



Performance Evaluation and Comparative Analysis of Lao-Chinese Machine Translation using Few-shot Prompting between SEA-LION and Qwen Models in Tourism-Railway Domains

Vimontha Khievongphachanh^{*1}, Ge Lina², Quan Guo², Khamhoung Chanthavong³

¹Department of Computer Engineering and Information Technology, Faculty of Engineering, National University of Lao, ²Guangxi Minxu University, China, ³Confucius Institute, National University of Lao

***Correspondence:** Vimontha

Khievongphachanh ,
Department of Computer
Engineering and Information
Technology, Faculty of
Engineering, National University
of Lao. Tel: +856 20 55592414,
E-mail: vimontha@fe-nuol.edu.la

Article Info:

Submitted: June 2, 2026

Revised: June 18, 2026

Accepted: June 24, 2026

Abstract

The growth of the Laos-China Railway project has rapidly boosted both the economy and tourism, making cross-lingual communication between Laos and China increasingly important. However, Lao is still considered a low-resource language, which limits the performance of machine translation tools. This study aims to evaluate and compare the translation efficiency between two open-weight Large Language Models (LLMs): SEA-LION-v3-8B and Qwen2.5-7B. The test used a 3-shot prompting technique without any fine-tuning. The experiments were conducted on Google Colab using 4-bit quantization. The test dataset included 720 sentences, divided into two domains: General and Travel-Railway, with sample sizes of 10, 20, 50, and 100 sentences. The translation quality was measured using the BLEU score. The results show that no single model performs best in every scenario. In the General domain for Lao-to-Chinese (Lao→Zh) translation, Qwen outperformed SEA-LION across all data sizes, reaching a top score of 21.62. On the other hand, in the Travel-Railway domain (Lao→Zh), SEA-LION did significantly better than Qwen, with a maximum score of 10.99. This outcome aligns with the Domain Adaptation theory. However, both models faced an imbalance in translation directions. The Chinese-to-Lao (Zh→Lao) translation scores were much lower, mainly due to tokenizer limitations and the lack of a Lao corpus. In conclusion, this research proves that using 7-8B parameter LLMs with 4-bit quantization is highly possible in resource-limited environments. It also provides a useful guideline for developing better tokenizers or fine-tuning models with specific domain data in the future.

Keywords: Machine Translation; Large Language Models; Low-Resource Language; Few-shot Prompting; Tourism and Railway Domain

1. ພາກສະເໜີ

ການຂະຫຍາຍຕົວຂອງໂຄງການລົດໄຟລາວ-ຈີນ ໄດ້ເພີ່ມຈຳນວນນັກທ່ອງທ່ຽວ, ນັກລົງທຶນ ແລະ ການແລກປ່ຽນວັດທະນະທຳຢ່າງກ້າວກະໂດດ, ເຮັດໃຫ້ຄວາມຕ້ອງການການສື່ສານຂ້າມພາສາລາວ-ຈີນ ມີຄວາມສຳຄັນເປັນພິເສດ. ເຖິງຢ່າງໃດກໍຕາມ, ພາສາລາວຖືກຈັດຢູ່ໃນກຸ່ມພາສາທີ່ມີຊັບພະຍາກອນຂໍ້ມູນຕໍ່າ (Low-resource

Language) ເຊິ່ງຍັງຂາດຂໍ້ມູນຄູ່ຂະໜານຂະໜາດໃຫຍ່ ແລະ ເຄື່ອງມືປະມວນຜົນພາສາທຳມະຊາດແບບດິຈິຕອນທີ່ຄົບຖ້ວນ, ສິ່ງຜິດໃຫ້ການພັດທະນາລະບົບແປພາສາອັດຕະໂນມັດປະຊຸມກັບຄວາມທ້າທາຍດ້ານຄວາມຖືກຕ້ອງ ແລະ ຄວາມສອດຄ່ອງກັບບໍລິບົດຂອງເນື້ອຫາ (Joshi et al., 2020; Ranathunga et al., 2023). ໂມດວພາສາຂະໜາດໃຫຍ່ (Large Language Models)

ແບບຫຼາຍພາສາ ໄດ້ເປີດໂອກາດໃໝ່ໃນການແກ້ໄຂບັນຫາການແປພາສາໂດຍບໍ່ຕ້ອງໃຊ້ຂໍ້ມູນຝຶກຝົນຈຳນວນຫຼາຍ (Costa-jussà et al., 2022; Winata et al., 2022).

ໃນຂົງເຂດອາຊີຕາເວັນອອກສຽງໃຕ້, ມີຫຼາຍບົດຄົ້ນຄວ້າທີ່ໄດ້ສຶກສາການພັດທະນາເຕັກໂນໂລຊີພາສາສໍາລັບພາສາທີ່ມີຊັບພະຍາກອນຕໍ່າ. ຕົວຢ່າງເຊັ່ນ, ການສຶກສາຂອງ Keohavong et al. (2023) ໄດ້ວິເຄາະຄວາມທ້າທາຍຂອງການແປພາສາໃນພາສາລາວ ໂດຍຊີ້ໃຫ້ເຫັນວ່າການຂາດຂໍ້ມູນຄູ່ຂະໜານທີ່ມີຄຸນນະພາບ ແລະ ຄວາມຊັບຊ້ອນຂອງໂຄງສ້າງໄວຍາກອນລາວ ເປັນປັດໄຈຫຼັກທີ່ຈຳກັດປະສິດທິພາບຂອງລະບົບ NMT. Sengxay et al. (2022) ໄດ້ທົດລອງພັດທະນາລະບົບແປພາສາລາວ-ຈີນໂດຍໃຊ້ເຕັກນິກ Transfer Learning ຈາກພາສາໄທ ແລະ ຈີນ, ເຊິ່ງສາມາດປັບປຸງຄະແນນ BLEU ໄດ້ປະມານ 15–20%. Winata et al. (2022) ໄດ້ສະເໜີແນວທາງການຖ່າຍໂອນຄວາມຮູ້ຂ້າມພາສາ (Cross-lingual Transfer) ສໍາລັບພາສາໃນອາຊີຕາເວັນອອກສຽງໃຕ້, ໂດຍຢືນຢັນວ່າ ໂມເດວທີ່ຖືກຝຶກຝົນດ້ວຍຂໍ້ມູນຫຼາຍພາສາສາມາດປັບຕົວໄດ້ດີກັບພາສາທີ່ມີຂໍ້ມູນໜ້ອຍ ໂດຍບໍ່ຕ້ອງໃຊ້ຂໍ້ມູນຝຶກຝົນໃໝ່ທັງໝົດ.

Chu and Wang (2022) ມີການປັບໂມເດວໃຫ້ເໝາະກັບໂດເມນສະເພາະ (Domain Specific) ຕ້ອງອີງໃສ່ການກັ່ນຕອງຂໍ້ມູນ ແລະ ການອອກແບບ Prompt ທີ່ຊັດເຈນ, ໂດຍສະເພາະໃນຂະແໜງການທ່ອງທ່ຽວ ແລະ ການຂົນສົ່ງທີ່ມີຄໍາສັບວິຊາການສະເພາະ. Zhang et al. (2024) ໄດ້ສຶກສາການ Fine-tune ໂມເດວຫຼາຍພາສາສໍາລັບຂໍ້ມູນການຂົນສົ່ງ ແລະ ການທ່ອງທ່ຽວໃນພາກພື້ນອາຊີ, ໂດຍພົບວ່າ ໂມເດວທີ່ຖືກປັບດ້ວຍຂໍ້ມູນໂດເມນສະເພາະສາມາດຫຼຸດອັດຕາຄວາມຜິດພາດໃນຂໍ້ມູນຕົວເລກ ແລະ ຊື່ສະຖານີໄດ້ຢ່າງຊັດເຈນ. ເຖິງແມ່ນວ່າຈະມີການສຶກສາກ່ຽວກັບການແປພາສາໃນອາຊີຕາເວັນອອກສຽງໃຕ້ ແຕ່ຍັງຂາດການປຽບທຽບແບບປະລິມານທີ່ເນັ້ນໃສ່ໂດເມນການທ່ອງທ່ຽວ ແລະ ລົດໄຟໂດຍກົງ. ດັ່ງນັ້ນ, ການຄົ້ນຄວ້າຄັ້ງນີ້ຈຶ່ງມີຈຸດປະສົງເພື່ອ: (1) ປຽບທຽບປະສິດທິພາບການແປພາສາລາວ-ຈີນ ແບບ Few-shot ລະຫວ່າງໂມເດວ SEA-LION ແລະ Qwen, (2) ວິເຄາະຜົນການແປ

ຂໍ້ຄວາມ ທົ່ວໄປ ແລະ ຂໍ້ຄວາມການທ່ອງທ່ຽວ-ລົດໄຟ, ແລະ (3) ປະເມີນຄວາມເປັນໄປໄດ້ໃນການດໍາເນີນການທົດລອງໃນສະພາບແວດລ້ອມຊັບພະຍາກອນຕໍ່າ.

2. ອຸປະກອນ ແລະ ວິທີການ

2.1 ວິທີການທົດລອງ

ການຄົ້ນຄວ້າໃຊ້ໂມເດວພາສາຂະໜາດໃຫຍ່ (LLMs) ແບບ Open-weight ຂະໜາດປະມານ 7-8 ພັນລ້ານພາຣາມິເຕີ ໂມເດວທີ່ໃຊ້ແມ່ນໂມເດວ SEA-LION-v3-8B (Base) ແລະ Qwen2.5-7B (Base) ໂດຍມີເງື່ອນໄຂການເຂົ້າເຖິງທີ່ແຕກຕ່າງກັນ:

-ໂມເດວ SEA-LION-v3-8B ຈະຕ້ອງສ້າງ Access Token ໃນ Hugging Face ແລະ ກົດຂໍສິດເຂົ້າເຖິງໜ້າໂມເດວກ່ອນຈຶ່ງຈະສາມາດໂຫຼດມາໃຊ້ໄດ້ AI Singapore (2025).

-ໂມເດວ Qwen2.5-7B ເປັນແບບ Public ທີ່ສາມາດໂຫຼດ ແລະ ນຳໃຊ້ໄດ້ທັນທີໂດຍບໍ່ຕ້ອງຂໍສິດ (Bai, 2024).

ການກຳນົດສະພາບແວດລ້ອມດ້ານຮາດແວ: ເພື່ອປະຢັດໜ່ວຍຄວາມຈຳກຣາຟິກ (VRAM) ໃນການປະມວນຜົນ, ວິທີການທົດລອງໄດ້ນຳໃຊ້ເຕັກນິກການຫຼຸດຈຳນວນບິດແບບ 4-bit Quantization ໃນການໂຫຼດໂມເດວ.

ການອອກແບບ Prompt : ການທົດສອບນີ້ໄດ້ນຳໃຊ້ເຕັກນິກ Few-shot Prompting ໂດຍບໍ່ຜ່ານຂັ້ນຕອນການປັບແຕ່ງໂມເດວ (Fine-tuning) ເຊິ່ງເປັນແນວທາງການຮຽນຮູ້ໃນບໍລິບົດ (In-context Learning) ຕາມແນວຄິດຂອງ Brown et al. (2020). ໂຄງສ້າງຂອງ Prompt ຖືກອອກແບບໂດຍການໃສ່ຕົວຢ່າງຄູ່ປະໂຫຍກການແປ (ລາວ-ຈີນ ແລະ ຈີນ-ລາວ) ຈຳນວນ 3 ຄູ່ ວາງໄວ້ກ່ອນຂໍ້ຄວາມທີ່ຕ້ອງການທົດສອບ ເພື່ອຊີ້ນຳໃຫ້ໂມເດວເຂົ້າໃຈຮູບແບບການແປ (ດັ່ງສະແດງໃນຕາຕະລາງ 1 ແລະ ຕາຕະລາງ 2).

ການຄວບຄຸມການສ້າງຂໍ້ຄວາມ: ວິທີການທົດລອງໄດ້ກຳນົດພະລາມິເຕີ stop_word ເປັນ "Chinese:" ຫຼື "Lao:" ຕາມທິດທາງການແປ ເພື່ອຄວບຄຸມບໍ່ໃຫ້ໂມເດວສ້າງຂໍ້ຄວາມເກີນຄວາມຈຳເປັນ, ຫຼຸດການເວົ້າຊ້ຳ, ແລະ

ເນັ້ນໃຫ້ໂມເດວແປ ຫຼື ສ້າງຂໍ້ຄວາມສັ້ນໆ ຢ່າງແມ້ນຍຳ ແລະ ວ່ອງໄວ.

2.2 ວິທີການວິເຄາະຂໍ້ມູນ

ຂໍ້ມູນທີ່ໄດ້ຮັບຈາກການທົດລອງຈະຖືກນຳມາ ວິເຄາະດ້ວຍວິທີການແບບປະລິມານ ແລະ ການວິເຄາະແບບ ປຽບທຽບປະສິດທິພາບ (Quantitative and Comparative Analysis) ເຊິ່ງໃນວຽກງານການປະມວນ ຜົນພາສາທຳມະຊາດ (NLP) ແລະ ການປະເມີນຜົນການ ແປພາສາໂດຍທົ່ວໄປສາມາດແບ່ງອອກເປັນ 2 ແນວທາງ ຫຼັກ ຄື: ການປະເມີນຜົນອັດຕະໂນມັດ (Automatic Evaluation) ເຊັ່ນ ມາດຕະຖານ BLEU ແລະ ChrF, ແລະ ການປະເມີນຜົນໂດຍມະນຸດ (Human Evaluation) (Papineni et al., 2002; Popović, 2015). ບົດຄົ້ນຄວ້າ ນີ້ໄດ້ເລືອກໃຊ້ການປະເມີນຜົນອັດຕະໂນມັດດ້ວຍ ມາດຕະຖານ BLEU (Bilingual Evaluation Understudy) (Post, 2018). ເປັນຕົວຊີ້ວັດຫຼັກ ໂດຍບໍ່ ໄດ້ນຳໃຊ້ການປະເມີນຜົນຈາກມະນຸດ ເນື່ອງຈາກການວິໄຈ ນີ້ເນັ້ນໃສ່ການວັດແທກຄວາມຖືກຕ້ອງທາງດ້ານຄຳສັບ ແລະ ຄວາມທ່ຽງຕົງຂອງຜົນການແປເປັນຫຼັກ ໂດຍ ມາດຕະຖານ BLEU ມີຂໍ້ດີດ້ານຄວາມເປັນກາງທາງສະຖິຕິ ທີ່ໃຊ້ເກນການຄິດໄລ່ທາງຄະນິດສາດທີ່ຄົງທີ່ ແລະ ເນື່ອງ ຈາກຊຸດຂໍ້ມູນທົດສອບມີຈຳນວນຫຼາຍເຖິງ 720 ປະໂຫຍກ ແລະ ມີຫຼາຍເງື່ອນໄຂການທົດລອງ, ການນຳໃຊ້ BLEU ຈຶ່ງຊ່ວຍໃຫ້ການປະມວນຜົນມີຄວາມວ່ອງໄວ, ມີ ປະສິດທິພາບ ເຮັດໃຫ້ສາມາດປຽບທຽບລະຫວ່າງໂມເດວ SEA-LION ແລະ Qwen ພາຍໃຕ້ສະພາບແວດລ້ອມ ແລະ ເກນການວັດແທກດຽວກັນ.

3. ຜົນໄດ້ຮັບ

ຈາກການທົດສອບປະສິດທິພາບການແປພາສາແບບ Few-shot ລະຫວ່າງໂມເດວ SEA-LION ແລະ Qwen ໂດຍອີງໃສ່ຕົວຊີ້ວັດຄະແນນ (Score) ໃນທິດທາງການແປ ລາວ-ຈີນ (Lao→Zh) ແລະ ຈີນ-ລາວ (Zh→Lao) ພາຍ ໃຕ້ການກຳນົດຂະໜາດຊຸດຂໍ້ມູນທີ່ແຕກຕ່າງກັນ (10, 20, 50, ແລະ 100 ປະໂຫຍກ), ຜົນການວິໄຈສາມາດສະຫຼຸບ ແລະ ວິເຄາະໃຫ້ເຫັນປະສິດທິພາບໃນໂດເມນທົ່ວໄປ (General Domain) ວ່າ ໃນທິດທາງການແປ ລາວ ຫາ

ຈີນ (Lao→Zh) ໂມເດວ Qwen ສະແດງປະສິດທິພາບ ແລະ ຄວາມແມ້ນຍຳສູງກວ່າ SEA-LION ຢ່າງເຫັນໄດ້ ຊັດເຈນໃນທຸກໆຂະໜາດຂອງຊຸດຂໍ້ມູນ ໂດຍສາມາດເຮັດ ຄະແນນໄດ້ສູງສຸດທີ່ 21.62 ໃນຊຸດຂໍ້ມູນຂະໜາດ 50 ປະໂຫຍກ, ແຕ່ໃນທາງກົງກັນຂ້າມ ສາມາດສັງເກດເຫັນວ່າ ໃນທິດທາງການແປ ຈີນ ຫາ ລາວ (Zh→Lao) ໂມເດວ SEA-LION ກັບສາມາດເຮັດຄະແນນໄດ້ດີກວ່າ Qwen ເມື່ອຈຳນວນຊຸດຂໍ້ມູນຕົວຢ່າງເພີ່ມຂຶ້ນຕັ້ງແຕ່ 20 ຫາ 100 ປະໂຫຍກ.

ສຳລັບການປຽບທຽບປະສິດທິພາບ ແລະ ວິເຄາະຜົນ ການແປໃນໂດເມນທ່ອງທ່ຽວ-ລົດໄຟ (Travel-Railway Domain) ພົບວ່າ ໃນທິດທາງການແປ ລາວ ຫາ ຈີນ (Lao→Zh) ໂມເດວ SEA-LION ມີປະສິດທິພາບເໝືອນ ກວ່າ Qwen ໂດຍມີຄະແນນສູງສຸດຢູ່ທີ່ 10.99 ໃນຊຸດຂໍ້ ມູນຂະໜາດ 100 ປະໂຫຍກ, ຢ່າງໃດກໍຕາມ ໃນທິດທາງ ການແປ ຈີນ ຫາ ລາວ (Zh→Lao) ພົບວ່າທັງສອງໂມເດວ ບໍ່ສາມາດແປພາສາໄດ້ເລີຍໃນໂດເມນນີ້ ເຊິ່ງສະທ້ອນໃຫ້ ເຫັນເຖິງຂໍ້ຈຳກັດຢ່າງຮຸນແຮງໃນການປະມວນຜົນຄຳສັບ ສະເພາະທາງ.

ເມື່ອປະເມີນຄວາມເປັນໄປໄດ້ໃນສະພາບແວດລ້ອມ ຊັບພະຍາກອນຕ່ຳ (Low-resource Environment) ສາມາດສະຫຼຸບໄດ້ວ່າ ການຮຽນຮູ້ແບບ Few-shot ທີ່ໃຊ້ຂໍ້ ມູນຕົວຢ່າງໜ້ອຍພຽງ 10 ຫາ 100 ປະໂຫຍກ ແມ່ນມີ ຄວາມເປັນໄປໄດ້ ແລະ ສາມາດນຳມາປະຍຸກໃຊ້ໄດ້ດີໃນໂດ ເມນທົ່ວໄປ ໂດຍສະເພາະໂມເດວ Qwen ໃນທິດທາງ Lao→Zh ແລະ SEA-LION ໃນທິດທາງ Zh→Lao, ແຕ່ສຳລັບໂດເມນສະເພາະທາງຢ່າງທ່ອງທ່ຽວ-ລົດໄຟ ຄະແນນຂອງທັງສອງໂມເດວແມ່ນຍັງຢູ່ໃນເກນທີ່ຕ່ຳຫຼາຍ ໃນການແປ ລາວ→Zh ແລະ ລົ້ມເຫຼວຈົນບໍ່ສາມາດແປໄດ້ ໃນການແປ ຈີນ→Lao ເຊິ່ງສິ່ງນີ້ຊີ້ໃຫ້ເຫັນວ່າ ໃນສະພາບ ແວດລ້ອມທີ່ຊັບພະຍາກອນພາສາລາວມີຈຳກັດ ການໃຊ້ ພຽງເຕັກນິກ Few-shot ຍັງບໍ່ພຽງພໍທີ່ຈະເຮັດໃຫ້ AI ເຂົ້າໃຈຄຳສັບສະເພາະທາງໄດ້ ແລະ ຈຳເປັນຕ້ອງໄດ້ມີການ ເພີ່ມຄັງຂໍ້ມູນຄູ່ຂະໜານ ຫຼື ການເຮັດ Fine-tuning ເພີ່ມ ຕື່ມໃນອະນາຄົດ..

4. ວິພາກຜົນ

ຈາກຜົນການທົດລອງປຽບທຽບປະສິດທິພາບການແປພາສາລາວ-ຈີນ ແບບ Few-shot ລະຫວ່າງໂມເດວ SEA-LION ແລະ Qwen ຊຶ່ງໃຫ້ເຫັນວ່າ ບໍ່ມີໂມເດວໃດທີ່ສາມາດສະແດງສັກກະຍະພາບໄດ້ເໝືອກວ່າໃນທຸກສະຖານະການ. ເມື່ອວິເຄາະແຍກຕາມໂດເມນຂໍ້ຄວາມ ພົບວ່າ ໂມເດວ Qwen ເຮັດໄດ້ດີໃນໂດເມນທົ່ວໄປ (General Domain) ເນື່ອງຈາກໄດ້ຮັບການຝຶກຝົນດ້ວຍຄຳຂໍ້ມູນທົ່ວໄປຂະໜາດໃຫຍ່, ແຕ່ເມື່ອປ່ຽນມາເປັນໂດເມນທ່ອງທ່ຽວ-ລົດໄຟ (Travel-Railway Domain) ໃນທິດທາງການແປລາວ ຫາ ຈີນ (Lao→Zh), ໂມເດວ SEA-LION ກັບສາມາດສະແດງປະສິດທິພາບໄດ້ດີກວ່າ. ປາກົດການນີ້ແມ່ນສອດຄ່ອງກັບທິດສະດີການປັບຕົວເຂົ້າກັບໂດເມນ (Domain Adaptation) ທີ່ລະບຸວ່າ ໂມເດວພາສາຂະໜາດໃຫຍ່ (LLMs) ເຖິງວ່າຈະມີຄວາມຮູ້ຮອບດ້ານ ແຕ່ມັກຈະມີປະສິດທິພາບຫຼຸດລົງເມື່ອຕ້ອງປະມວນຜົນຄຳສັບສະເພາະດ້ານ ຖ້າບໍ່ໄດ້ຮັບການປັບແຕ່ງ (Fine-tuning) ໃຫ້ເຂົ້າກັບບໍລິບົດສະເພາະຂອງພາກພື້ນ (Sennrich et al., 2016; Chu & Wang, 2018).

ນອກຈາກນັ້ນ, ໃນການປະມົນຄວາມເປັນໄປໄດ້ໃນສະພາບແວດລ້ອມຊັບພະຍາກອນຕ່ຳ, ຜົນການທົດລອງທິດທາງການແປຈາກ ລາວ ຫາ ຈີນ (Lao→Zh) ມີຄະແນນສູງກວ່າການແປຈາກ ຈີນ ຫາ ລາວ (Zh→Lao) ຢ່າງເຫັນໄດ້ຊັດເຈນນັ້ນ ແມ່ນສອດຄ່ອງກັບແນວຄິດການພັດທະນາລະບົບພາສາສຳລັບກຸ່ມພາສາທີ່ມີຊັບພະຍາກອນຕ່ຳ (Low-resource languages); ເຊິ່ງມີການຊີ້ໃຫ້ເຫັນວ່າ ໂຄງສ້າງຂອງ Tokenizer ແລະ ການຂາດແຄນຄັງຂໍ້ມູນພາສາປາຍທາງ ຈະສົ່ງຜົນເຮັດໃຫ້ຂັ້ນຕອນການສ້າງປະໂຫຍກ ເປັນພາສາທີ່ມີຊັບພະຍາກອນຕ່ຳມີຄວາມຖືກຕ້ອງຫຼຸດລົງ ເມື່ອທຽບໃສ່ການແປໄປຫາພາສາທີ່ມີຊັບພະຍາກອນສູງ (Yong et al., 2023; Zhao et al., 2025). ຂໍ້ຈຳກັດຂອງການປະມວນຜົນພາສາລາວໃນສະພາບແວດລ້ອມດັ່ງກ່າວ ຍັງສອດຄ່ອງກັບການສຶກສາຂອງ Khievongphachanh et al. (2023) ທີ່ກ່າວວ່າລະບົບປັນຍາປະດິດ ຫຼື ເຄື່ອງມືມາດຕະຖານລະດັບສາກົນມັກຈະມີປະສິດທິພາບຫຼຸດລົງເມື່ອຕ້ອງປະມວນຜົນກັບພາສາລາວ ຫາກຂາດການຝຶກຝົນ (Train) ຫຼື ການປັບແຕ່ງ

ຄັງຂໍ້ມູນ ໃຫ້ເຂົ້າກັບລັກສະນະສະເພາະຂອງຕົວອັກສອນ, ການຕັດຄຳ, ແລະ ບໍລິບົດທາງພາສາລາວ.

5. ສະຫຼຸບ

ຜົນການທົດລອງແປພາສາແບບ Few-shot ພົບວ່າ ບໍ່ມີໂມເດວໃດທີ່ມີປະສິດທິພາບເໝືອກວ່າໃນທຸກສະຖານະການ ໂດຍໂມເດວ Qwen2.5-7B ສາມາດເຮັດໄດ້ດີໃນໂດເມນທົ່ວໄປ (General Domain) ທິດທາງລາວ→ຈີນ ຍ້ອນໄດ້ຮັບການຝຶກຝົນດ້ວຍຄຳຂໍ້ມູນຂະໜາດໃຫຍ່, ໃນຂະນະທີ່ SEA-LION-v3-8B ເຮັດໄດ້ດີກວ່າໃນໂດເມນທ່ອງທ່ຽວ-ລົດໄຟ (Travel-Railway Domain) ທິດທາງ ລາວ→ຈີນ ເຊິ່ງສະທ້ອນເຖິງຄວາມໄດ້ປຽບຂອງໂມເດວທີ່ຖືກອອກແບບມາສຳລັບກຸ່ມພາສາໃນພາກພື້ນອາຊີຕາເວັນອອກສຽງໃຕ້ໂດຍສະເພາະຢ່າງໃດກໍຕາມ, ທັງສອງໂມເດວຍັງຄົງປະສົບບັນຫາຄວາມບໍ່ສົມດຸນໃນທິດທາງການແປ ໂດຍທິດທາງ ຈີນ→ລາວ ມີຄະແນນ BLEU ຕ່ຳກວ່າ ລາວ→ຈີນ ຍ້ອນຂໍ້ຈຳກັດຂອງ Tokenizer ແລະ ການຂາດແຄນຄັງຂໍ້ມູນພາສາລາວ; ນອກຈາກນັ້ນ, ການທົດລອງພາຍໃຕ້ສະພາບແວດລ້ອມທີ່ມີ VRAM ຈຳກັດ (Google Colab) ບົບໃຫ້ຕ້ອງໃຊ້ເຕັກນິກ 4-bit Quantization ເຊິ່ງອາດຫຼຸດຄວາມລະອຽດໃນການແປລົງເລັກນ້ອຍ, ປະກອບກັບຊຸດຂໍ້ມູນທົດສອບໃນປັດຈຸບັນຍັງບໍ່ທັນກວມເອົາສຳນວນທ້ອງຖິ່ນ ແລະ ຄຳສັບເຕັກນິກທັງໝົດ. ເນື່ອງຈາກການສຶກສານີ້ເນັ້ນໃສ່ພຽງແຕ່ເຕັກນິກ Few-shot Prompting ໂດຍບໍ່ໄດ້ມີການປັບແຕ່ງນ້ຳໜັກ, ການຄົ້ນຄວ້າໃນອະນາຄົດຈຶ່ງຄວນມຸ່ງເນັ້ນໄປທີ່ການເຮັດ Fine-tuning ດ້ວຍຂໍ້ມູນຄູ່ຂະໜານສະເພາະດ້ານທີ່ມີຄຸນນະພາບສູງ ຄວບຄູ່ໄປກັບການພັດທະນາ Tokenizer ທີ່ຮອງຮັບພາສາລາວໂດຍກົງເພື່ອຍົກລະດັບຄວາມແມ່ນຢຳໃຫ້ສູງຂຶ້ນ. ການຄົ້ນຄວ້ານີ້ໄດ້ຝຸດໃຫ້ເຫັນເຖິງຄວາມເປັນໄປໄດ້ໃນການພັດທະນາ AI ຂັ້ນສູງດ້ວຍຕົ້ນທຶນຕ່ຳ ເຊິ່ງຜົນການຄົ້ນຄວ້າທີ່ໄດ້ ສາມາດນຳໄປຕໍ່ຍອດເປັນເຄື່ອງມືແປພາສາອັດຕະໂນມັດ ເພື່ອອຳນວຍຄວາມສະດວກ ແລະ ຫຼຸດຜ່ອນຂໍ້ຜິດພາດໃນການສື່ສານໃຫ້ແກ່ນັກທ່ອງທ່ຽວ, ພະນັກງານ, ແລະ ຜູ້ປະກອບໄດ້ຢ່າງມີປະສິດທິພາບ.

6. ຂໍ້ຂັດແຍ່ງ

ຂ້າພະເຈົ້າໃນນາມຜູ້ຄົນຄວາມວິທະຍາສາດ ຂໍ
ປະຕິຍານຕົນວ່າ ຂໍ້ມູນທັງໝົດທີ່ມີໃນບົດຄວາມວິຊາການ
ດັ່ງກ່າວນີ້ ແມ່ນບໍ່ມີຂໍ້ຂັດແຍ່ງທາງຜົນປະໂຫຍດກັບ
ພາກສ່ວນໃດ ແລະ ບໍ່ໄດ້ເອື້ອປະໂຫຍດໃຫ້ກັບພາກສ່ວນ
ໃດພາກສ່ວນໜຶ່ງ, ກໍລະນີມີການລະເມີດ ໃນຮູບການໃດ
ໜຶ່ງ ຂ້າພະເຈົ້າມີຄວາມຍິນດີ ທີ່ຈະຮັບຜິດຊອບແຕ່ພຽງຜູ້
ດຽວ.

7. ເອກະສານອ້າງອີງ

- AI Singapore. (2025). *Llama-SEA-LION-v3-8B: Southeast Asian languages in one network* [Model card]. Hugging Face.
<https://huggingface.co/aisingapore/Llama-SEA-LION-v3-8B>
- Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., & Yang, A. (2024). *Qwen2.5 technical report* (arXiv:2412.15115). arXiv.
<https://doi.org/10.48550/arXiv.2412.15115>
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.
- Chu, C., & Wang, R. (2022). Domain adaptation for neural machine translation: A survey. *Computational Linguistics*, 48(3), 651-692. https://doi.org/10.1162/coli_a_00429
- Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., ... & Wang, G. (2022). *No language left behind: Scaling human-centered machine translation* (arXiv:2207.04672). arXiv.
<https://doi.org/10.48550/arXiv.2207.04672>
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282-6293.
<https://doi.org/10.18653/v1/2020.acl-main.560>
- Keohavong, S., Vongphachanh, K., & Phetsarath, S. (2023). Machine translation challenges in low-resource Southeast Asian languages: A Lao perspective. *Souphanouvong University Journal of Multidisciplinary Research and Development*, 9(2), 112-125.
- Sengxay, S., Yu, Z., & Liu, C. H. (2022). Enhancing Lao-Chinese machine translation via transfer learning with Thai-Lao linguistic similarity. In *Proceedings of the International Conference on Asian Language Processing (IALP 2022)* (pp. 45–52). IEEE
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311-318.
<https://doi.org/10.3115/1073083.1073135>
- Post, M. (2018). A call for clarity in reporting BLEU scores. *Proceedings of the 3rd Conference on Machine Translation: Research Papers*, 186-191.
<https://doi.org/10.18653/v1/W18-6319>
- Ranathunga, S., Lee, E. A., Skute, M. P., & Dias, G. (2023). Low-resource machine translation: A survey. *Computational Linguistics*, 49(4), 1045-1098.
https://doi.org/10.1162/coli_a_00492
- Sennrich, R., Haddow, B., & Birch, A. (2016). Improving neural machine translation models with monolingual data. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 86-96.
<https://doi.org/10.18653/v1/P16-1009>
- Winata, G. I., Lin, Z., & Fung, P. (2022). Cross-lingual transfer for low-resource Southeast Asian languages. *Proceedings of the Workshop on Language Technology for Equal Access*, 45-52.
- Yong, X., et al. (2023). Overcoming tokenization bias for low-resource

languages in large language models.
Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), 4512-4526.

Association for Computational Linguistics.

Zhang, Y., Chen, L., Wang, H., & Li, M.
 (2024). *Fine-tuning multilingual LLMs for domain-specific translation in transportation and tourism*
 (arXiv:2403.11287). arXiv.
<https://doi.org/10.48550/arXiv.2403.11287>

Khieovongphachanh, V., Sounakhen, K.,
 Saypadith, S., Mekthanavanh, V.,
 Luangxaysana, K., & Chanthavong, X.
 (2023). Enhancing Lao vehicle license
 plate recognition through
 jTessBoxEditor and Tesseract optical
 character recognition (OCR).

Souphanouvong University Journal Multidisciplinary Research and Development, 9(4), 239–250.

<http://www.su-journal.com/index.php/su/article/view/523>

Popović, M. (2015). chrF: character n-gram F-score for automatic machine translation evaluation. *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 392-395.

Zhao, X., Liu, C. H., Vylomova, E., Pirinen, F., Washington, J., & Oco, N. (Eds.).
 (2025). *Proceedings of the Eighth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2025)*. Association for Computational Linguistics.
<https://aclanthology.org/volumes/2025.loresmt-1/>

ຕາຕະລາງທີ 1. ການອອກແບບ Prompt ແບບ Few-shot ຄຳສັບທົ່ວໄປ

ທິດທາງ: ລາວ → ຈີນ	ທິດທາງ: ຈີນ → ລາວ
Lao: ສະບາຍດີ	Chinese: 你好
Chinese: 你好	Lao: ສະບາຍດີ
Lao: ຂອບໃຈ	Chinese: 谢谢
Chinese: 谢谢	Lao: ຂອບໃຈ
Lao: ໂຊກດີ	Chinese: 祝你好运
Chinese: 祝你好运	Lao: ໂຊກດີ
Lao: {text}	Chinese: {text}
Chinese:	Lao:

ຕາຕະລາງທີ 2. ການອອກແບບ Prompt ແບບ Few-shot ຄຳສັບທ່ອງທ່ຽວ-ລົດໄຟ

ທິດທາງ: ລາວ → ຈີນ	ທິດທາງ: ຈີນ → ລາວ
Lao: ສະຖານີລົດໄຟ	Chinese: 火车站
Chinese: 火车站	Lao: ສະຖານີລົດໄຟ
Lao: ປີ້ລົດໄຟ	Chinese: 火车票
Chinese: 火车票	Lao: ປີ້ລົດໄຟ
Lao: ສະຖານທີ່ທ່ອງທ່ຽວ	Chinese: 旅游景点
Chinese: 旅游景点	Lao: ສະຖານທີ່ທ່ອງທ່ຽວ
Lao: {text}	Chinese: {text}
Chinese:	Lao:

ຕາຕະລາງທີ 3. ຕົວຢ່າງ ການປະໂຫຍກທົດສອບການແປ

ປະໂຫຍກທົ່ວໄປ (General Sentences)

ລາວ (Lao)	ຈີນ (Chinese)	ຈີນ (Chinese)	ລາວ (Lao)
ເຈົ້າສະບາຍດີບໍ່?	你好吗?	ບໍ່ຕ້ອງເກງໃຈ	不用客气
ມື້ນີ້ເປັນແນວໃດ?	今天怎么样?	ຂໍໃຫ້ມີຄວາມສຸກໃນມື້ນີ້	祝你今天愉快
ເຈົ້າຊື່ຫຍັງ?	你叫什么名字?	ເຂົ້າໃຈແລ້ວ	明白了

ປະໂຫຍກດ້ານການທ່ອງທ່ຽວ-ລຶດໄຟ

ລາວ (Lao)	ຈີນ (Chinese)	ຈີນ (Chinese)	ລາວ (Lao)
ປີ້ລຶດໄຟ	火车票	ວິວທິວທັດ	风景
ສະຖານີ ນະຄອນຫຼວງວຽງຈັນ	万象站	ທ່ອງທ່ຽວ	旅游
ທາງເຂົ້າສະຖານີ	进站口	ວັດ	寺庙

ຕາຕະລາງທີ 4. ຄະແນນ BLUE ການປະເມີນປະສິດທິພາບຕາມໂຕເມນ ແລະ ຈຳນວນປະໂຫຍກ

ໂຕເມນ	SEA-LION (10ປະໂຫຍກ)	SEA-LION (20ປະໂຫຍກ)	Qwen (10ປະໂຫຍກ)	Qwen (20ປະໂຫຍກ)
ທົ່ວໄປ (Lao→Zh)	7.93	10.72	17.65	12.19
ທົ່ວໄປ (Zh→Lao)	2.65	3.00	3.38	1.36
ທ່ອງທ່ຽວ-ລຶດໄຟ (Lao→Zh)	10.72	3.48	2.34	1.10
ທ່ອງທ່ຽວ-ລຶດໄຟ (Zh→Lao)	2.83	1.49	8.17	4.40

ໂຕເມນ	SEA-LION (50ປະໂຫຍກ)	SEA-LION (100ປະໂຫຍກ)	Qwen (50ປະໂຫຍກ)	Qwen (100ປະໂຫຍກ)
ທົ່ວໄປ (Lao→Zh)	13.61	12.07	21.62	18.93
ທົ່ວໄປ (Zh→Lao)	2.87	1.61	0.88	0.90
ທ່ອງທ່ຽວ-ລຶດໄຟ (Lao→Zh)	6.35	10.99	1.56	1.93
ທ່ອງທ່ຽວ-ລຶດໄຟ (Zh→Lao)	0.75	0.79	1.43	0.63