ThaiAcadLaws: A Legal Text Classification on Academic Obligation Domain using Text-to-Text Translating Encoder with GPT-3 Fine-tuning Decoder

Chantarat Kingsaeng[1], Lawankorn Mookdarsanit[2] and Pakpoom Mookdarsanit[3*]

[1, 3*]Computer Science and Artificial Intelligence, Faculty of Science
Chandrakasem Rajabhat University
Bangkok, Thailand
[1]chantarat.k@chandra.ac.th, [3*]pakpoom.m@chandra.ac.th

[2]Business Information Systems, Faculty of Management Science
Chandrakasem Rajabhat University
Bangkok, Thailand
[2]lawankorn.s@chandra.ac.th

Abstract—This paper contributed a legal AI for Thai academic obligation domain, named ThaiAcadLaws as a new paradigm of Thai-NLP research areas. Since the academic obligation defined by Thai higher education had 5 tasks: teaching, researching, academic service, cultural art preservation and other tasks, ThaiAcadLaws architecture consisted of text-to-text translating encoder and GPT-3 fine-tuning decoder was designed for classifying those 5 target classes. The obligation data was 2,560 Thai texts. Each class had 512 texts. The text-to-text translating encoder was based on SCB-MT-EN-TH 2020 zero-shot learning to translate the text, while GPT-3 fine-tuning decoder was few-shot learning from the translated text. The state-of-the-art results were 0.75 averaged accuracy for legal text classification. ***(Abstract)***

Keywords-Thai Text Classification; Legal AI; Fine-tuned GPT-3; Legal Informatics; Few-shot learning (key words)

I. INTRODUCTION

Legal informatics has been defined by the American Library Association (ALA) as the use of computer technology to automate legal tasks or environments [1]. In the artificial general intelligence (AGI) era, computers with big data are affected by not only office automation with documentation or web-based information systems but also autonomous intelligence with expert and recommender systems [2], as "legal AI".

All verdicts of the Thai Dika Court were initially collected systematically in a TCXML web-based application and dataset [3]. Legal informatics was changed from conventional web-based applications to semantic webs that use ontology as a knowledge-based system. Ontology was designed to conveniently facilitate text retrieval [4] in the form of law knowledge [5] and/or supreme court sentence retrieval [6] (text retrieval was the main originality of "prompting" or "prompt engineering" in the AGI era). Some researchers used the ATOB algorithm [7–8] to create the ontological structure for legal informatics. In big data analytics, there were so many legal reasoning experts [9–10] and/or illustration systems [11–12]. As to the machine learning algorithms, the TCXML dataset was modeled to discover the main knowledge of legal reasoning [13] in 2008, and the GUI was used as a decision support system for lawyers [14]. Artificial neural networks (ANNs) were so powerful for identifying criminal law sentences [15–16] that they played the main role in deep learning. In the deep learning era, a Bi-GRUs with Attention layer was used to predict the court's judgment [17], which was adapted from natural language processing (NLP) with embedding techniques to formulate each word or phrase into a vector representation. In 2023, the transfer adaptation learning of large language models (LLM) was first applied to a law retrieval system with text augmentation based on few-shot fine-tuning in the downstream task [18].

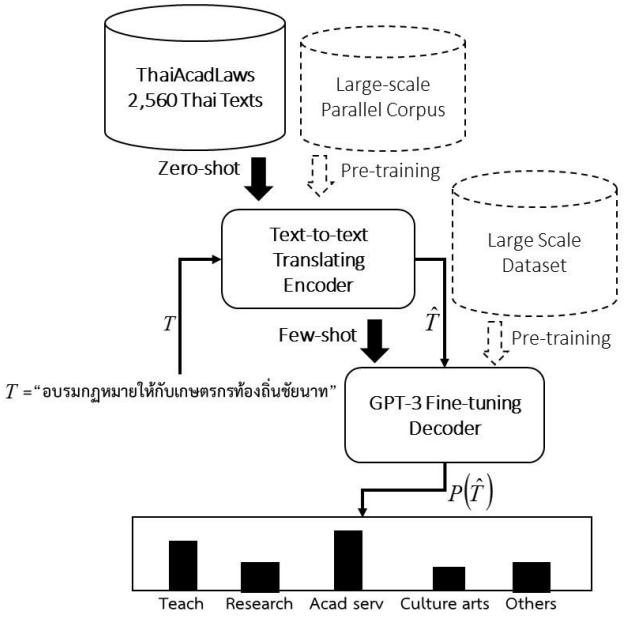

Figure 1. ThaiAcadLaws architecture, the input is a Thai text (*T*) to define the probaility of academic obligation domain

To another move on, this paper contributed a novel "ThaiAcadLaws" to classify the 5 academic obligations in Thailand (teaching, researching, academic service, cultural art preservation and other tasks), based on 2,560 Thai texts. The academic obligation laws were used by the 69,000 lecturers in the Thai higher education system. A Thai text input was fed into the text-to-text translating encoder based on the SCB-MT-EN-TH-2020 transformer [19]. And the fine-tuned GPT-3 decoder [20] classified the target class from the translated text. To the best of our knowledge, this paper was intentionally proposed to the three main contributions.

- This paper firstly introduced a novel ThaiAcadLaws for the academic obligation domain.
- The GPT pre-trained model with NLP was used for fine-tuning inThai laws data.
- ThaiAcadLaws would be one of Thai legal AI applications in Thai language and NLP.

The organization of this paper was Thai-NLP towards legal AI in the section 2. Section 3 explained ThaiAcadLaws architecture. The experimental results and conclusion were in section 4 and 5.

II. THAI-NLP TOWARDS LEGAL AI

Prior to Thai natural language processing (Thai-NLP), computers and Thai languages [21] had been research topics for more than 20 years. Initially, the research vision was to originally make computers process the information in Thai, especially Thai optical recognition (Thai-OCR) and Thai fonts [22-23]. Some algorithms were proposed for Thai-OCR problems; Bi-LSTMs with connectionist temporal classification (CTC) showed the state-of-the-art report [24]. Since Thai-OCR seemed to be no longer challenging (as well as the Captcha was disrupted by the reCaptcha [25]), Thai handwritten recognition was later a main local problem. BEST hackathon [26] has been still organized by Thailand's National Electronic Computer and Technology Centre (NECTEC) for challenging Thai researchers and students to compete with their designed handwritten recognition algorithms as a new way of preserving language in digital form. Toward generative artificial intelligence (GenAI), those Thai handwritten characters were not only recognized but also generated, called ThaiWritableGAN [27], which could be a new hackathon to allow Thai researchers and students to compete their proposed Thai calligraphic algorithms.

In the AGI age, Thai-NLP was to make computers not only retrieve Thai textual information but also understand the textual Thai language. S-Sense [28] was a Thai text classification baseline for sentiment analysis, developed by NECTEC researchers. Several Thai text classification applications were available based on Thai texts, e.g., fake news detection [29-30], soft skills classification [31], HR intelligence [32], Thai stock sentiment classification [33], and hate speech detection [34-35]. Some researchers further applied Thai-OCR with text classification [36] to detect hate speech from memes. Image and textual caption were in a strong relationship, and the computer algorithms could learn and understand the image content from a caption. There were many million images without Thai captions (as well as captions without images). Thai image captioning (Thai-IC) was first proposed to solve this crisis problem [37] by VGGNet-LSTM (with a little vanishing gradient), and the gradient was fixed by the Transformer encoder [38]. As to the enhanced image captioning, the Thai prompt was designed to create an image [39] (called "Thai text-to-image") based on a pre-trained language model (pLM) with stable diffusion. Thai-IC was not only an image-text relationship but also a Thai textual song lyrics recommender system [40].

The introduced ThaiAcadLaws was such a Thai text classification towards legal AI that was proposed to make computer understand Thai text content (as well as [17]). Since there were many researches in Thai legal informatics, most of them were just information retrieval based on ontology by textual query (but not from the text understanding). This paper proposed a deep learning algorithm to identify the 5 academic obligations based on the Thai text input.

III. THAIACADLAWS ARCHITECTURE

ThaiAcadLaws architecture was shown in Fig.1. The academic obligation data was firstly collected. Then, those data was to text-to-text translating encoder based on SCB-MT-EN-TH-2020 zero-shot learning; and it was finally sent to GPT-3 fine-tuning to classify the target class.

A. The Academic Obligation Data

There were 2,560 Thai texts used in this paper that were crawled and cleaned based on the university projects' topics and/or their projects' purposes during the 2015–2021 annual project report. There were 5 target classes that consisted of teaching, researching, academic service, cultural art preservation, and other tasks. Each Thai text was manually supervised or tagged by a human. Each target class had 512 texts, by (1). The target classes were academic obligation domains that were used by the 69,000 lecturers as the main laws of the Thai higher education system.

(1)

B. Text-to-Text Translating Encoder

SCB-MT-EN-TH-2020 transformer was used as text-to-text encoder to translate Thai ( ) to English ( ).

Text-to-text translating encoder in ThaiAcadLaws' architecture ( ) was direct zero-shot learning that did not need tokenizing algorithm to segment words in Thai text; and represent Thai word in term of vector (Thai Word2Vec). All 2,560 Thai texts were fed to text-to-text translating encoder, defined by (2). And those Thai texts with their target classes were later fed to GPT-3 fine-tuning decoder (or few-shot learning) in Fig.2.

(2)

Originally, the SCB-MT-EN-TH-2020 transformer was [19] an open large-scale machine translation that had 1,001,752 Thai-English parallel corpus, collected from and curated from various secondary sources, including Wikipedia, news, SMS messages, dialogs, web-crawled data, and government documents (available at https://huggingface.co/datasets/airesearch/scb_mt_enth_2020). The machine translation was evaluated by the universal sentence encoder multi-lingual (USEM) and bilingual evaluation understudy (BLEU).

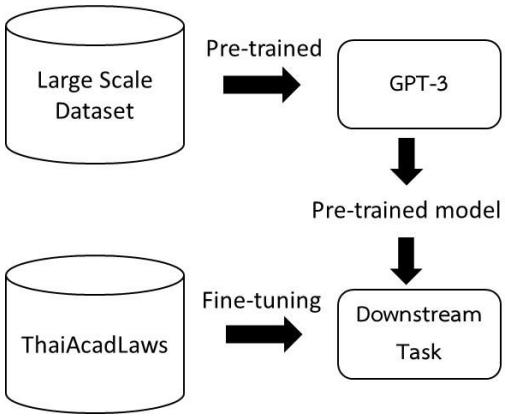

Figure 2. GPT-3 Fine-tuning (or few-shot learning) after text-to-text translating encoder by SCB-MT-EN-TH-2020 zero-shot learning

C. GPT-3 Fine-tuning Decoder

The generative pre-trained transformer (GPT) [41] was just a decoder in Transformer with 12 layers, called the "autoregressive language model" that was defined by (3).

(3)

where in a text , was the previous token before token that could be illustrated in Fig.3.

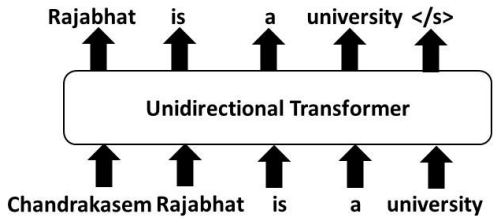

Figure 3. GPT autogressive large language model

GPT was pre-trained by 40 GB of text data from the book corpus in 117 million parameters. GPT-2 [42] was trained on a much larger dataset of 8 million web pages (40 GB of text) from the internet with 1.5 billion parameters. GPT-3 [20] was trained on 570 GB of text data from diverse sources with 175 billion parameters, which was much larger than GPT-2, which was scaled up to a 96-layer Transformer architecture. This paper applied GPT-3 as fine-tuning decoder, as shown in Fig.4.

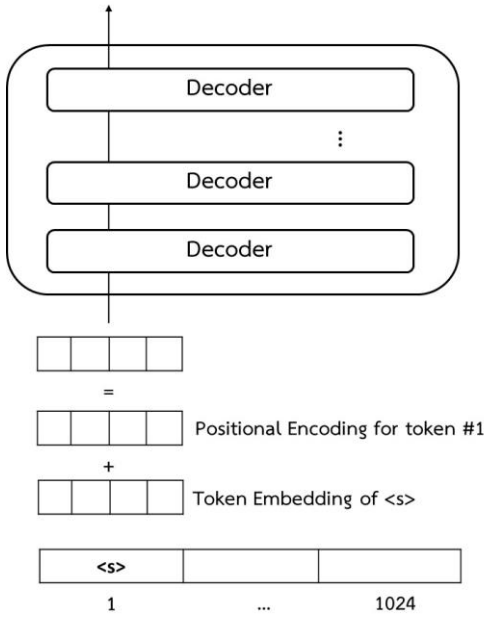

Figure 4. GPT-3 architecture as ThaiAcadLaws fine-tuning (or few-shot) decoder

IV. EXPERIMENTATIONAL RESULTS

ThaiAcadLaws architecture was designed to classify a Thai text in 5 different target classes: teaching, researching, academic service, cultural art preservation, and other tasks. The encoder architecture was measured by BLEU and the decoder was done by accuracy.

A. Zero-shot encoder's BLEU

The bilingual evaluation understudy (BLEU) measured the adequacy and fluency of translated text ( ). BLEU compared between machine and human translation.

$$BLEU = \min\left(1, \frac{length_{machine}}{length_{human}}\right) \times \left(\prod_{i=1}^{n} precision_{i-gram}\right) \quad (4)$$

The BLEU results was shown in Table 1 that adopt precision metrics in n-Gram model (where n =1, 2, 3, 4), by (4)

TABLE I. ENCODER BLEU COMPARRISON BETWEEN N-GRAM

n-Gram	Zero-shot Encoder
BLEU-1	0.72
BLEU-2	0.58
BLEU-3	0.46
BLEU-4	0.31

B. Few-shot decoder's Accuracy

The GPT-3 fine-tuning (or few-shot learning) decoder was proposed to classify a translated text into 5 target classes that was measured by accuracy in (5).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

The classification results was shown in Table 2. The "Researching" obligation was the highest accuracy since the task was clearly different from other obligations. Academic service and teaching sometimes seemed to not be different. However, other tasks were the lowest accuracy as many texts were not clear definition of tasks.

TABLE II. DECODER ACCURACY RESULTS

Obligation	Few-shot Decoder
Teaching	0.75
Researching	0.85
Academic service	0.71
Cultural art preservation	0.78
Other tasks	0.67
Average	0.75

V. CONCLUSION

Since Thai academic staffs had to do 5 legal obligations consisted of teaching, researching, academic service, cultural art preservation and other tasks, this paper contributed ThaiAcadLaws architecture to classify a Thai text into 5 target classes. ThaiAcadLaws had text-to-text encoder based on SCB-MT-EN-TH-2020 transformer for zero-shot translation. And GPT-3 fine-tuning decoder for few-shot classification, respectively. The encoder and decoder were measured by BLEU and accuracy. The BLEU was evaluated in different n-Gram. And the averaged accuracy was 0.75. For future works and extensions, the text preparation with synonyms could be directly represented by a vector that might be leverage the higher accuracy.

ACKNOWLEDGMENT

ThaiAcadLaws research was designed to classify a Thai text into 5 different academic obligations: teaching, researching, academic service, cultural art preservation, and other tasks. All resources were supported by Chandrakasem Rajabhat University.

REFERENCES

[1] S. Erdelez and S. O'Hare, "Legal Informatics: Application of Information Technology in Law," Annual Review of Information Science and Technology, vol. 32, pp. 367, 1997.

[2] L. Soimart and P. Mookdarsanit, "An Admission Recommendation of High-school Students using Apriori Algorithm," in *Proceedings of the 6th International Conference on Sciences and Social Sciences*, Sep. 2016.

[3] S. Thammaboosadee and U. Silparcha, "TCXML for Collection of Verdicts of Thai Dika Court," in *Proceedings of the National Conference on Information Technology*, Nov. 2006, pp. 179-186.

[4] T. Tantisripreecha and N. Soonthornphisaj, "A study of Thai succession law ontology on supreme court sentences retrieval," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, 2010.

[5] P. Osathitporn, N. Soonthornphisaj, and W. Vatanawood, "A scheme of criminal law knowledge acquisition using ontology," in *Proceedings of the 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, Jun. 2017, pp. 29-34.

[6] T. Tantisripreecha and N. Soonthornphisaj, "Supreme court sentences retrieval using Thai law ontology," *Intelligent Control and Computer Engineering*, pp. 177-189, 2011.

[7] V. Boonchom and N. Soonthornphisaj, "Legal ontology construction using ATOB algorithm," in *Proceedings of Business Information Systems Workshops: BIS 2010 International Workshops*, Berlin, Germany, May 3-5, 2010, Revised Papers, vol. 13, pp. 268-279.

[8] V. Boonchom and N. Soonthornphisaj, "ATOB algorithm: an automatic ontology construction for Thai legal sentences retrieval," *Journal of Information Science*, vol. 38, no. 1, pp. 37-51, Feb. 2012.

[9] T. Tantisripreecha and N. Soonthornphisaj, "Creating rules using abduction for legal reasoning by logic programming," in *Proceedings of Business Information Systems Workshops: BIS 2011 International Workshops and BPSC International Conference*, Poznań, Poland, Jun. 15-17, 2011, Revised Papers, vol. 14, pp. 282-293.

[10] T. Tantisripreecha, K. Satoh, and N. Soonthornphisaj, "Legal reasoning engine for civil court procedure," in *Proceedings of Intelligent Computing Methodologies: 10th International Conference, ICIC 2014*, Taiyuan, China, Aug. 3-6, 2014, Proceedings, vol. 10, pp. 500-512.

[11] T. Tantisripreecha and N. Soonthornphisaj, "LASTC: Legal Advisory System for Thai Cheque Law," in *Proceedings of New Perspectives in Information Systems and Technologies*, vol. 1, 2014, pp. 503-512.

[12] T. Tantisripreecha and N. Soonthornphisaj, "LegalEX: An expert system for law firm," *Intelligent Decision Technologies*, vol. 10, no. 3, pp. 315-328, Jan. 2016.

[13] S. Thammaboosadee and U. Silparcha, "A framework for criminal judicial reasoning system using data mining techniques," in *Proceedings of the 2nd IEEE International Conference on Digital Ecosystems and Technologies*, Feb. 2008, pp. 518-523.

[14] S. Thammaboosadee and U. Sulparcha, "A GUI prototype for the framework of criminal judicial reasoning system," *Journal of International Commercial Law and Technology.*, vol. 4, pp. 224, 2009.

[15] S. Thammaboosadee, B. Watanapa, and N. Charoenkitkarn, "A framework of multi-stage classifier for identifying criminal law sentences," *Procedia Computer Science*, vol. 13, pp. 53-59, Jan. 2012.

[16] S. Thammaboosadee and B. Watanapa, "Identification of criminal case diagnostic issues: a modular ANN approach," *International Journal of Information Technology & Decision Making*, vol. 12, no. 3, pp. 523-546, May 2013.

[17] K. Kowsrihawat, P. Vateekul, and P. Boonkwan, "Predicting judicial decisions of criminal cases from Thai Supreme Court using bi-directional GRU with attention mechanism," in *Proceedings of the 5th Asian Conference on Defense Technology (ACDT)*, Oct. 2018, pp. 50-55.

[18] T. Chusri, S. Arsaibun, P. Chokesuwattanaskul, E. Chuangsuwanich, and A. T. Rutherford, "Few-Shot Law Retrieval System for Supreme Court Cases," in *Proceedings of the 20th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, Jun. 2023, pp. 84-89.

[19] L. Lowphansirikul, C. Polpanumas, A. T. Rutherford, and S. Nutanong, "A large English–Thai parallel corpus from the web and machine-generated text," *Language Resources and Evaluation*, vol. 56, no. 2, pp. 477-499, Jun. 2022.

[20] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877-1901, 2020.

[21] H. T. Koanantakool, T. Karoonboonyanan, and C. Wutiwiwatchai, "Computers and the Thai language," *IEEE Annals of the History of Computing*, vol. 31, no. 1, pp. 46-61, Mar. 2009.

[22] P. Mookdarsanit and L. Mookdarsanit, "ThaiWrittenNet: Thai Handwritten Script Recognition Using Deep Neural Networks," *Azerbaijan Journal of High Performance Computing*, vol. 3, no. 1, pp. 75-93, 2020.

[23] O. Surinta and L. Schomaker, "Overview of handwritten Thai character recognition," *Lecture Notes Online*, 2010.

[24] T. Emsawas and B. Kijsirikul, "Thai printed character recognition using long short-term memory and vertical component shifting," in *Proceedings of PRICAI 2016: Trends in Artificial Intelligence: 14th Pacific Rim International Conference on Artificial Intelligence*, Phuket, Thailand, Aug. 22-26, 2016, Proceedings, vol. 14, pp. 106-115.

[25] L. Mookdarsanit and P. Mookdarsanit, "An Adversarial Perturbation Technique against reCaptcha Image Attacks," *Journal of Science and Technology Buriram Rajabhat University*, vol. 4, no. 1, 2020.

[26] K. Kosawat, M. Boriboon, P. Chootrakool, A. Chotimongkol, S. Klaithin, S. Kongyoung, K. Kriengket, S. Phaholphinyo, S. Purodakananda, T. Thanakulwarapas, and C. Wutiwiwatchai, "BEST 2009: Thai word segmentation software contest," in *Proceedings of the 8th International Symposium on Natural Language Processing*, Oct. 2009, pp. 83-88.

[27] L. Mookdarsanit and P. Mookdarsanit, "ThaiWritableGAN: Handwriting Generation under Given Information," *International Journal of Computing and Digital Systems*, vol. 10, no. 1, pp. 689-699, 2021.

[28] C. Haruechaiyasak, A. Kongthon, P. Palingoon, and K. Trakultaweekoon, "S-sense: A sentiment analysis framework for social media sensing," in *Proceedings of the IJCNLP 2013 Workshop on Natural Language Processing for Social Media (SocialNLP)*, Oct. 2013, pp. 6-13.

[29] P. Mookdarsanit and L. Mookdarsanit, "The COVID-19 fake news detection in Thai social texts," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 2, pp. 988-998, Apr. 2021.

[30] S. Aphiwongsophon and P. Chongstitvatana, "Identifying misinformation on Twitter with a support vector machine," *Engineering & Applied Science Research*, vol. 47, no. 3, Jul. 2020.

[31] L. Mookdarsanit and P. Mookdarsanit, "Thai NLP-based Text Classification of the 21st-century Skills toward Educational Curriculum and Project Design," *International Journal of Applied Computer Technology and Information Systems*, vol. 11, no. 2, pp. 62-67, 2022.

[32] L. Mookdarsanit and P. Mookdarsanit, "The Insights in Computer Literacy toward HR Intelligence: Some Associative Patterns between IT Subjects and Job Positions," *Journal of Science and Technology RMUTSB*, vol. 4, no. 2, pp. 12-23, 2020.

[33] A. Chattupan and P. Netisopakul, "Thai stock news sentiment classification using wordpair features," in *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, 2015, pp. 188-195.

[34] P. Mookdarsanit and L. Mookdarsanit, "TGF-GRU: A Cyber-bullying Autonomous Detector of Lexical Thai across Social Media," *NKRAFA Journal of Science and Technology*, vol. 15, no. 1, pp. 50-58, 2019.

[35] S. Hemtanon, K. Phetkrachang, and W. Yangyuen, "Classification and keyword extraction of online harassment text in Thai social network," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 6, pp. 3837-3842, Dec. 2023.

[36] L. Mookdarsanit and P. Mookdarsanit, "Combating the hate speech in Thai textual memes," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 3, pp. 1493-1502, 2021.

[37] P. Mookdarsanit and L. Mookdarsanit, "Thai-IC: Thai Image Captioning based on CNN-RNN Architecture," *International Journal of Applied Computer Technology and Information Systems*, vol. 10, no. 1, pp. 40-45, 2020.

[38] K. Dittakan, K. Prompitak, P. Thungklang, and C. Wongwattanakit, "Image caption generation using transformer learning methods: a case study on Instagram image," *Multimedia Tools and Applications*, vol. 83, pp. 46397–46417, Oct. 2023.

[39] P. Mookdarsanit and L. Mookdarsanit, "Thai Text-to-Image Prompt Engineering by Pre-trained Large Language with Stable Diffusion Model," *Azerbaijan Journal of High Performance Computing*, vol. 6, no. 2, pp. 171-190, 2023.

[40] N. Sanguansub, P. Kamolrungwarakul, S. Poopair, K. Techaphonprasit, and T. Siriborvornratanakul, "Song lyrics recommendation for social media captions using image captioning, image emotion, and caption-lyric matching via universal sentence embedding," *Social Network Analysis and Mining*, vol. 13, article no. 95, Jun. 2023.

[41] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," OpenAI, 2018.A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, Feb. 2019.