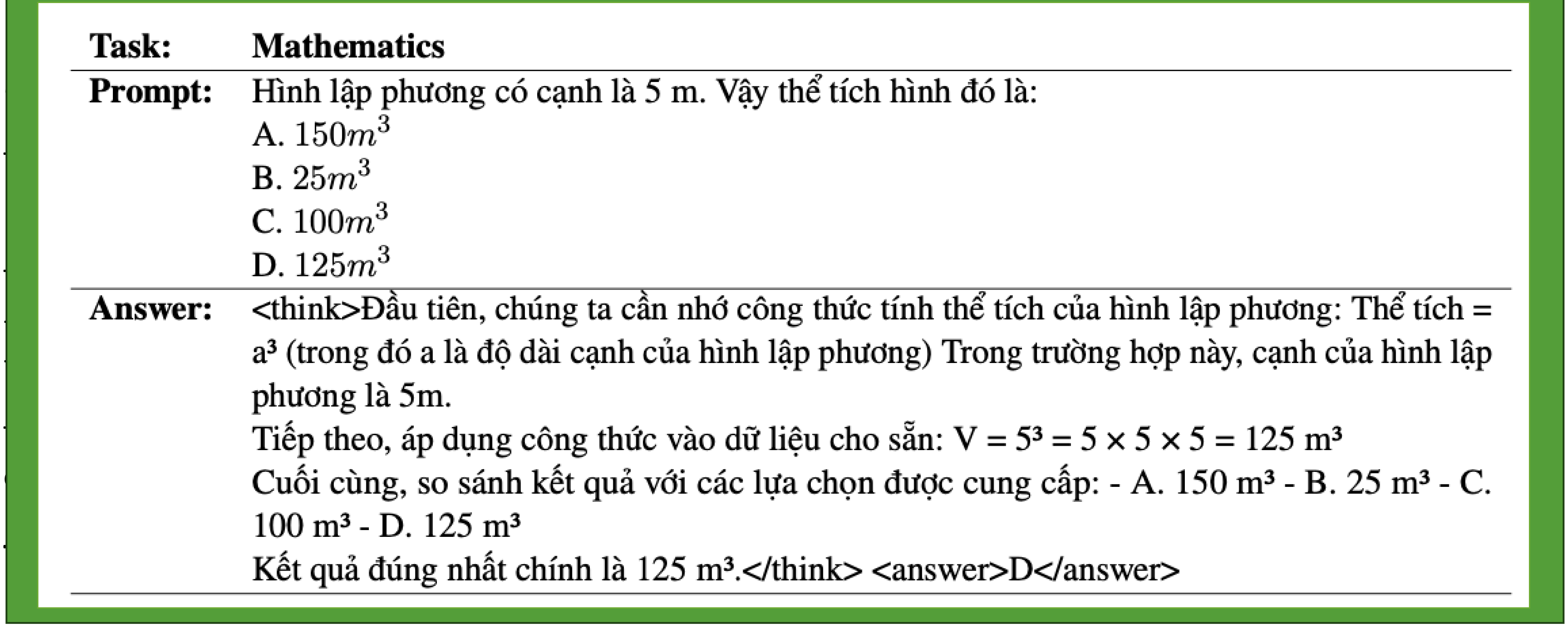
Web search

The reasoning process can be defined as a function:

$$f: I \times C \rightarrow R \times A$$

REASONING DATA CURATION

Pre-trained models are **too general** and lack **domain-specific knowledge** for industries such as healthcare, finance, or eCommerce. **Fine-tuning** is necessary to align models with specific industry requirements.



Task:	**Mathematics**
Prompt:	Hình lập phương có cạnh là 5 m. Vậy thể tích hình đó là: A. $150m^3$ B. $25m^3$ C. $100m^3$ D. $125m^3$
Answer:	<think>Đầu tiên, chúng ta cần nhớ công thức tính thể tích của hình lập phương: Thể tích = a^3 (trong đó a là độ dài cạnh của hình lập phương) Trong trường hợp này, cạnh của hình lập phương là 5m. Tiếp theo, áp dụng công thức vào dữ liệu cho sẵn: $V = 5^3 = 5 \times 5 \times 5 = 125$ m^3 Cuối cùng, so sánh kết quả với các lựa chọn được cung cấp: - A. 150 m^3 - B. 25 m^3 - C. 100 m^3 - D. 125 m^3 Kết quả đúng nhất chính là 125 m^3.</think> <answer>D</answer>

Task Type Diversity

Broad range of task such as arithmetic, world problems, commonsense inference, symbolic logic, etc.

Linguistic Complexity

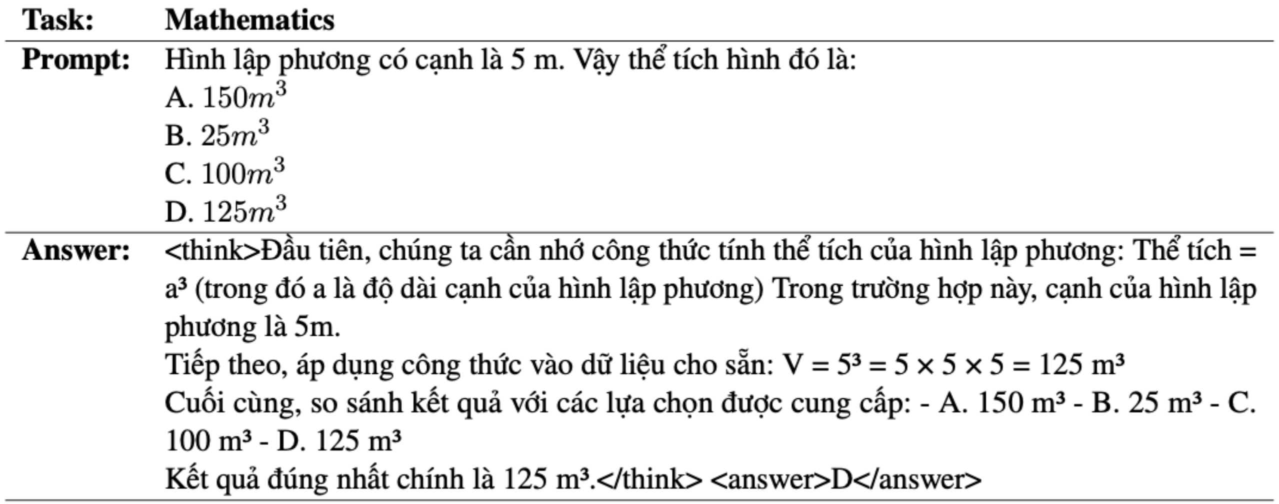
Sample across verifying syntactic and lexical complexities to challenge the model understanding of Vietnamese

Reasoning Depth

Prioritize tasks that require multi-step deduction, analogical thinking, counterfactual reasoning

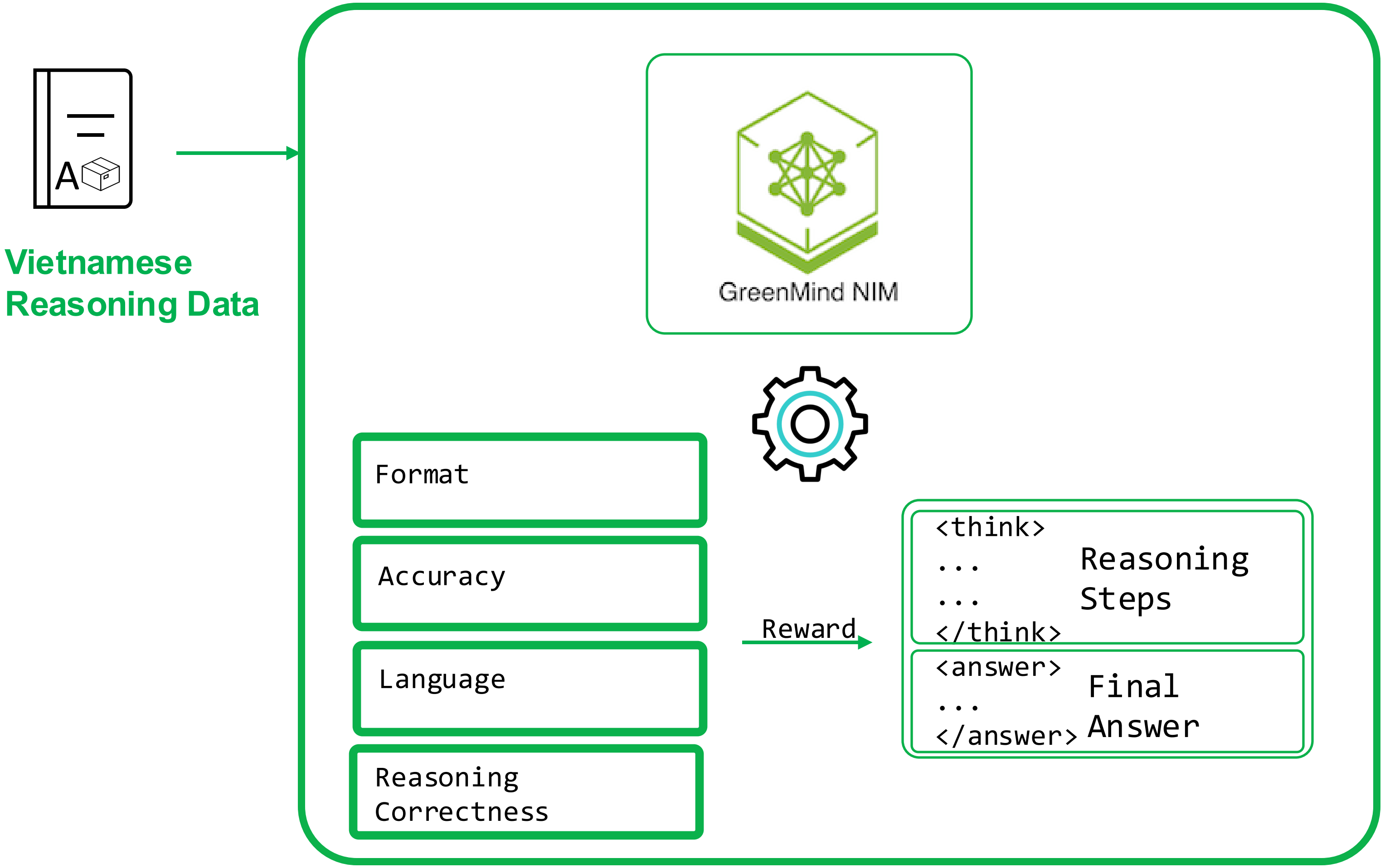
Verifiability

Each instruction-answer pair is manually verified or derived from trusted Vietnamese educational and encyclopedic sources



Task:	**Mathematics**
Prompt:	Hình lập phương có cạnh là 5 m. Vậy thể tích hình đó là: A. $150m^3$ B. $25m^3$ C. $100m^3$ D. $125m^3$
Answer:	<think>Đầu tiên, chúng ta cần nhớ công thức tính thể tích của hình lập phương: Thể tích = a^3 (trong đó a là độ dài cạnh của hình lập phương) Trong trường hợp này, cạnh của hình lập phương là 5m. Tiếp theo, áp dụng công thức vào dữ liệu cho sẵn: $V = 5^3 = 5 \times 5 \times 5 = 125$ m^3 Cuối cùng, so sánh kết quả với các lựa chọn được cung cấp: - A. 150 m^3 - B. 25 m^3 - C. 100 m^3 - D. 125 m^3 Kết quả đúng nhất chính là 125 m^3.</think> <answer>D</answer>

GROUP RELATIVELY POLICY OPTIMIZATION



Vietnamese Reasoning Data

GreenMind NIM

Format

Accuracy

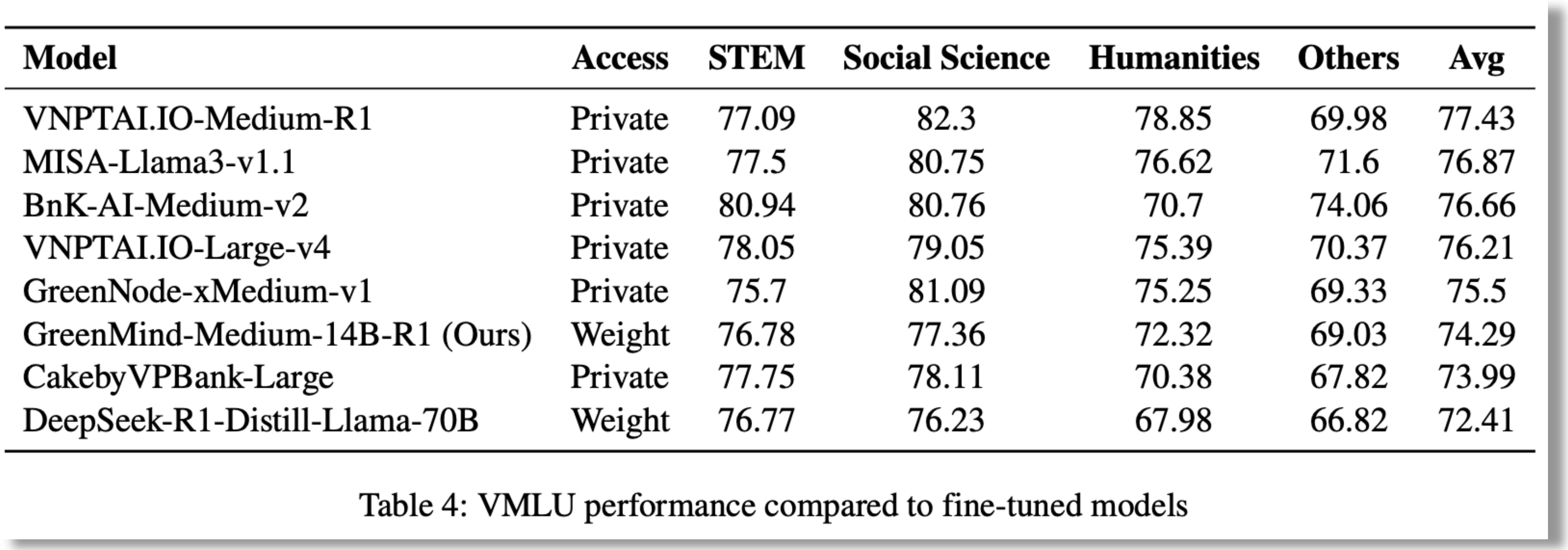
Language

Reasoning Correctness

Reward

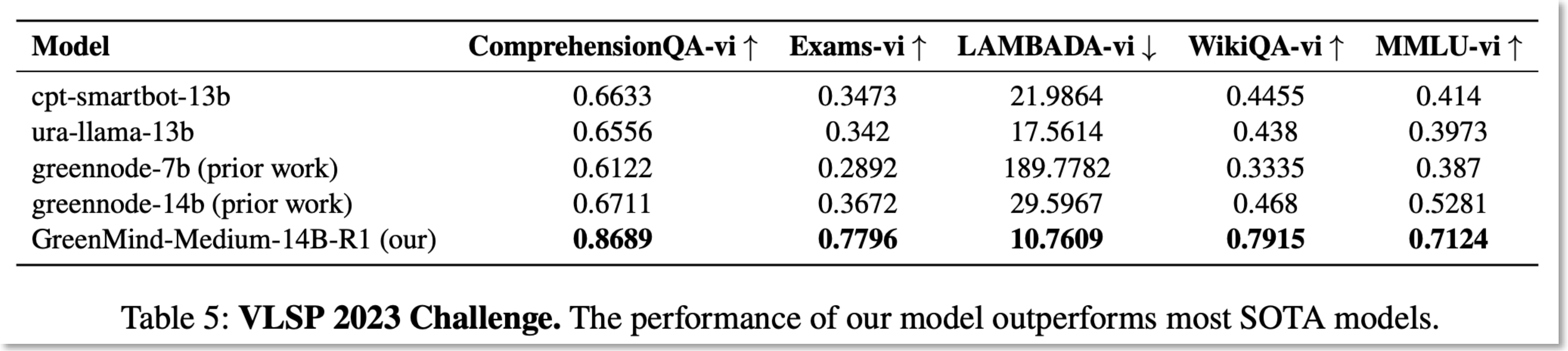
```
<think>
...        Reasoning
...        Steps
</think>
<answer>    Final
...
</answer>   Answer
```

RESULT AND INSIGHT



Model	Access	STEM	Social Science	Humanities	Others	Avg
VNPTAI.IO-Medium-R1	Private	77.09	82.3	78.85	69.98	77.43
MISA-Llama3-v1.1	Private	77.5	80.75	76.62	71.6	76.87
BnK-AI-Medium-v2	Private	80.94	80.76	70.7	74.06	76.66
VNPTAI.IO-Large-v4	Private	78.05	79.05	75.39	70.37	76.21
GreenNode-xMedium-v1	Private	75.7	81.09	75.25	69.33	75.5
GreenMind-Medium-14B-R1 (Ours)	Weight	76.78	77.36	72.32	69.03	74.29
CakebyVPBank-Large	Private	77.75	78.11	70.38	67.82	73.99
DeepSeek-R1-Distill-Llama-70B	Weight	76.77	76.23	67.98	66.82	72.41

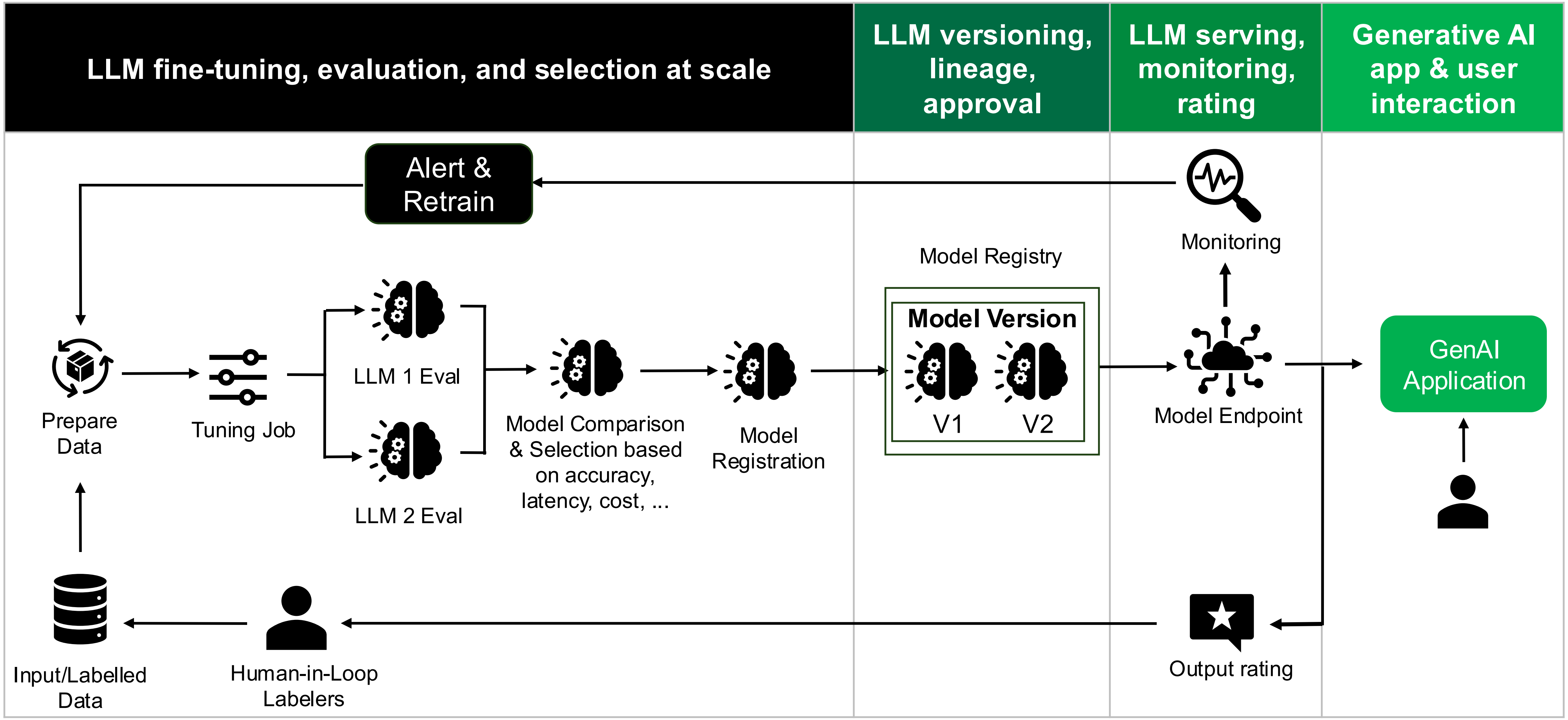
Table 4: VMLU performance compared to fine-tuned models



Model	ComprehensionQA-vi ↑	Exams-vi ↑	LAMBADA-vi ↓	WikiQA-vi ↑	MMLU-vi ↑
cpt-smartbot-13b	0.6633	0.3473	21.9864	0.4455	0.414
ura-llama-13b	0.6556	0.342	17.5614	0.438	0.3973
greennode-7b (prior work)	0.6122	0.2892	189.7782	0.3335	0.387
greennode-14b (prior work)	0.6711	0.3672	29.5967	0.468	0.5281
GreenMind-Medium-14B-R1 (our)	**0.8689**	**0.7796**	**10.7609**	**0.7915**	**0.7124**

Table 5: **VLSP 2023 Challenge.** The performance of our model outperforms most SOTA models.

GREENNODE AI PLATFORM ↘



LLM fine-tuning, evaluation, and selection at scale

LLM versioning, lineage, approval

LLM serving, monitoring, rating

Generative AI app & user interaction

Alert & Retrain

Monitoring

Model Registry

Prepare Data

Tuning Job

LLM 1 Eval

LLM 2 Eval

Model Comparison & Selection based on accuracy, latency, cost, ...

Model Registration

Model Version

V1

V2

Model Endpoint

GenAI Application

Input/Labelled Data

Human-in-Loop Labelers

Output rating

Embedding VN-MTEB ↘

Presented by

Vo Trong Thu
Head of AI
thu@greennode.ai

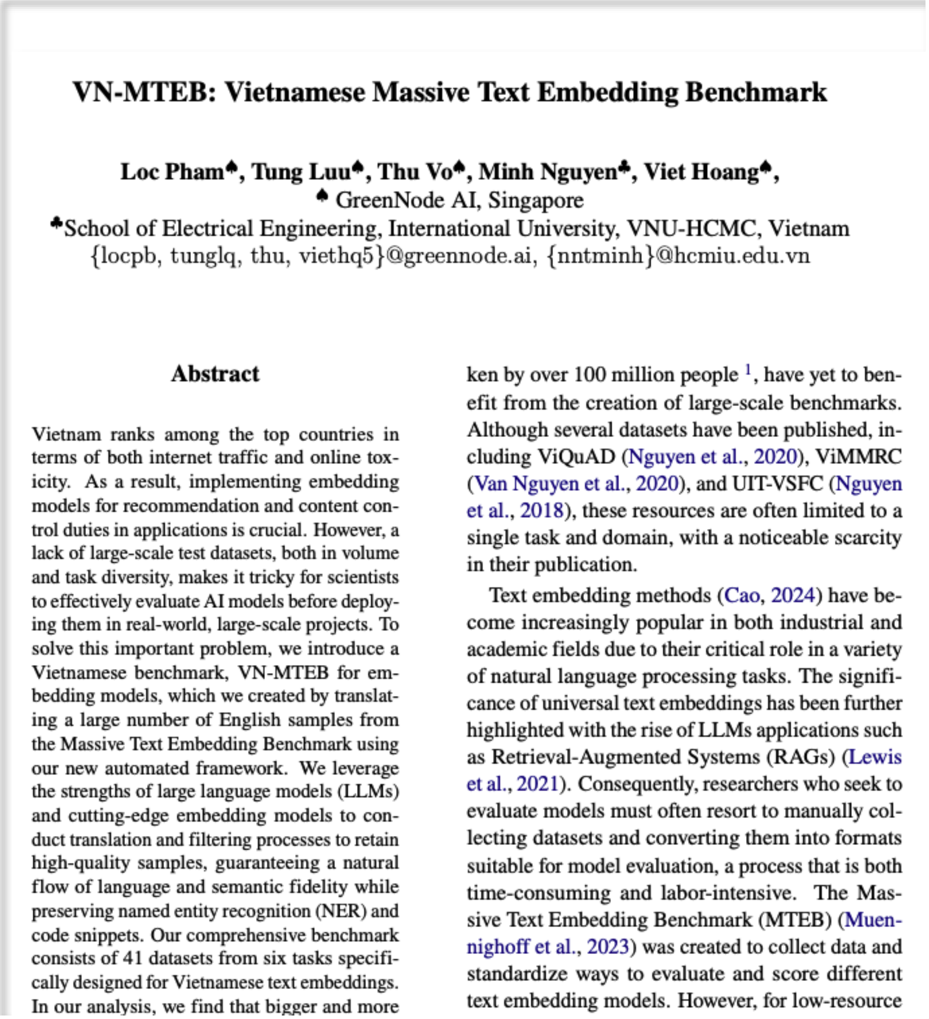
Pham Bao Loc
Senior AI Engineer
locpb@greennode.ai

GreenNode AI Presentation

Training a Vietnamese LLM for Complex Reasoning

Chain-of-Thought Data and Self-Correction Cycles

MTEB



VN-MTEB: Vietnamese Massive Text Embedding Benchmark

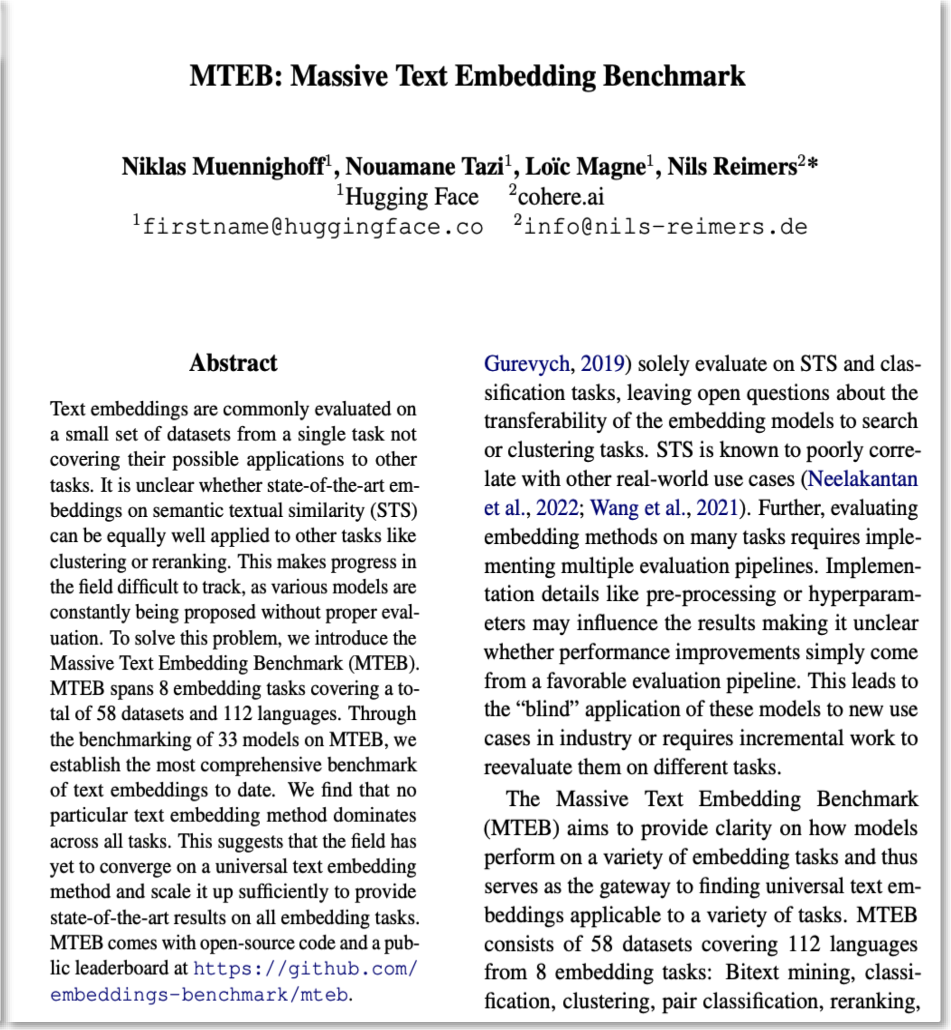
Loc Pham♠, Tung Luu♠, Thu Vo♠, Minh Nguyen♣, Viet Hoang♠,
♠ GreenNode AI, Singapore
♣School of Electrical Engineering, International University, VNU-HCMC, Vietnam
{locpb, tunglq, thu, viethq5}@greennode.ai, {nntminh}@hcmiu.edu.vn

Abstract

Vietnam ranks among the top countries in terms of both internet traffic and online toxicity. As a result, implementing embedding models for recommendation and content control duties in applications is crucial. However, a lack of large-scale test datasets, both in volume and task diversity, makes it tricky for scientists to effectively evaluate AI models before deploying them in real-world, large-scale projects. To solve this important problem, we introduce a Vietnamese benchmark, VN-MTEB for embedding models, which we created by translating a large number of English samples from the Massive Text Embedding Benchmark using our new automated framework. We leverage the strengths of large language models (LLMs) and cutting-edge embedding models to conduct translation and filtering processes to retain high-quality samples, guaranteeing a natural flow of language and semantic fidelity while preserving named entity recognition (NER) and code snippets. Our comprehensive benchmark consists of 41 datasets from six tasks specifically designed for Vietnamese text embeddings. In our analysis, we find that bigger and more

ken by over 100 million people [1], have yet to benefit from the creation of large-scale benchmarks. Although several datasets have been published, including ViQuAD (Nguyen et al., 2020), ViMMRC (Van Nguyen et al., 2020), and UIT-VSFC (Nguyen et al., 2018), these resources are often limited to a single task and domain, with a noticeable scarcity in their publication.

Text embedding methods (Cao, 2024) have become increasingly popular in both industrial and academic fields due to their critical role in a variety of natural language processing tasks. The significance of universal text embeddings has been further highlighted with the rise of LLMs applications such as Retrieval-Augmented Systems (RAGs) (Lewis et al., 2021). Consequently, researchers who seek to evaluate models must often resort to manually collecting datasets and converting them into formats suitable for model evaluation, a process that is both time-consuming and labor-intensive. The Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2023) was created to collect data and standardize ways to evaluate and score different text embedding models. However, for low-resource



MTEB: Massive Text Embedding Benchmark

Niklas Muennighoff[1], Nouamane Tazi[1], Loïc Magne[1], Nils Reimers[2]*
[1]Hugging Face [2]cohere.ai
[1]`firstname@huggingface.co` [2]`info@nils-reimers.de`

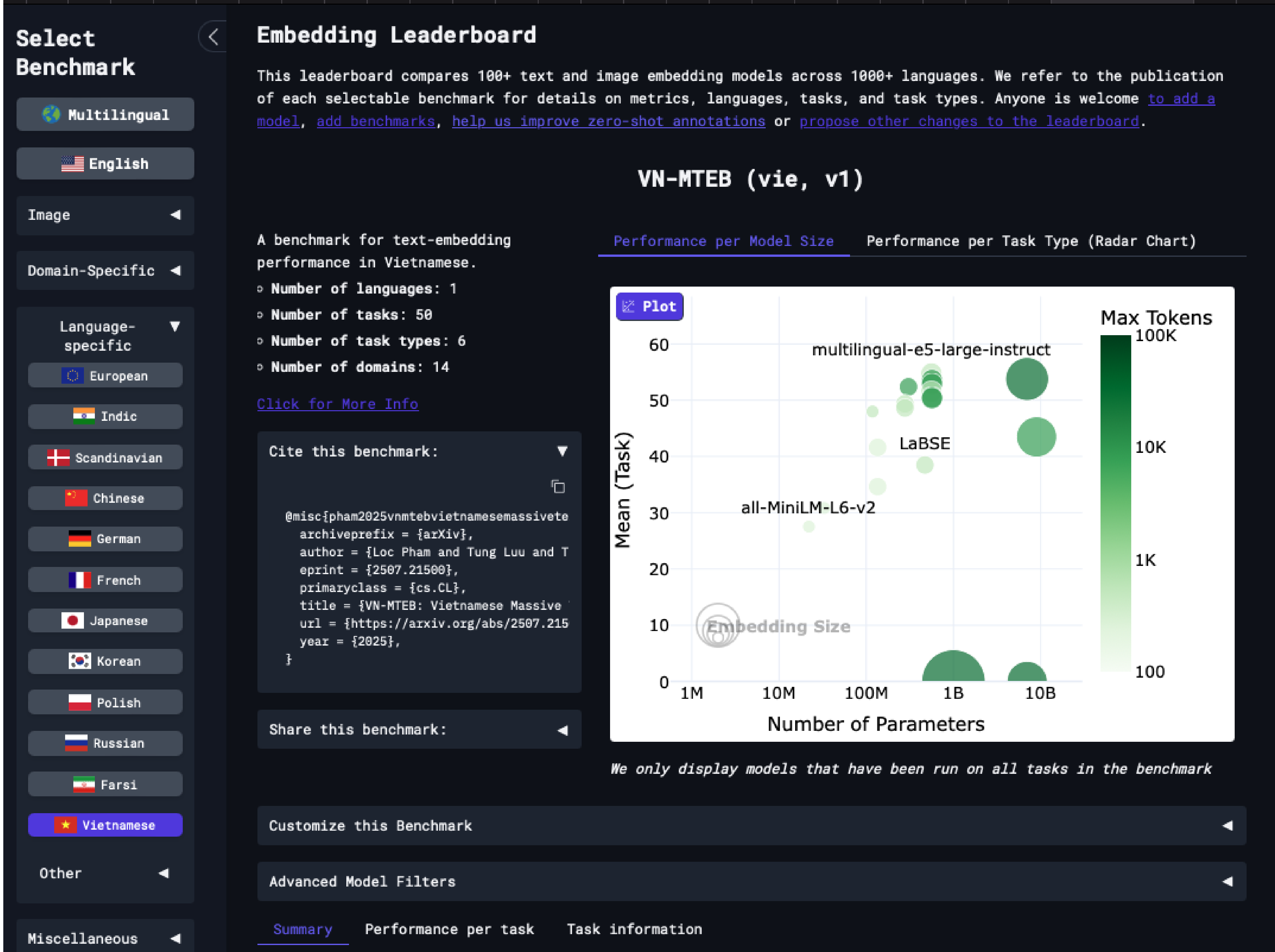
Abstract

Text embeddings are commonly evaluated on a small set of datasets from a single task not covering their possible applications to other tasks. It is unclear whether state-of-the-art embeddings on semantic textual similarity (STS) can be equally well applied to other tasks like clustering or reranking. This makes progress in the field difficult to track, as various models are constantly being proposed without proper evaluation. To solve this problem, we introduce the Massive Text Embedding Benchmark (MTEB). MTEB spans 8 embedding tasks covering a total of 58 datasets and 112 languages. Through the benchmarking of 33 models on MTEB, we establish the most comprehensive benchmark of text embeddings to date. We find that no particular text embedding method dominates across all tasks. This suggests that the field has yet to converge on a universal text embedding method and scale it up sufficiently to provide state-of-the-art results on all embedding tasks. MTEB comes with open-source code and a public leaderboard at `https://github.com/embeddings-benchmark/mteb`.

Gurevych, 2019) solely evaluate on STS and classification tasks, leaving open questions about the transferability of the embedding models to search or clustering tasks. STS is known to poorly correlate with other real-world use cases (Neelakantan et al., 2022; Wang et al., 2021). Further, evaluating embedding methods on many tasks requires implementing multiple evaluation pipelines. Implementation details like pre-processing or hyperparameters may influence the results making it unclear whether performance improvements simply come from a favorable evaluation pipeline. This leads to the “blind” application of these models to new use cases in industry or requires incremental work to reevaluate them on different tasks.

The Massive Text Embedding Benchmark (MTEB) aims to provide clarity on how models perform on a variety of embedding tasks and thus serves as the gateway to finding universal text embeddings applicable to a variety of tasks. MTEB consists of 58 datasets covering 112 languages from 8 embedding tasks: Bitext mining, classification, clustering, pair classification, reranking,

MOTIVATION ↘



Embedding Leaderboard

This leaderboard compares 100+ text and image embedding models across 1000+ languages. We refer to the publication of each selectable benchmark for details on metrics, languages, tasks, and task types. Anyone is welcome to add a model, add benchmarks, help us improve zero-shot annotations or propose other changes to the leaderboard.

VN-MTEB (vie, v1)

A benchmark for text-embedding performance in Vietnamese.

- **Number of languages:** 1
- **Number of tasks:** 50
- **Number of task types:** 6
- **Number of domains:** 14

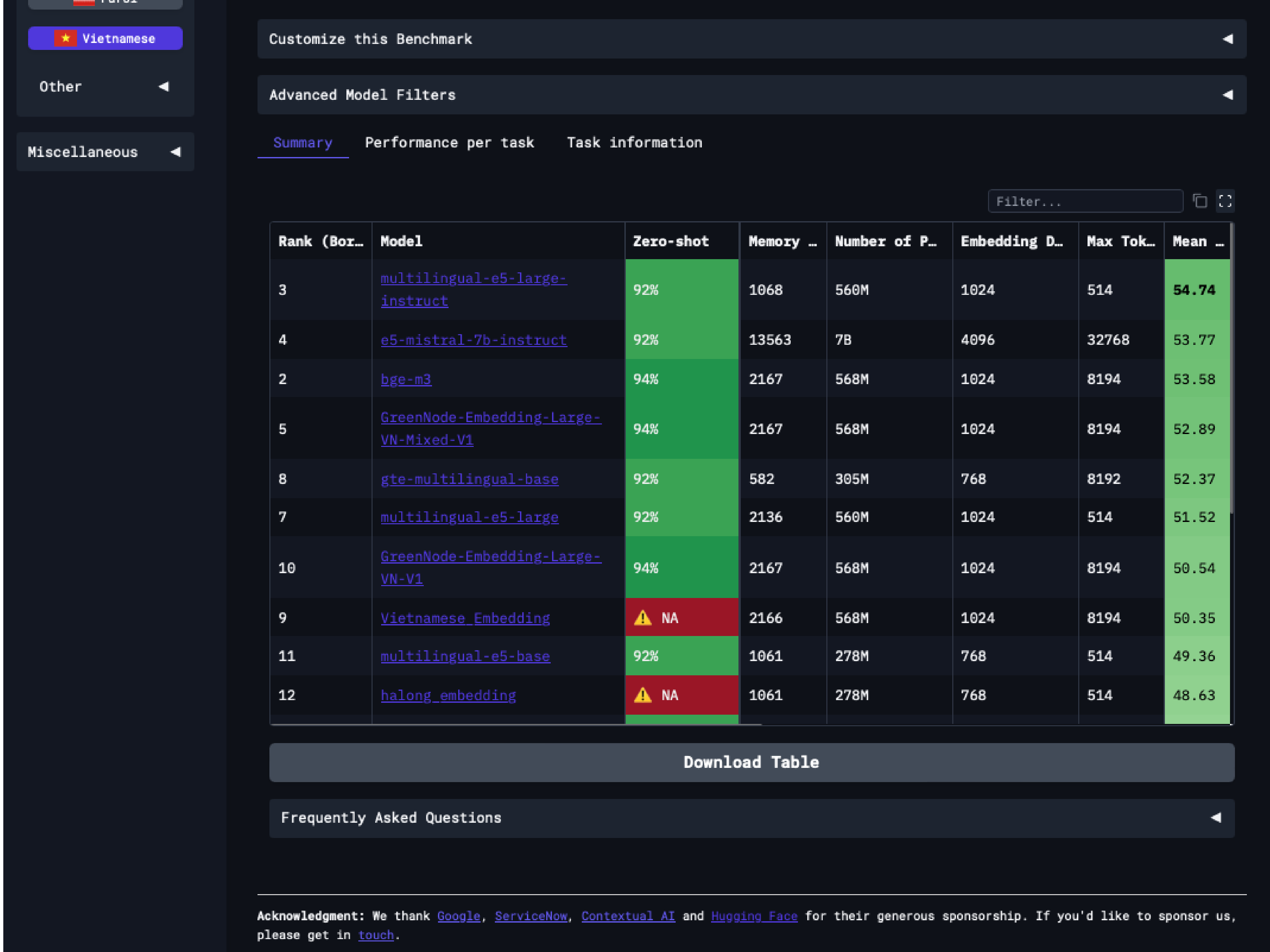
Click for More Info

Cite this benchmark:

```
@misc{pham2025vnmtebvietnamesemassivete
  archiveprefix = {arXiv},
  author = {Loc Pham and Tung Luu and T
  eprint = {2507.21500},
  primaryclass = {cs.CL},
  title = {VN-MTEB: Vietnamese Massive
  url = {https://arxiv.org/abs/2507.215
  year = {2025},
}
```

We only display models that have been run on all tasks in the benchmark

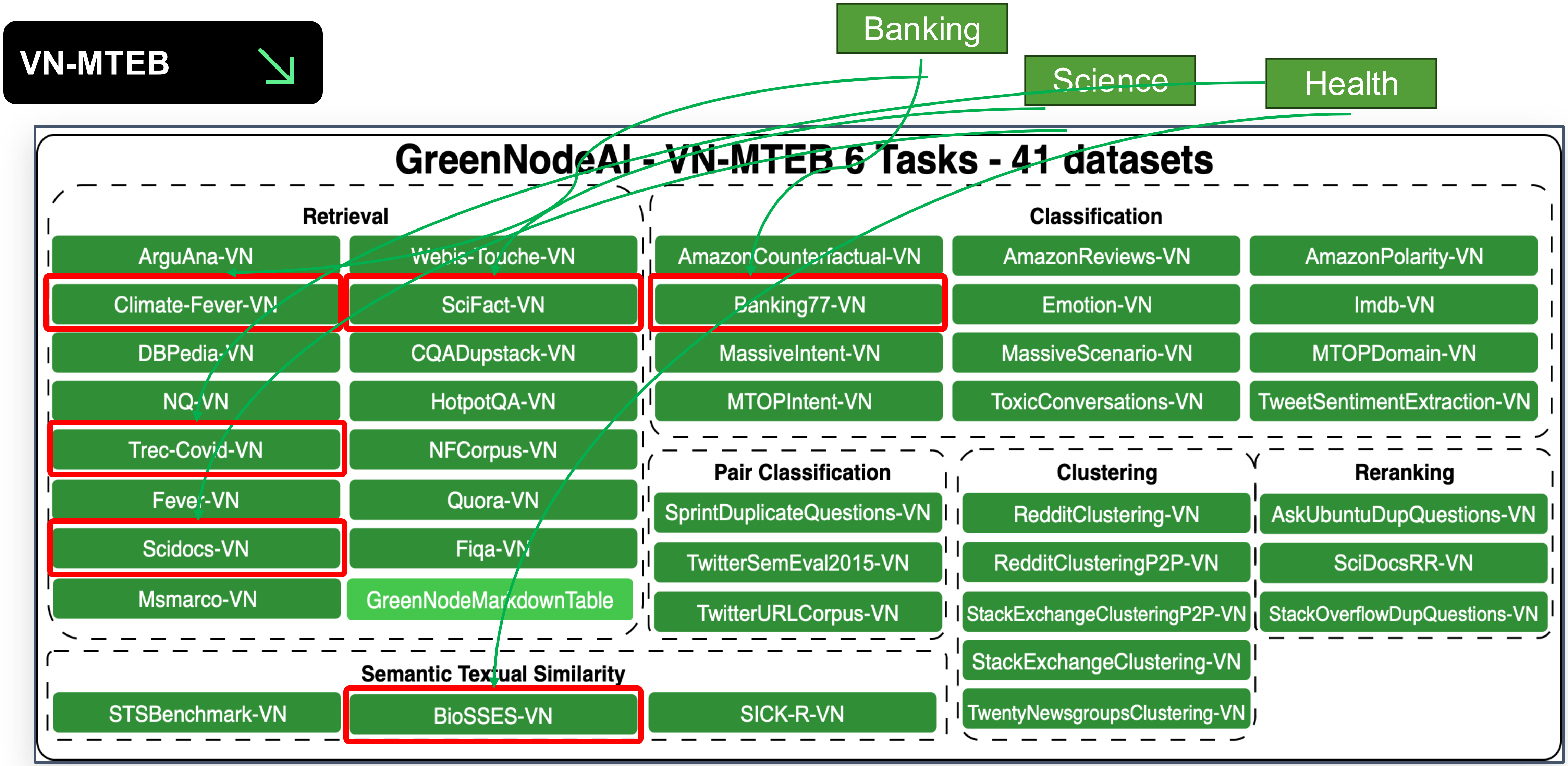
MOTIVATION ↘



Rank (Bor…	Model	Zero-shot	Memory …	Number of P…	Embedding D…	Max Tok…	Mean …
3	multilingual-e5-large-instruct	92%	1068	560M	1024	514	**54.74**
4	e5-mistral-7b-instruct	92%	13563	7B	4096	32768	53.77
2	bge-m3	94%	2167	568M	1024	8194	53.58
5	GreenNode-Embedding-Large-VN-Mixed-V1	94%	2167	568M	1024	8194	52.89
8	gte-multilingual-base	92%	582	305M	768	8192	52.37
7	multilingual-e5-large	92%	2136	560M	1024	514	51.52
10	GreenNode-Embedding-Large-VN-V1	94%	2167	568M	1024	8194	50.54
9	Vietnamese_Embedding	⚠ NA	2166	568M	1024	8194	50.35
11	multilingual-e5-base	92%	1061	278M	768	514	49.36
12	halong_embedding	⚠ NA	1061	278M	768	514	48.63

Acknowledgment: We thank Google, ServiceNow, Contextual AI and Hugging Face for their generous sponsorship. If you'd like to sponsor us, please get in touch.

VN-MTEB ↘



Banking

Science

Health

GreenNodeAI - VN-MTEB 6 Tasks - 41 datasets

Retrieval

- ArguAna-VN
- Climate-Fever-VN
- DBPedia-VN
- NQ-VN
- Trec-Covid-VN
- Fever-VN
- Scidocs-VN
- Msmarco-VN
- Webis-Touche-VN
- SciFact-VN
- CQADupstack-VN
- HotpotQA-VN
- NFCorpus-VN
- Quora-VN
- Fiqa-VN
- GreenNodeMarkdownTable

Classification

- AmazonCounterfactual-VN
- Banking77-VN
- MassiveIntent-VN
- MTOPIntent-VN
- AmazonReviews-VN
- Emotion-VN
- MassiveScenario-VN
- ToxicConversations-VN
- AmazonPolarity-VN
- Imdb-VN
- MTOPDomain-VN
- TweetSentimentExtraction-VN

Pair Classification

- SprintDuplicateQuestions-VN
- TwitterSemEval2015-VN
- TwitterURLCorpus-VN

Clustering

- RedditClustering-VN
- RedditClusteringP2P-VN
- StackExchangeClusteringP2P-VN
- StackExchangeClustering-VN
- TwentyNewsgroupsClustering-VN

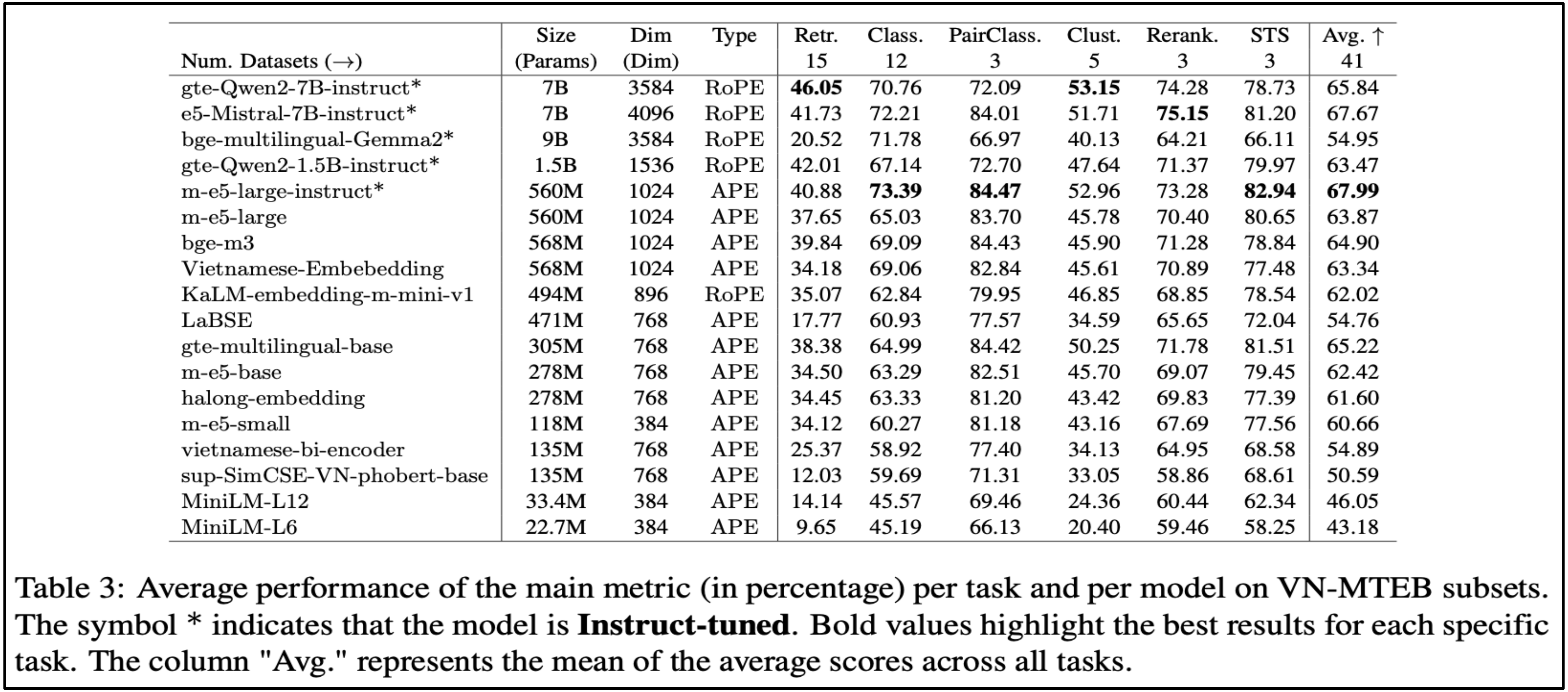
Reranking

- AskUbuntuDupQuestions-VN
- SciDocsRR-VN
- StackOverflowDupQuestions-VN

Semantic Textual Similarity

- STSBenchmark-VN
- BioSSES-VN
- SICK-R-VN

BENCHMARK RESULT & CONCLUSION



Num. Datasets (→)	Size (Params)	Dim (Dim)	Type	Retr. 15	Class. 12	PairClass. 3	Clust. 5	Rerank. 3	STS 3	Avg. ↑ 41
gte-Qwen2-7B-instruct*	7B	3584	RoPE	**46.05**	70.76	72.09	**53.15**	74.28	78.73	65.84
e5-Mistral-7B-instruct*	7B	4096	RoPE	41.73	72.21	84.01	51.71	**75.15**	81.20	67.67
bge-multilingual-Gemma2*	9B	3584	RoPE	20.52	71.78	66.97	40.13	64.21	66.11	54.95
gte-Qwen2-1.5B-instruct*	1.5B	1536	RoPE	42.01	67.14	72.70	47.64	71.37	79.97	63.47
m-e5-large-instruct*	560M	1024	APE	40.88	**73.39**	**84.47**	52.96	73.28	**82.94**	**67.99**
m-e5-large	560M	1024	APE	37.65	65.03	83.70	45.78	70.40	80.65	63.87
bge-m3	568M	1024	APE	39.84	69.09	84.43	45.90	71.28	78.84	64.90
Vietnamese-Embebedding	568M	1024	APE	34.18	69.06	82.84	45.61	70.89	77.48	63.34
KaLM-embedding-m-mini-v1	494M	896	RoPE	35.07	62.84	79.95	46.85	68.85	78.54	62.02
LaBSE	471M	768	APE	17.77	60.93	77.57	34.59	65.65	72.04	54.76
gte-multilingual-base	305M	768	APE	38.38	64.99	84.42	50.25	71.78	81.51	65.22
m-e5-base	278M	768	APE	34.50	63.29	82.51	45.70	69.07	79.45	62.42
halong-embedding	278M	768	APE	34.45	63.33	81.20	43.42	69.83	77.39	61.60
m-e5-small	118M	384	APE	34.12	60.27	81.18	43.16	67.69	77.56	60.66
vietnamese-bi-encoder	135M	768	APE	25.37	58.92	77.40	34.13	64.95	68.58	54.89
sup-SimCSE-VN-phobert-base	135M	768	APE	12.03	59.69	71.31	33.05	58.86	68.61	50.59
MiniLM-L12	33.4M	384	APE	14.14	45.57	69.46	24.36	60.44	62.34	46.05
MiniLM-L6	22.7M	384	APE	9.65	45.19	66.13	20.40	59.46	58.25	43.18

Table 3: Average performance of the main metric (in percentage) per task and per model on VN-MTEB subsets. The symbol * indicates that the model is **Instruct-tuned**. Bold values highlight the best results for each specific task. The column "Avg." represents the mean of the average scores across all tasks.

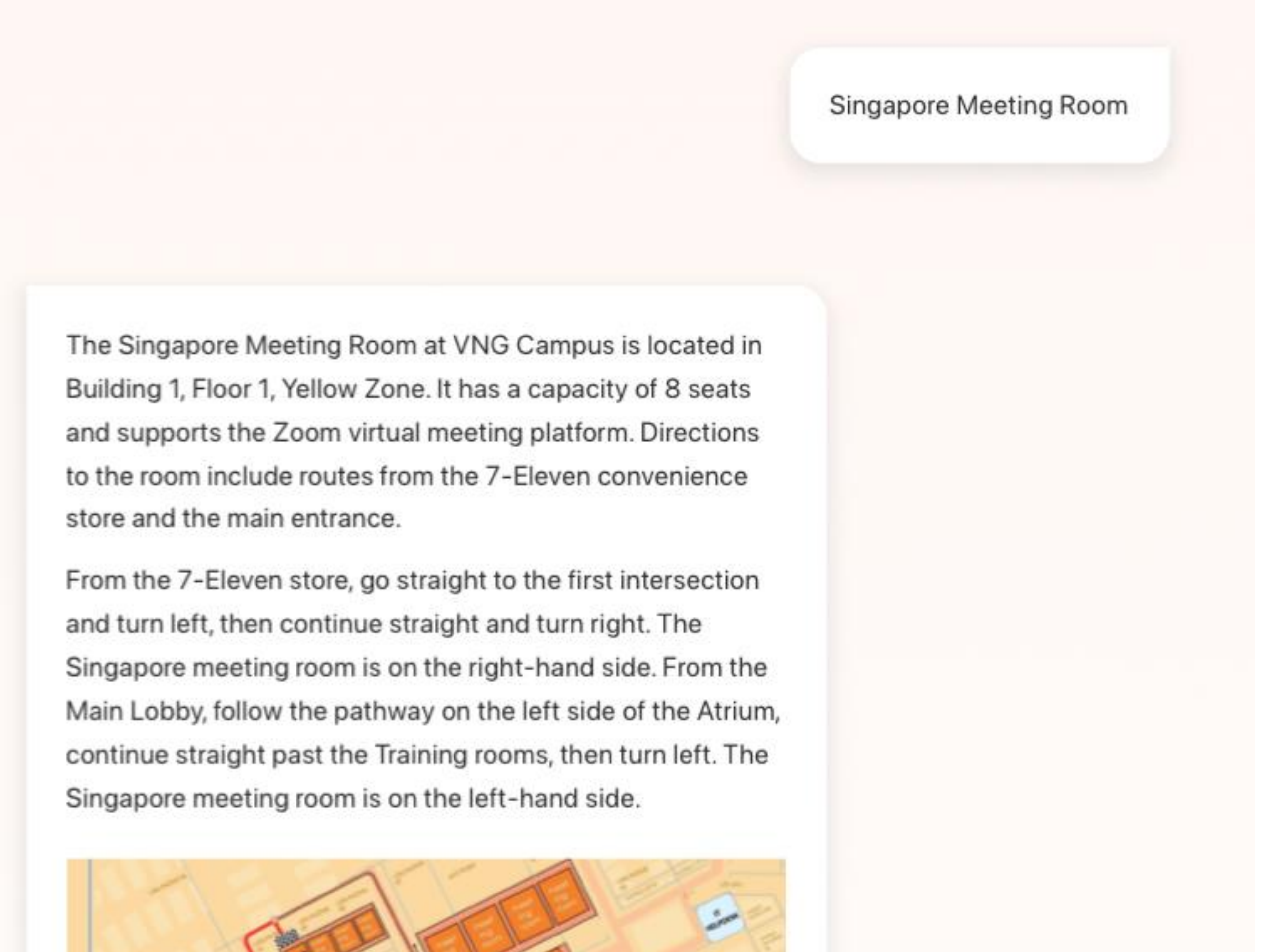
MOTIVATION ↘

AI Chatbot using RAG

Singapore Meeting Room

MOTIVATION

AI Chatbot using RAG



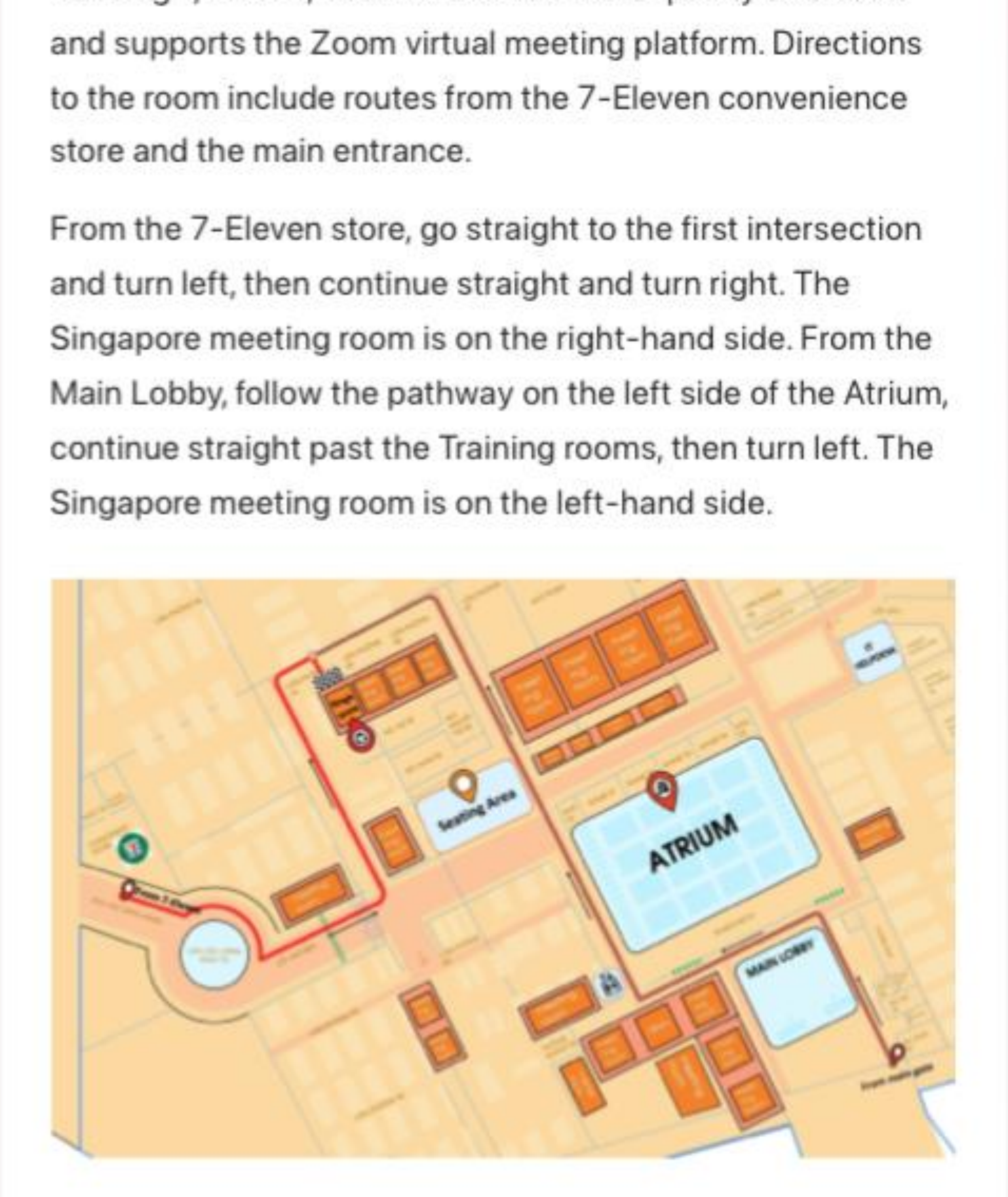
Singapore Meeting Room

The Singapore Meeting Room at VNG Campus is located in Building 1, Floor 1, Yellow Zone. It has a capacity of 8 seats and supports the Zoom virtual meeting platform. Directions to the room include routes from the 7-Eleven convenience store and the main entrance.

From the 7-Eleven store, go straight to the first intersection and turn left, then continue straight and turn right. The Singapore meeting room is on the right-hand side. From the Main Lobby, follow the pathway on the left side of the Atrium, continue straight past the Training rooms, then turn left. The Singapore meeting room is on the left-hand side.

MOTIVATION ↘

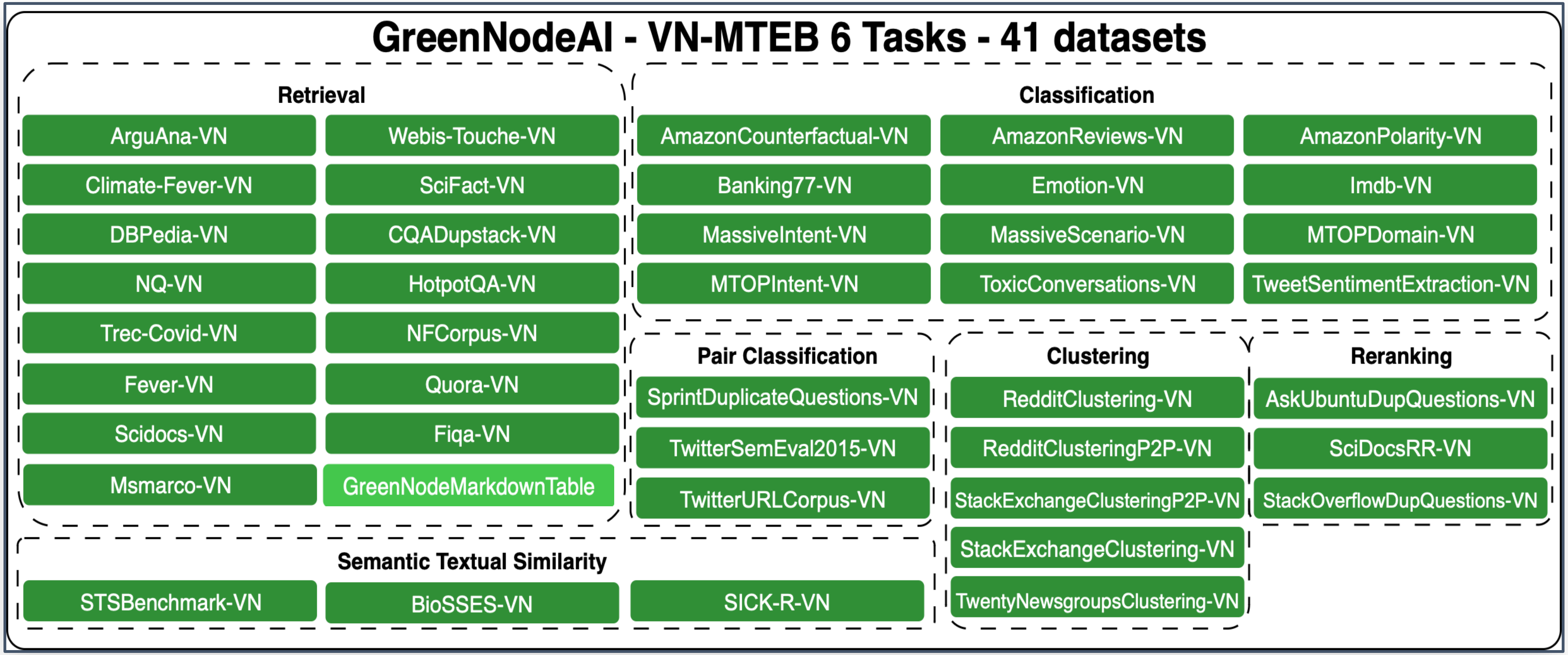
AI Chatbot using RAG



and supports the Zoom virtual meeting platform. Directions to the room include routes from the 7-Eleven convenience store and the main entrance.

From the 7-Eleven store, go straight to the first intersection and turn left, then continue straight and turn right. The Singapore meeting room is on the right-hand side. From the Main Lobby, follow the pathway on the left side of the Atrium, continue straight past the Training rooms, then turn left. The Singapore meeting room is on the left-hand side.

VN-MTEB



GreenNodeAI - VN-MTEB 6 Tasks - 41 datasets

Retrieval

- ArguAna-VN
- Webis-Touche-VN
- Climate-Fever-VN
- SciFact-VN
- DBPedia-VN
- CQADupstack-VN
- NQ-VN
- HotpotQA-VN
- Trec-Covid-VN
- NFCorpus-VN
- Fever-VN
- Quora-VN
- Scidocs-VN
- Fiqa-VN
- Msmarco-VN
- GreenNodeMarkdownTable

Classification

- AmazonCounterfactual-VN
- AmazonReviews-VN
- AmazonPolarity-VN
- Banking77-VN
- Emotion-VN
- Imdb-VN
- MassiveIntent-VN
- MassiveScenario-VN
- MTOPDomain-VN
- MTOPIntent-VN
- ToxicConversations-VN
- TweetSentimentExtraction-VN

Pair Classification

- SprintDuplicateQuestions-VN
- TwitterSemEval2015-VN
- TwitterURLCorpus-VN

Clustering

- RedditClustering-VN
- RedditClusteringP2P-VN
- StackExchangeClusteringP2P-VN
- StackExchangeClustering-VN
- TwentyNewsgroupsClustering-VN

Reranking

- AskUbuntuDupQuestions-VN
- SciDocsRR-VN
- StackOverflowDupQuestions-VN

Semantic Textual Similarity

- STSBenchmark-VN
- BioSSES-VN
- SICK-R-VN

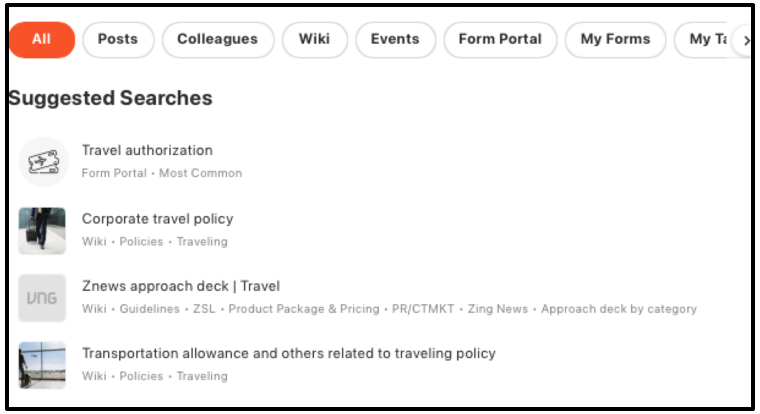
CHALLENGE ↘

GreenNodeMarkdownTable

File format: docx, pdf, csv.

Document format: plain text, paragraph, tables

Document language: Vietnamese, English



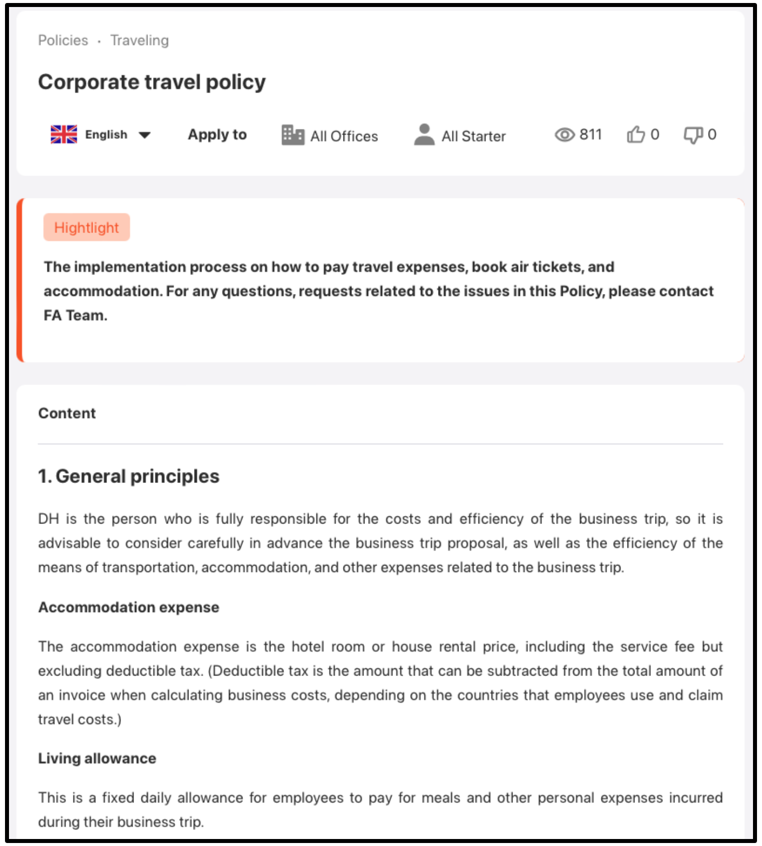
Suggested Searches

Travel authorization
Form Portal • Most Common

Corporate travel policy
Wiki • Policies • Traveling

Znews approach deck | Travel
Wiki • Guidelines • ZSL • Product Package & Pricing • PR/CTMKT • Zing News • Approach deck by category

Transportation allowance and others related to traveling policy
Wiki • Policies • Traveling



Corporate travel policy

The implementation process on how to pay travel expenses, book air tickets, and accommodation. For any questions, requests related to the issues in this Policy, please contact FA Team.

Content

1. General principles

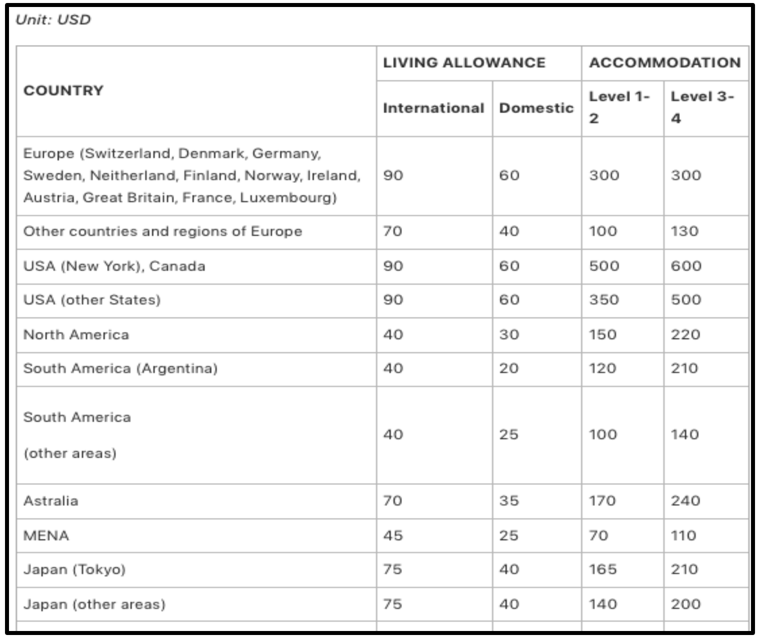
DH is the person who is fully responsible for the costs and efficiency of the business trip, so it is advisable to consider carefully in advance the business trip proposal, as well as the efficiency of the means of transportation, accommodation, and other expenses related to the business trip.

Accommodation expense

The accommodation expense is the hotel room or house rental price, including the service fee but excluding deductible tax. (Deductible tax is the amount that can be subtracted from the total amount of an invoice when calculating business costs, depending on the countries that employees use and claim travel costs.)

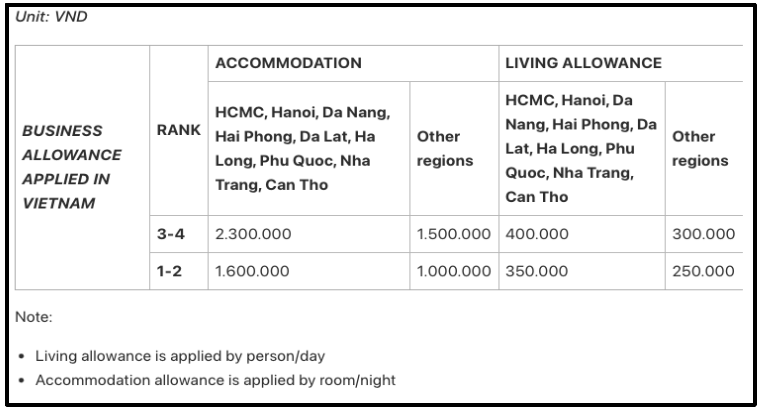
Living allowance

This is a fixed daily allowance for employees to pay for meals and other personal expenses incurred during their business trip.



Unit: USD

COUNTRY	LIVING ALLOWANCE		ACCOMMODATION	
	International	Domestic	Level 1-2	Level 3-4
Europe (Switzerland, Denmark, Germany, Sweden, Neitherland, Finland, Norway, Ireland, Austria, Great Britain, France, Luxembourg)	90	60	300	300
Other countries and regions of Europe	70	40	100	130
USA (New York), Canada	90	60	500	600
USA (other States)	90	60	350	500
North America	40	30	150	220
South America (Argentina)	40	20	120	210
South America (other areas)	40	25	100	140
Astralia	70	35	170	240
MENA	45	25	70	110
Japan (Tokyo)	75	40	165	210
Japan (other areas)	75	40	140	200



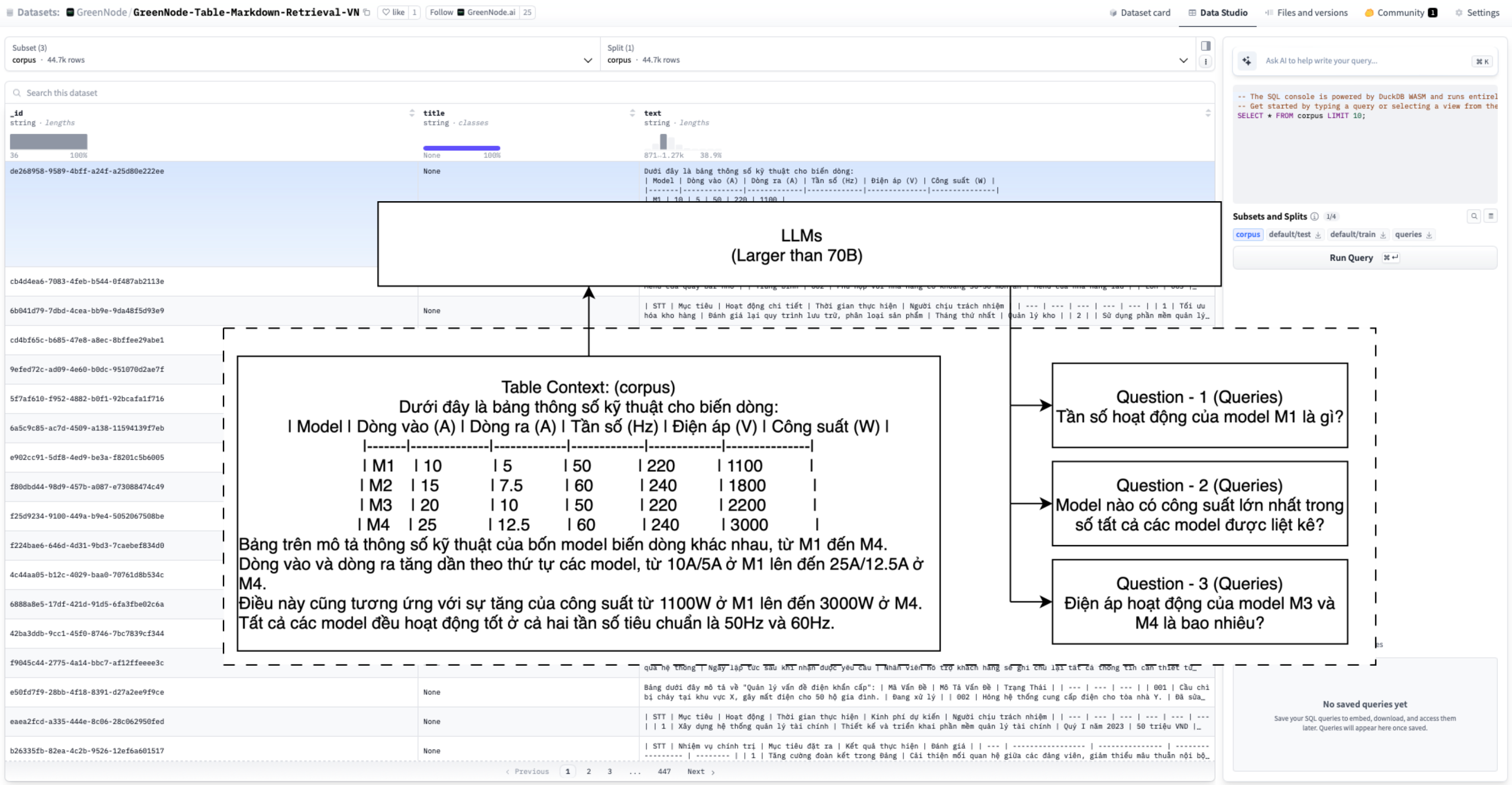
Unit: VND

BUSINESS ALLOWANCE APPLIED IN VIETNAM	RANK	ACCOMMODATION		LIVING ALLOWANCE	
		HCMC, Hanoi, Da Nang, Hai Phong, Da Lat, Ha Long, Phu Quoc, Nha Trang, Can Tho	Other regions	HCMC, Hanoi, Da Nang, Hai Phong, Da Lat, Ha Long, Phu Quoc, Nha Trang, Can Tho	Other regions
	3-4	2.300.000	1.500.000	400.000	300.000
	1-2	1.600.000	1.000.000	350.000	250.000

Note:

- Living allowance is applied by person/day
- Accommodation allowance is applied by room/night

SAMPLE DATA



```
-- The SQL console is powered by DuckDB WASM and runs entirel
-- Get started by typing a query or selecting a view from the
SELECT * FROM corpus LIMIT 10;
```

_id string · lengths 36 100%	title string · classes None 100%	text string · lengths 871↔1.27k 38.9%
de268958-9589-4bff-a24f-a25d80e222ee	None	Dưới đây là bảng thông số kỹ thuật cho biến dòng: | Model | Dòng vào (A) | Dòng ra (A) | Tần số (Hz) | Điện áp (V) | Công suất (W) | |-------|-------------|------------|------------|-------------|--------------| | M1 | 10 | 5 | 50 | 220 | 1100 |
cb4d4ea6-7083-4feb-b544-0f487ab2113e		
6b041d79-7dbd-4cea-bb9e-9da48f5d93e9	None	| STT | Mục tiêu | Hoạt động chi tiết | Thời gian thực hiện | Người chịu trách nhiệm | | --- | --- | --- | --- | --- | | 1 | Tối ưu hóa kho hàng | Đánh giá lại quy trình lưu trữ, phân loại sản phẩm | Tháng thứ nhất | Quản lý kho | | 2 | | Sử dụng phần mềm quản lý…
cd4bf65c-b685-47e8-a8ec-8bffee29abe1		
9efed72c-ad09-4e60-b0dc-951070d2ae7f		
5f7af610-f952-4882-b0f1-92bcafa1f716		
6a5c9c85-ac7d-4509-a138-11594139f7eb		
e902cc91-5df8-4ed9-be3a-f8201c5b6005		
f80dbd44-98d9-457b-a087-e73088474c49		
f25d9234-9100-449a-b9e4-5052067508be		
f224bae6-646d-4d31-9bd3-7caebef834d0		
4c44aa05-b12c-4029-baa0-70761d8b534c		
6888a8e5-17df-421d-91d5-6fa3fbe02c6a		
42ba3ddb-9cc1-45f0-8746-7bc7839cf344		
f9045c44-2775-4a14-bbc7-af12ffeeee3c		quả hệ thống | Ngay lập tức sau khi nhận được yêu cầu | Nhân viên hỗ trợ khách hàng sẽ ghi chú lại tất cả thông tin cần thiết từ…
e50fd7f9-28bb-4f18-8391-d27a2ee9f9ce	None	Bảng dưới đây mô tả về "Quản lý vấn đề điện khẩn cấp": | Mã Vấn Đề | Mô Tả Vấn Đề | Trạng Thái | | --- | --- | --- | | 001 | Cầu chì bị cháy tại khu vực X, gây mất điện cho 50 hộ gia đình. | Đang xử lý | | 002 | Hỏng hệ thống cung cấp điện cho tòa nhà Y. | Đã sửa…
eaea2fcd-a335-444e-8c06-28c062950fed	None	| STT | Mục tiêu | Hoạt động | Thời gian thực hiện | Kinh phí dự kiến | Người chịu trách nhiệm | | --- | --- | --- | --- | --- | --- | | 1 | Xây dựng hệ thống quản lý tài chính | Thiết kế và triển khai phần mềm quản lý tài chính | Quý I năm 2023 | 50 triệu VND |…
b26335fb-82ea-4c2b-9526-12ef6a601517	None	| STT | Nhiệm vụ chính trị | Mục tiêu đặt ra | Kết quả thực hiện | Đánh giá | | --- | ---------------- | -------------- | ----------------- | -------- | | 1 | Tăng cường đoàn kết trong Đảng | Cải thiện mối quan hệ giữa các đảng viên, giảm thiểu mâu thuẫn nội bộ…

LLMs
(Larger than 70B)

Table Context: (corpus)
Dưới đây là bảng thông số kỹ thuật cho biến dòng:

Model	Dòng vào (A)	Dòng ra (A)	Tần số (Hz)	Điện áp (V)	Công suất (W)
M1	10	5	50	220	1100
M2	15	7.5	60	240	1800
M3	20	10	50	220	2200
M4	25	12.5	60	240	3000

Bảng trên mô tả thông số kỹ thuật của bốn model biến dòng khác nhau, từ M1 đến M4.
Dòng vào và dòng ra tăng dần theo thứ tự các model, từ 10A/5A ở M1 lên đến 25A/12.5A ở M4.
Điều này cũng tương ứng với sự tăng của công suất từ 1100W ở M1 lên đến 3000W ở M4.
Tất cả các model đều hoạt động tốt ở cả hai tần số tiêu chuẩn là 50Hz và 60Hz.

Question - 1 (Queries)
Tần số hoạt động của model M1 là gì?

Question - 2 (Queries)
Model nào có công suất lớn nhất trong số tất cả các model được liệt kê?

Question - 3 (Queries)
Điện áp hoạt động của model M3 và M4 là bao nhiêu?

GreenMind NIM for Enterprise Data flywheel

Presented by

Vo Trong Thu
Head of AI
thu@greennode.ai

Pham Bao Loc
Senior AI Engineer
locpb@greennode.ai

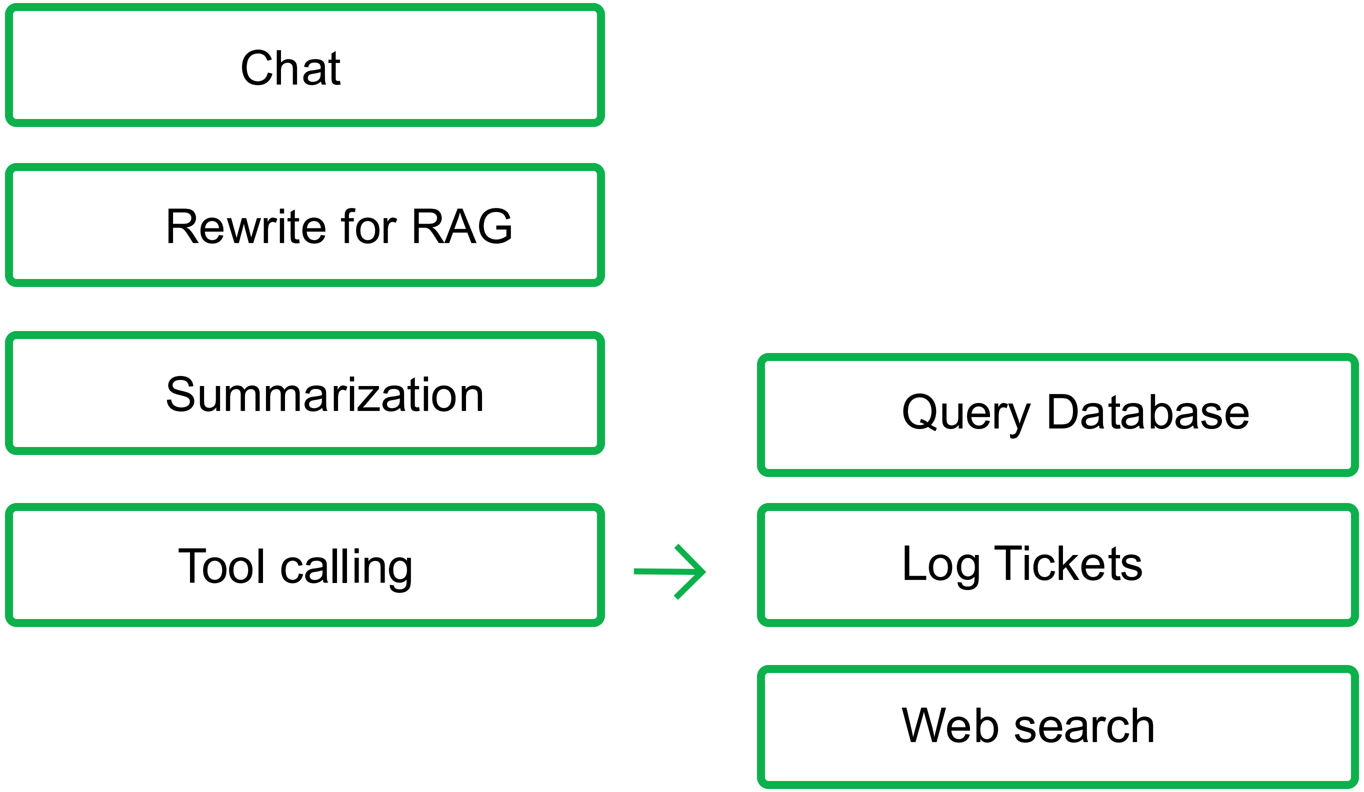
GreenNode AI Presentation

Training a Vietnamese LLM for Complex Reasoning
Chain-of-Thought Data and Self-Correction Cycles

WHAT IS DATA FLYWHEEL? ↘

LLM

High cost, high “accuracy”



Chat

Rewrite for RAG

Summarization

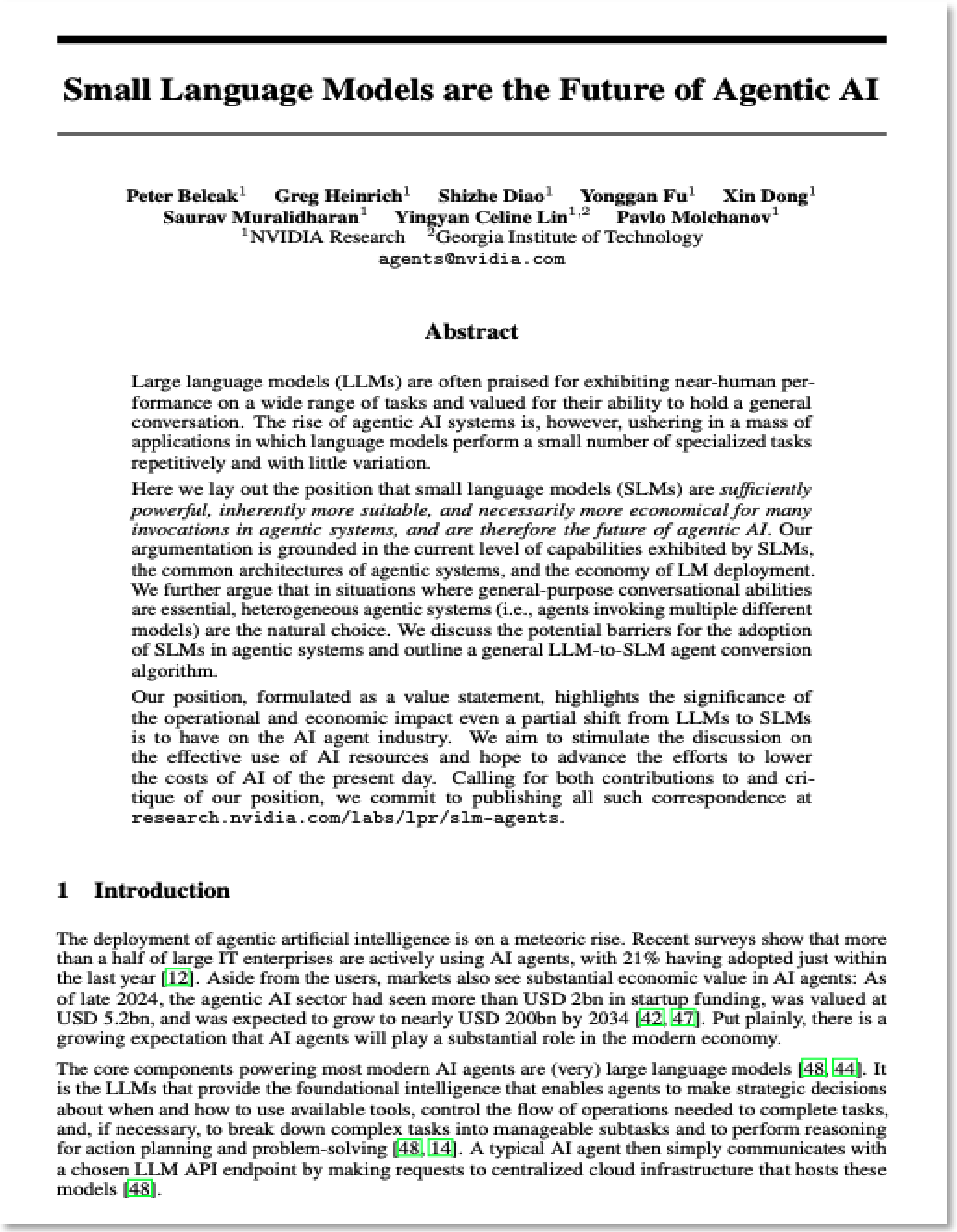
Tool calling →

Query Database

Log Tickets

Web search

NVIDIA LATEST PAPER ↘



Small Language Models are the Future of Agentic AI

Peter Belcak[1] **Greg Heinrich**[1] **Shizhe Diao**[1] **Yonggan Fu**[1] **Xin Dong**[1]
Saurav Muralidharan[1] **Yingyan Celine Lin**[1,2] **Pavlo Molchanov**[1]
[1]NVIDIA Research [2]Georgia Institute of Technology
`agents@nvidia.com`

Abstract

Large language models (LLMs) are often praised for exhibiting near-human performance on a wide range of tasks and valued for their ability to hold a general conversation. The rise of agentic AI systems is, however, ushering in a mass of applications in which language models perform a small number of specialized tasks repetitively and with little variation.

Here we lay out the position that small language models (SLMs) are *sufficiently powerful, inherently more suitable, and necessarily more economical for many invocations in agentic systems, and are therefore the future of agentic AI*. Our argumentation is grounded in the current level of capabilities exhibited by SLMs, the common architectures of agentic systems, and the economy of LM deployment. We further argue that in situations where general-purpose conversational abilities are essential, heterogeneous agentic systems (i.e., agents invoking multiple different models) are the natural choice. We discuss the potential barriers for the adoption of SLMs in agentic systems and outline a general LLM-to-SLM agent conversion algorithm.

Our position, formulated as a value statement, highlights the significance of the operational and economic impact even a partial shift from LLMs to SLMs is to have on the AI agent industry. We aim to stimulate the discussion on the effective use of AI resources and hope to advance the efforts to lower the costs of AI of the present day. Calling for both contributions to and critique of our position, we commit to publishing all such correspondence at `research.nvidia.com/labs/lpr/slm-agents`.

1 Introduction

The deployment of agentic artificial intelligence is on a meteoric rise. Recent surveys show that more than a half of large IT enterprises are actively using AI agents, with 21% having adopted just within the last year [12]. Aside from the users, markets also see substantial economic value in AI agents: As of late 2024, the agentic AI sector had seen more than USD 2bn in startup funding, was valued at USD 5.2bn, and was expected to grow to nearly USD 200bn by 2034 [42, 47]. Put plainly, there is a growing expectation that AI agents will play a substantial role in the modern economy.

The core components powering most modern AI agents are (very) large language models [48, 44]. It is the LLMs that provide the foundational intelligence that enables agents to make strategic decisions about when and how to use available tools, control the flow of operations needed to complete tasks, and, if necessary, to break down complex tasks into manageable subtasks and to perform reasoning for action planning and problem-solving [48, 14]. A typical AI agent then simply communicates with a chosen LLM API endpoint by making requests to centralized cloud infrastructure that hosts these models [48].

WHAT IS DATA FLYWHEEL? ↘

Can we "train" the SLM to be as strong as the LLM?

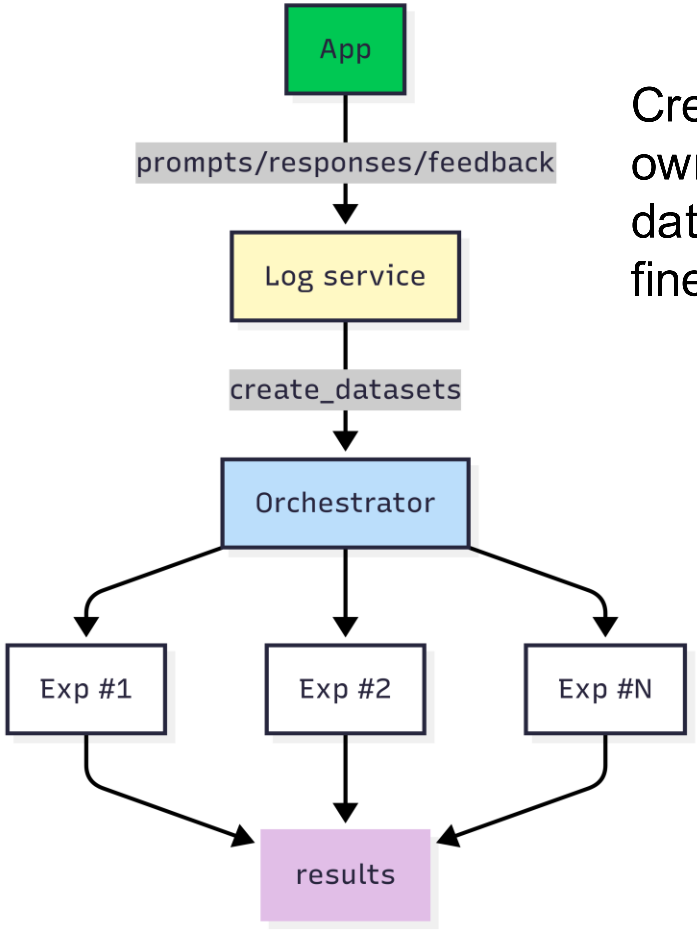
LLM

High cost, high "accuracy"

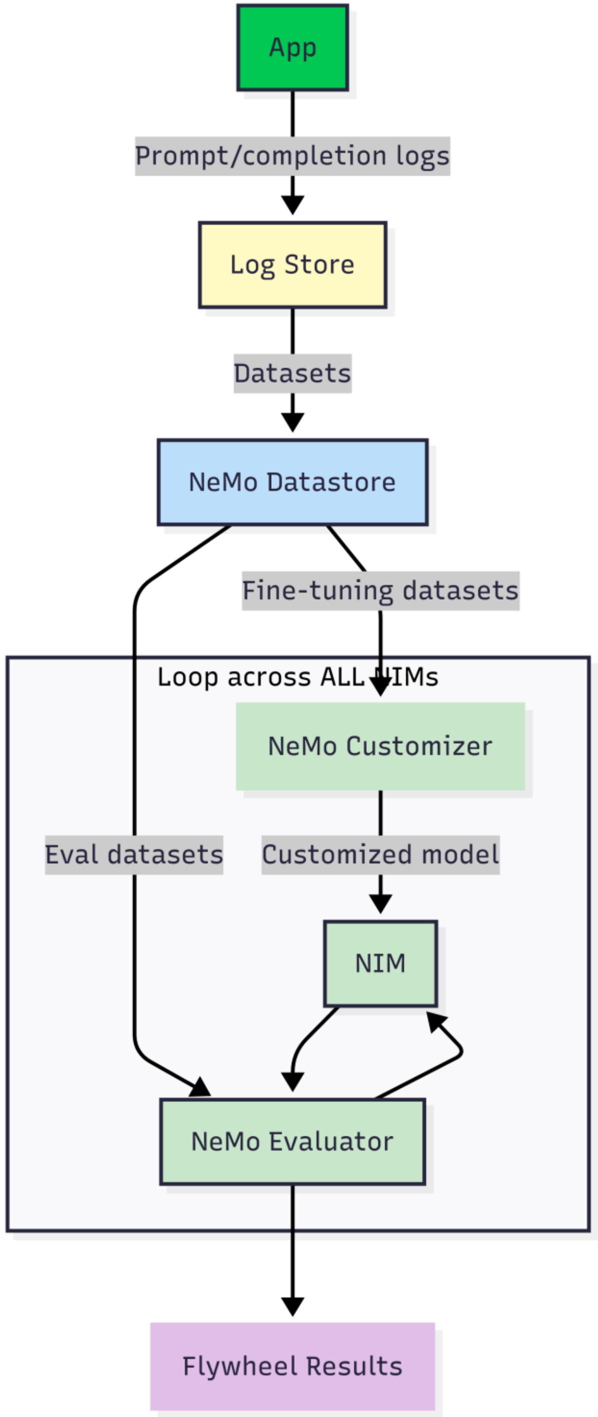
SLM

Low cost, low "accuracy"

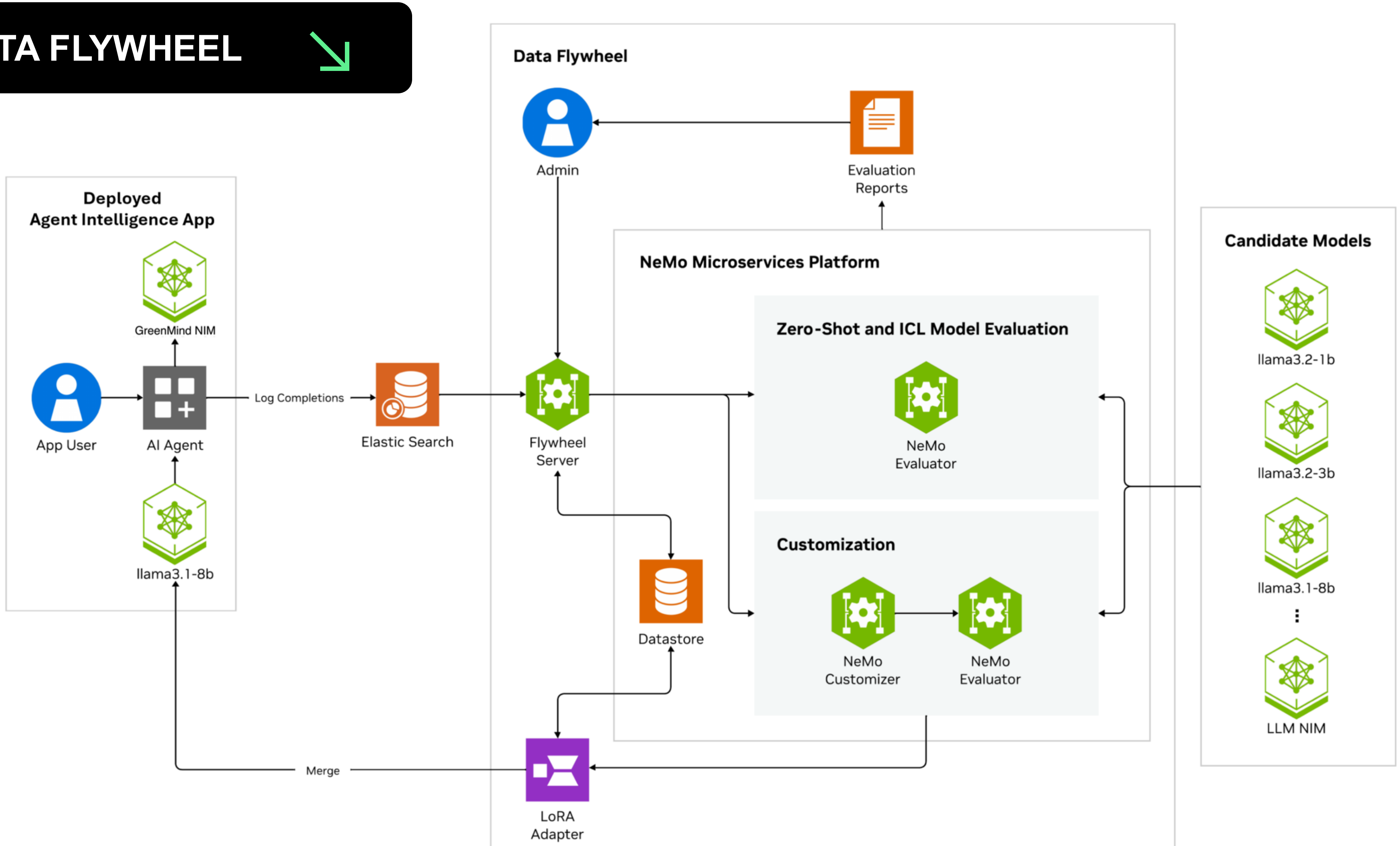
WHAT IS DATA FLYWHEEL?


App
prompts/responses/feedback
Log service
create_datasets
Orchestrator
Exp #1
Exp #2
Exp #N
results

Create, using your own collected dataset to finetuning the LLM


App
Prompt/completion logs
Log Store
Datasets
NeMo Datastore
Fine-tuning datasets
Loop across ALL NIMs
NeMo Customizer
Eval datasets
Customized model
NIM
NeMo Evaluator
Flywheel Results

NVIDIA DATA FLYWHEEL



GREENNODE

One-stop AI Solution

designed for your business success

With experience serving thousands of global customers across various industries, GreenNode is confident in delivering AI solutions tailored to your business needs. From infrastructure to applications, we ensure optimal cost and deployment time for maximum efficiency.

GreenNode
in South-East Asia

AI Cloud Infrastructure is currently located in Bangkok, Hanoi and Ho Chi Minh City

#AI #Cloud #Security

Serving enterprise clients & AI startups in US, EMEA, APAC

AI Cloud Infrastructure is currently located in Bangkok, Hanoi and Ho Chi Minh City

Compliant with ISO 27000, PCI DSS, and TVRA standards

Thank You ↘ For Listening