Towards Khmer-Native AI Models

Rina Buoy, PhD

Techo Startup Center, Ministry of Economy and Finance, Cambodia
Osaka Metropolitan University, Japan

Disclaimer: any opinions or viewpoints are solely mine. All slides are human-generated.

Profile & Experience

Teaching Experience

Courses related to Artificial Intelligence for Master's degree programs at AUPP, UP, CamTech, NUM, and RUPP

Work Experience

2020–Present: Senior AI Researcher at Techo Startup Center (TSC), Ministry of Economy and Finance (MEF)

2012–2020: Head of Research & Development (R&D) at Three60 Energy (Malaysia)

Education Background

Ph.D.: Artificial Intelligence (AI) from Osaka Metropolitan University (OMU)

Master's: Geoscience from Universiti Teknologi PETRONAS and IT Engineering from RUPP

Bachelor's: Petroleum Engineering from Universiti Teknologi PETRONAS



@RINABUOY

https://academy.rinabuoy.com

FB: Dr Rina Buoy

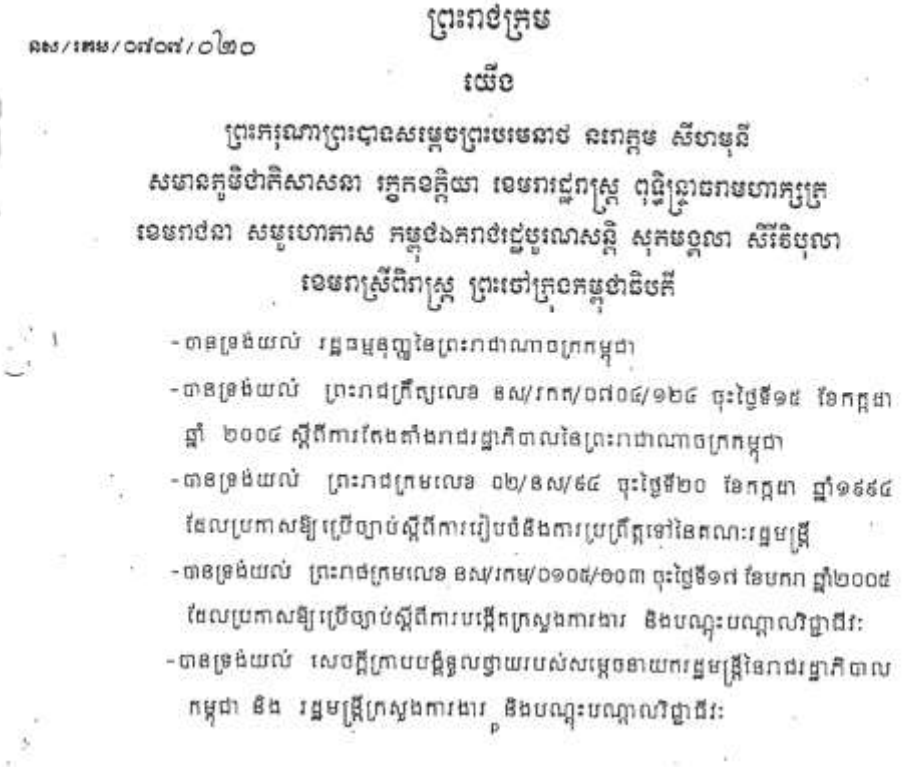
Artificial Intelligence (AI)

- LLMs, AI chatbots, and AI agents took over the world



- Some of the frontier models (particularly proprietary) have demonstrated remarkable performance with Khmer language.

For Example, Khmer Document Parsing (Gemini)



នស/រកម/០៧០៧/០២០

ព្រះរាជក្រម

យើង

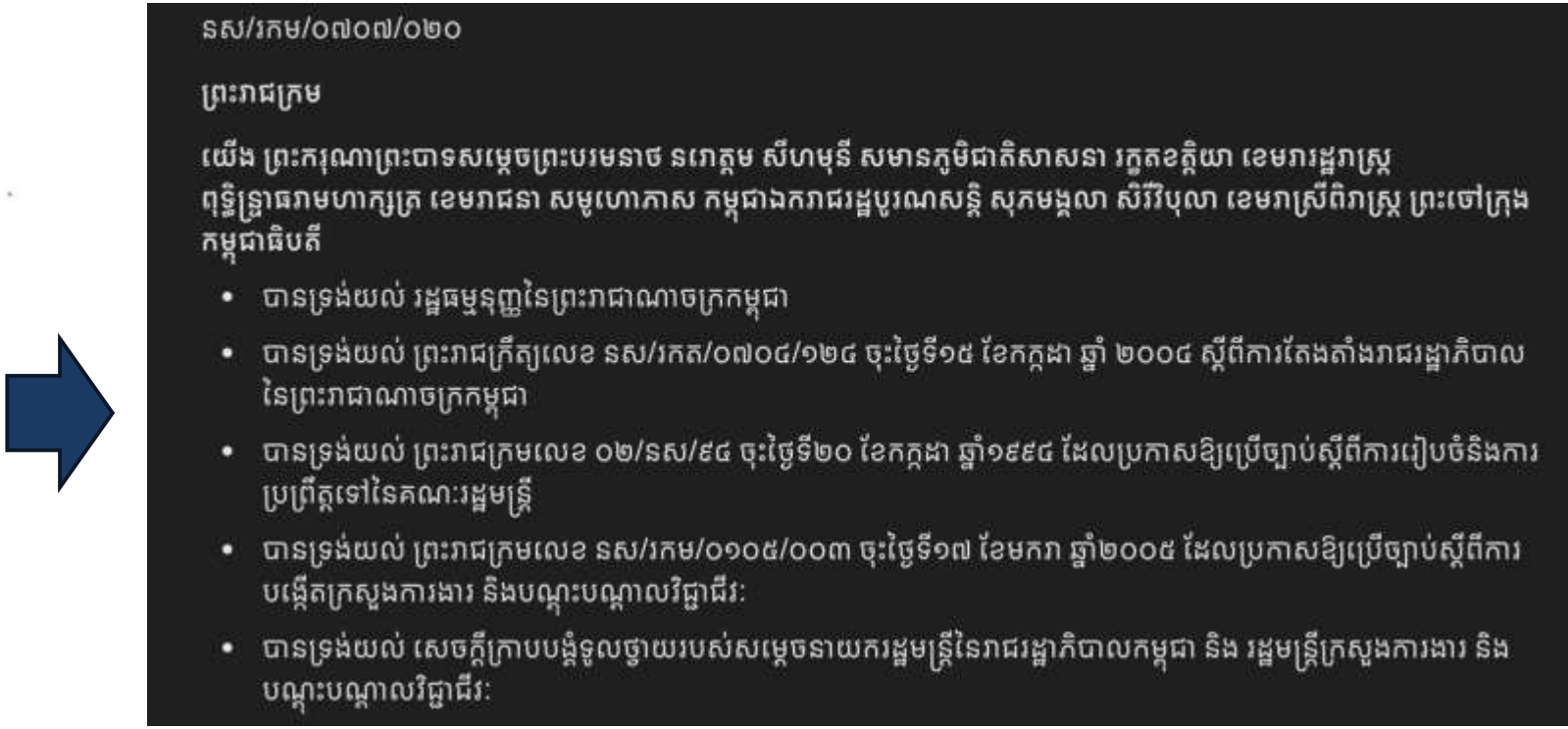
ព្រះករុណាព្រះបាទសម្តេចព្រះបរមនាថ នរោត្តម សីហមុនី

សមានភូមិជាតិសាសនា រក្ខតខត្តិយា ខេមរារដ្ឋរាស្ត្រ ពុទ្ធិន្ទ្រាធរាមហាក្សត្រ

ខេមរាជនា សមូហោភាស កម្ពុជឯករាជរដ្ឋបូរណាសន្តិ សុភមង្គលា សិរីវិបុលា

ខេមរាស្រីពិរាស្ត្រ ព្រះចៅក្រុងកម្ពុជាធិបតី

- បានទ្រង់យល់ រដ្ឋធម្មនុញ្ញនៃព្រះរាជាណាចក្រកម្ពុជា
- បានទ្រង់យល់ ព្រះរាជក្រឹត្យលេខ នស/រកត/០៧០៤/១២៤ ចុះថ្ងៃទី១៤ ខែកក្កដា ឆ្នាំ ២០០៤ ស្តីពីការតែងតាំងរាជរដ្ឋាភិបាលនៃព្រះរាជាណាចក្រកម្ពុជា
- បានទ្រង់យល់ ព្រះរាជក្រមលេខ ០២/នស/៩៤ ចុះថ្ងៃទី២០ ខែកក្កដា ឆ្នាំ១៩៩៤ ដែលប្រកាសឱ្យប្រើច្បាប់ស្តីពីការរៀបចំនិងការប្រព្រឹត្តទៅនៃគណៈរដ្ឋមន្ត្រី
- បានទ្រង់យល់ ព្រះរាជក្រមលេខ នស/រកម/០១០៥/០០៣ ចុះថ្ងៃទី១៧ ខែមករា ឆ្នាំ២០០៥ ដែលប្រកាសឱ្យប្រើច្បាប់ស្តីពីការបង្កើតក្រសួងការងារ និងបណ្តុះបណ្តាលវិជ្ជាជីវៈ
- បានទ្រង់យល់ សេចក្តីក្រាបបង្គំទូលថ្វាយរបស់សម្តេចនាយករដ្ឋមន្ត្រីនៃរាជរដ្ឋាភិបាលកម្ពុជា និង រដ្ឋមន្ត្រីក្រសួងការងារ និងបណ្តុះបណ្តាលវិជ្ជាជីវៈ



នស/រកម/០៧០៧/០២០

ព្រះរាជក្រម

យើង ព្រះករុណាព្រះបាទសម្តេចព្រះបរមនាថ នរោត្តម សីហមុនី សមានភូមិជាតិសាសនា រក្ខតខត្តិយា ខេមរារដ្ឋរាស្ត្រ ពុទ្ធិន្ទ្រាធរាមហាក្សត្រ ខេមរាជនា សមូហោភាស កម្ពុជាឯករាជរដ្ឋបូរណាសន្តិ សុភមង្គលា សិរីវិបុលា ខេមរាស្រីពិរាស្ត្រ ព្រះចៅក្រុងកម្ពុជាធិបតី

- បានទ្រង់យល់ រដ្ឋធម្មនុញ្ញនៃព្រះរាជាណាចក្រកម្ពុជា
- បានទ្រង់យល់ ព្រះរាជក្រឹត្យលេខ នស/រកត/០៧០៤/១២៤ ចុះថ្ងៃទី១៤ ខែកក្កដា ឆ្នាំ ២០០៤ ស្តីពីការតែងតាំងរាជរដ្ឋាភិបាលនៃព្រះរាជាណាចក្រកម្ពុជា
- បានទ្រង់យល់ ព្រះរាជក្រមលេខ ០២/នស/៩៤ ចុះថ្ងៃទី២០ ខែកក្កដា ឆ្នាំ១៩៩៤ ដែលប្រកាសឱ្យប្រើច្បាប់ស្តីពីការរៀបចំនិងការប្រព្រឹត្តទៅនៃគណៈរដ្ឋមន្ត្រី
- បានទ្រង់យល់ ព្រះរាជក្រមលេខ នស/រកម/០១០៥/០០៣ ចុះថ្ងៃទី១៧ ខែមករា ឆ្នាំ២០០៥ ដែលប្រកាសឱ្យប្រើច្បាប់ស្តីពីការបង្កើតក្រសួងការងារ និងបណ្តុះបណ្តាលវិជ្ជាជីវៈ
- បានទ្រង់យល់ សេចក្តីក្រាបបង្គំទូលថ្វាយរបស់សម្តេចនាយករដ្ឋមន្ត្រីនៃរាជរដ្ឋាភិបាលកម្ពុជា និង រដ្ឋមន្ត្រីក្រសួងការងារ និងបណ្តុះបណ្តាលវិជ្ជាជីវៈ

Nonetheless, we cannot Gemini all documents (e.g., private, classified, confidential documents).

Yet, *Translate Khmer text into English: កុំដំឡូងមួលពេក។*

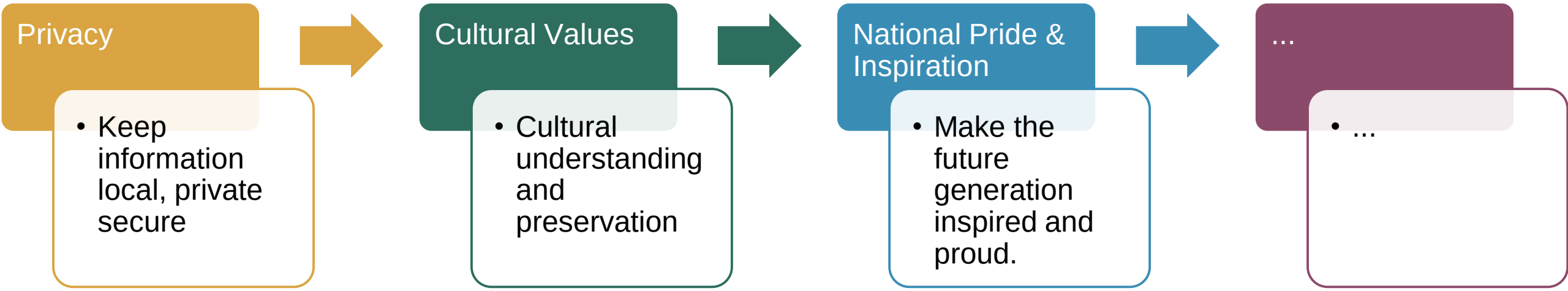
Don't make it too round. - Gemini

"Don't put too many potatoes." (literally "don't [add] round potatoes too much.") – Claude Sonnet 5.0

“Don’t make the potatoes too round.” - ChatGPT

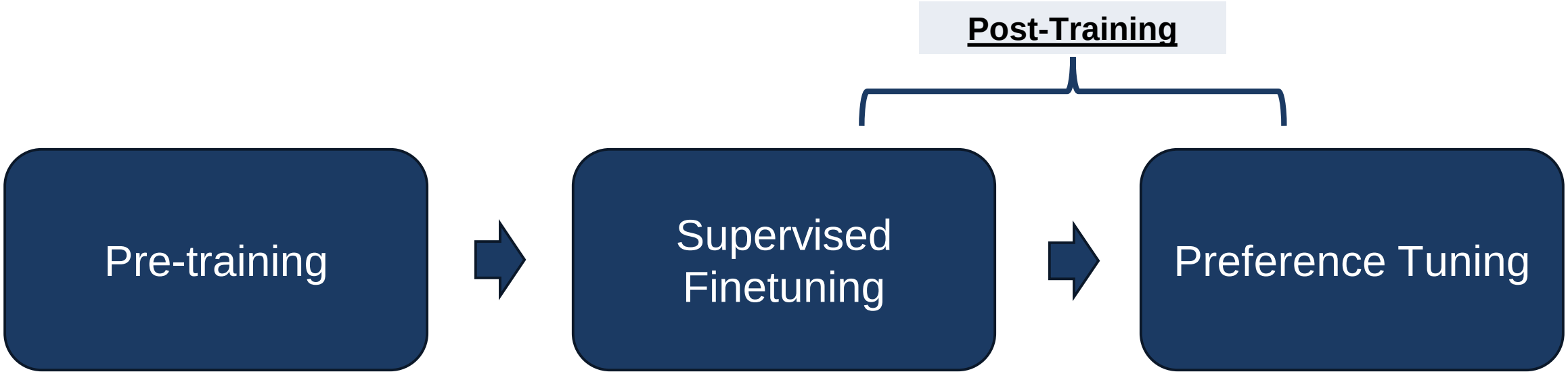
All these frontier models fail on this surprisingly simple translation.

Thus, A Khmer-Native AI Is Needed Because:



Thus, developing a Khmer-native AI model is not a question of 'if' but a question of 'when' and 'how'.

But, Training a Large Model Is Not Easy

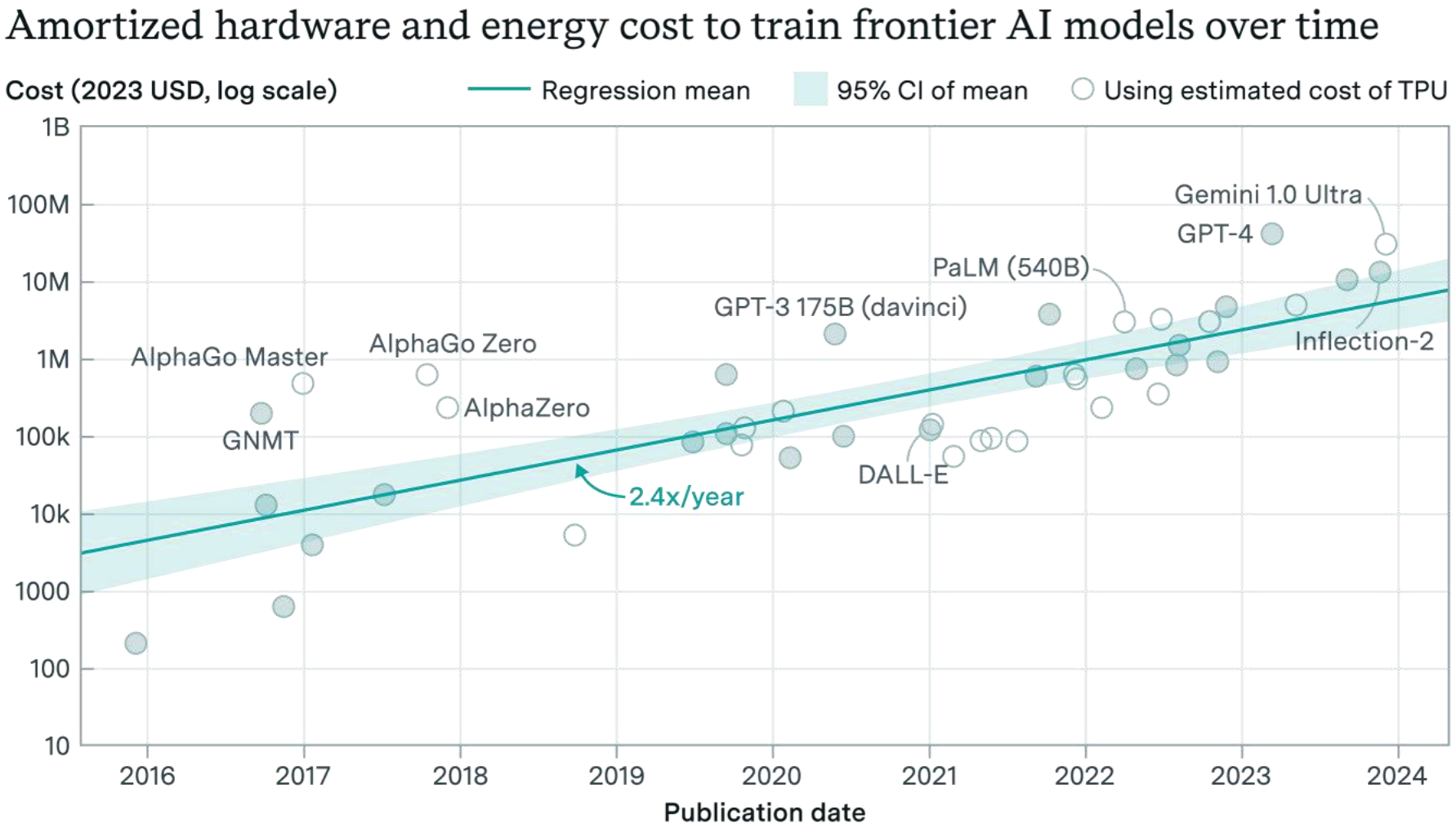


- Large-scale data (~15T tokens)
- Large-scale compute
- Cost-intensive

- Less data
- Less compute
- Low cost

While compute and talents can be developed, Khmer remains a low-resource language.

Cost of Training a Frontier Model



The rising costs of training frontier AI models

Not many labs can afford this.

Then, We Can Machine Translate

AIF Insights No. 14 (2025) | Machine Translation: Bridging Khmer Language and Large Language Models (LLMs)

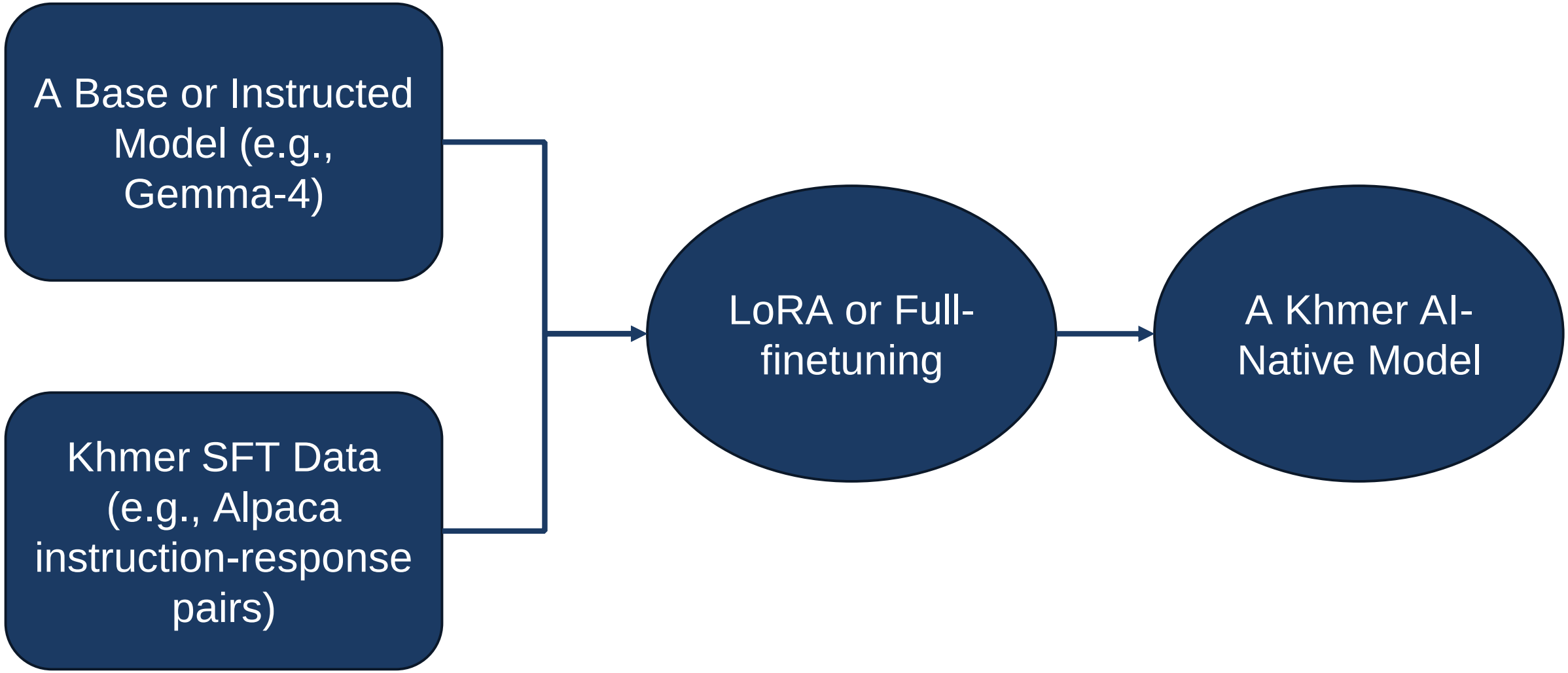
Release: 2025-05-09

Author: BUOY Rina, PhD (Informatics) | CHENDA Sovisal, MA (Advanced Computing Systems) | TAING Nguonly, PhD (Computer Science) | KONG Marry, PhD (Telecommunications Technology and Information Technology)

While the development of large language models (LLMs), specifically for the Khmer language, is constrained by both computational resources and the availability of quality training data, this paper presents a strategic direction highly relevant to artificial intelligence and machine learning engineers and developers working with the Khmer language. This paper highlights the significant role of high-quality machine translation in advancing Khmer natural language processing (KNLP). It argues that improving machine translation between Khmer and high- resource languages is currently the most viable approach to fully leverage LLMs and other advanced models for the Khmer language, at least in the short term.

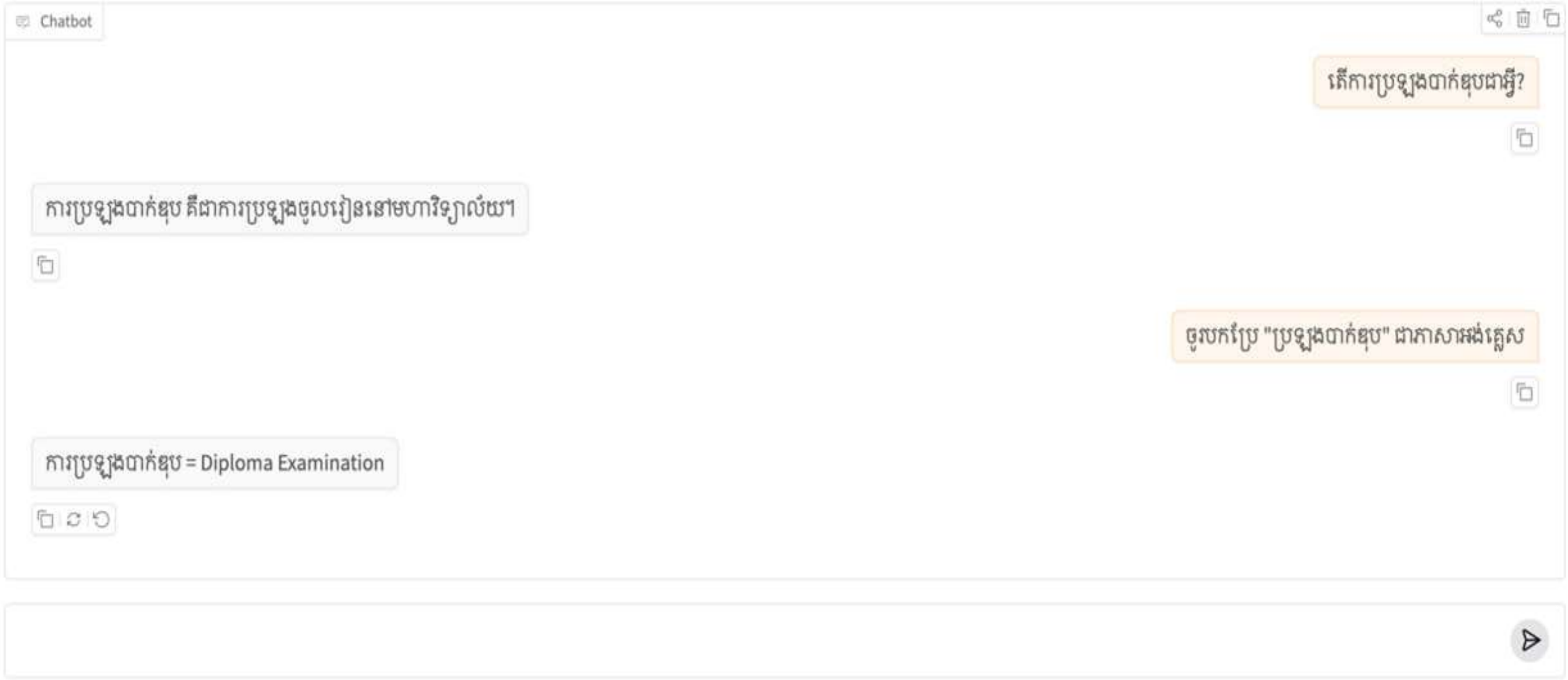
Working robustly in some cases, yet not a Khmer-native approach.

Then, Finetuning is the Solution



Preliminary encouraging results.

Preliminary Experimental Results



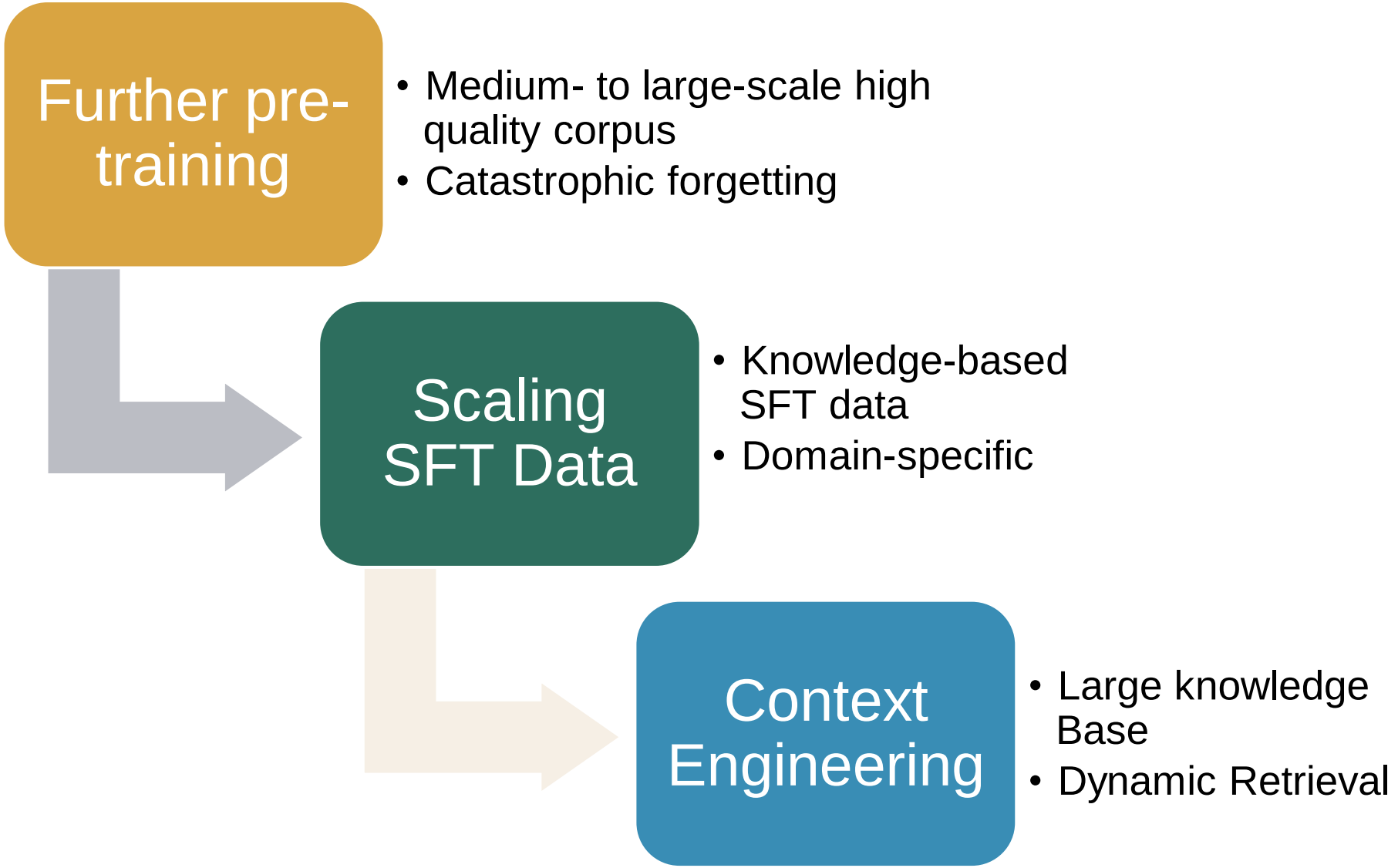
Natively understanding Khmer instructions and respond in Khmer fluently.

Yet, Still Hallucinate …



The model may understand instructions but lacks the necessary Khmer knowledge to answer.

The Fixes & Way Forward



Thank you

Any comments/questions, or interest in our personalized AI courses and mastery program,

email : rinabuoy@gmail.com