Azerbaijan Journal of High Performance Computing, Vol 6, Issue 2, 2023, pp. 171-190
https://doi.org/10.32010/26166127.2023.6.2.171.190

# Thai Text-to-Image Prompt Engineering by Pre-trained Large Language with Stable Diffusion Model

Pakpoom Mookdarsanit, Lawankorn Mookdarsanit

*Chandrakasem Rajabhat University, Bangkok, Thailand, pakpoom.m@chandra.ac.th, lawankorn.s@chandra.ac.th*

*Correspondence: Pakpoom Mookdarsanit, Chandrakasem Rajabhat University, Bangkok, Thailand, pakpoom.m@chandra.ac.th,

## Abstract

Text-to-image (T2I) generation is a new area of large language models (LLMs), a type of prompt engineering involving inputting a textual description to generate an image. To shift a new paradigm of Thai natural language processing (Thai-NLP), this paper first presents state-of-the-art Thai Text-to-Image prompt engineering (TH-T2I) to translate Thai text into a semantic image according to the semantic Thai textual description. The pre-trained SCB-MT-EN-TH model is employed for Text-to-Text (T2T) translation. Moreover, the image generation is done according to a semantic text prompt by a stable diffusion model. The T2T is evaluated by Bi-lingual Evaluation Understudy (BLEU), while T2I is done by Inception and Frechet Inception Distance (FID). The images generated by TH-T2I were of high quality, as measured by Inception and FID. TH-T2I contributes to a T2I baseline model in Thai, preserving the Thai cultural language on digital heritage.

**Keyword:** Text-to-Image Translation, Thai Prompt Engineering, Stable Diffusion Model, Image Generation

### *1. Introduction*

Although many large language models (LLMs) have been recently introduced, some low-resource languages must be generative artificial intelligence (AGI). Unlike English (one of the high-resource languages), Thai is one of the low-resource languages (Arreerard, Mander & Piao, 2022). that is locally spoken in Thailand and the surrounding Mekong Golden Triangle region. As to some LLMs and Thai, there were three well-known Transformer-based models: WangchanBERTa - a pre-trained Thai text categorization (Lowphansirikul, Polpanumas, Jantrakulchai & Nutanong, 2021)., as well as PhayaThaiBERT (Sriwirote, Thapiang, Timtong & Rutherford, 2023). and SCB-MT-EN-TH model (Lowphansirikul, Polpanumas, Rutherford & Nutanong, 2020). - a large-scale Thai-English machine translation (Lowphansirikul, Polpanumas, Rutherford & Nutanong, 2022). To enhance the conventional T2T translation model (e.g., text categorization, TH-EN machine translation) into T2I, Thai Text-to-Image (TH-T2I), prompt engineering was presented as one of the new challenges of Thai natural language processing areas that

are proposed to translate the Thai text into an image. For the academic movement beyond education (Mookdarsanit & Mookdarsanit, 2022)., TH-T2I could be a baseline model in T2I on Thai for future Thai AI competitions (e.g., BEST Hackathon organized by NECTEC) that were proposed to preserve Thai as the digital heritage inherited by the next generation of Thai AGI researchers. As well as planting the Thai digital Treebank, future digital donations might be in Thai speech, handwriting, or textual comments on social media to make Thai a fruitful resource and library comparable to other high-resource languages.

### *1.1. Thai-NLP Timeline*

Thai (one of the Kra-dai family) has been used as a local language in Thailand and the Mekong Golden Triangle region for longer than 720 years since the Sukothai stone tablet inscribed in Thai scripts by King Ramkhamhaeng (Inthajakra, Prachyapruit & Chantavanich, 2016). The tablet has been declared a World Heritage by UNESCO since 2003. Thai vocabulary acquired from Sanskrit, Pali, Khmer, and Mon. Linguistically, Thai is a native language spoken by 70 million people. Thai is a tonal language in speech that implies different meanings, which is one of the challenges in the 5G penetration test for Thai tonal speech. To give an example, the word “Ma” in Thai has five different tones: the first tone (Thai: มา, v.) means coming, the second tone (Thai: ม่า, n.) means grandmother, the third tone (Thai: ม้า, n.) as horse, the fourth (Thai: ม๊า, n.) as mother and fifth tone (Thai: หมา, n.) as dog, respectively. In addition, all scripts within a Thai text have no space between 2 words or phrases (Lapjaturapit, Viriyayudhakom & Theeramunkong, 2018). that needs some tokenization algorithm for separating between words/phrases (Klahan, Pannoi, Uewichitrapochana & Wiangsripanawan, 2018).

The first idea of Thai-English (TH-EN) machine translation (Koanantakool, Karoonboonyanan & Wutiwiwatchai, 2009). The Multilingual Machine Translation project (MMT) was proposed to automatically interpret between 2 languages in 1987 by Thailand's National Electronics and Computer Technology (NECTEC) as the origin of Thai natural language processing (Thai-NLP). The first LEXiTRON by NECTEC was launched in 1995 and had 11,000 Thai and 9,000 English entries (Sornlertlamvanich, 2019). Simultaneously, not only was the first Thai font introduced, known as Thai optical character recognition (Thai-OCR), but also the more complex Thai handwritten digit recognition (Thai-HDR). Both Thai-OCR (Emsawas & Kijsirikul, 2016). and Thai-HDR (Mookdarsanit & Mookdarsanit, 2020b). It could be seen as the Thai NLP meets Computer Vision. In the deep learning age, Thai-OCR had no more challenges (recognizing Thai characters from the image and understanding semantic text). Thai-OCR was expanded to meme image categorization (Mookdarsanit & Mookdarsanit, 2021a). Or scene text detection (Kobchaisawat, Chalidabhongse & Satoh, 2020). Thai-HDR recognizes the variety of Thai handwriting styles and generates different Thai handwriting styles written by AGI (Mookdarsanit & Mookdarsanit, 2021b).

From the previous literature, Thai-NLP has lasted longer than 35 years (Sornlertlamvanich, 2019). Researchers from NECTEC, Thammasat University, and Chulalongkorn University published many state-of-the-art Thai-NLP papers, respectively. Since the NECTEC was

founded with the vision of human language and computer (to make computers understand the Thai language in terms of "text," "speech", or "image"), many projects were provided for Thai industry (Tapsai, Unger & Meesad, 2020)., e.g., AI for Thai, OAM (a framework for design an ontology), Blackbeard Treebank, VAJA (as Text-to-Speech translation), BEST Hackathon. In 2000, a Thai character cluster (TCC) was designed to digitally group Thai characters (Theeramunkong, Sornlertlamvanich, Tanhermhong & Chinnan, 2000) by researchers from Thammasat University which highly impacted new Thai font construction (e.g., Chulabhorn Likit Font - dedicated to Princess Chulabhorn's continuously work for cancer patients in Thailand). Many researchers from Chulalongkorn University also played the leading role in contributing to Thai-NLP publications and available projects: AKARAWISUT (as a Thai plagiarism detector), GOWAJEE (Thai Speech Recognition), Thai-dependency Treebank, and Thai Speech-emotion Dataset.

Above all, Thai-NLP areas have been continuously developed. They could be categorized into text categorization (Mookdarsanit & Mookdarsanit, 2019)., sentiment analysis (Haruechaiyasak, Kongthon, Palingoon & Trakultaweekoon, 2013)., part of speech tagging (Boonkwan & Supnithi, 2017)., human resource language intelligence (Mookdarsanit & Mookdarsanit, 2020b)., plagiarism detection (Taerungruang & Aroonmanakun, 2018)., news summarization (Ketui, Theeramunkong & Onsuwan, 2013)., fake news detection (Mookdarsanit & Mookdarsanit, 2021b)., and TH-EN machine translation (Lowphansirikul, Polpanumas, Rutherford & Nutanong, 2022). To shift a new paradigm of Thai-NLP area, this paper presented a novel Thai Text-to-Image (TH-T2I) prompt engineering to generate an image according to Thai text as contributing a new baseline LLM enhancing from TH-EN machine translation.

### *1.2. Presented TH-T2I as a New Paradigm of Thai-NLP*

TH-EN machine translation could be seen as a text-to-text (T2T) translation that took a long time for the model to be completed entirely. Fluency and adequacy were the T2T translation measurements. From 2006 to 2016, there were attempting to design the completed TH-EN machine translation in the domain of law (Tirasaroj, 2016). and stock exchange (Ruangrajitpakorn, 2006). from Chulalongkorn University. However, statistical machine translation (SMT) was not enough for a significant linguistic data era, and TH-EN machine translation required deep neural network models, known as large language models (LLMs). Researchers from the Vidyasirimedhi Institute of Science and Technology (VISTEC) played the leading role in LLMs for Thai-NLP. In 2020, VISTEC researchers introduced a Transformer-based translator called the "SCB-MT-EN-TH model" (Lowphansirikul, Polpanumas, Rutherford & Nutanong, 2022). that were pre-trained by 1,001,752 TH-EN parallel texts. Although either ChatGPT or Google Bard were pre-trained by larger-scale text, the SCB-MT-EN-TH model was constructed by native researchers (Lowphansirikul, Polpanumas, Rutherford & Nutanong, 2020).

Some TH-EN large-language translation machine models were available as Text-to-Text (T2T) prompt engineering. To shift a new paradigm of Thai-NLP, this paper expanded

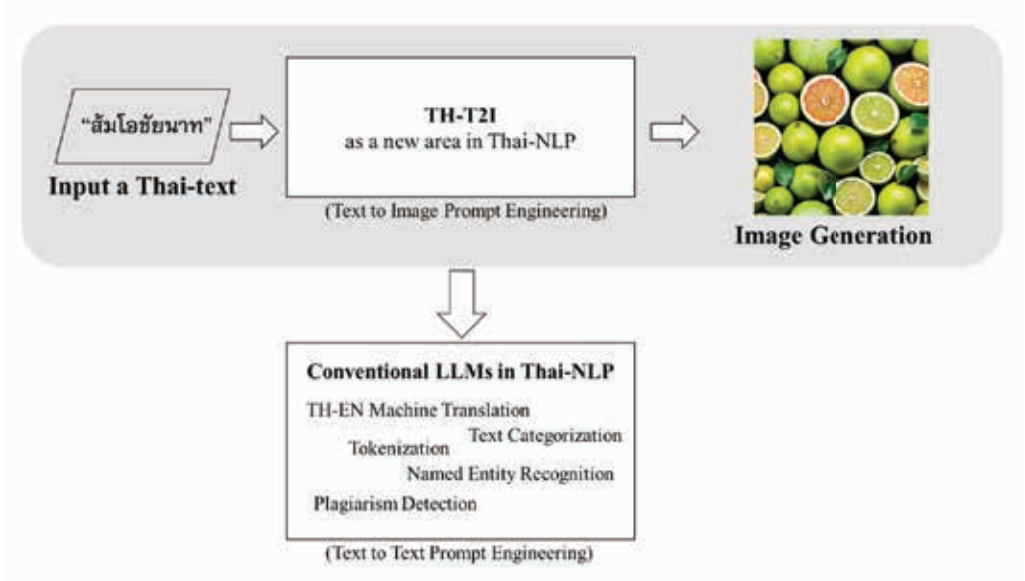


*Fig. 1: They presented TH-T2I as a new paradigm of Thai-NLP research.*

T2T to Text-to-Image (T2I) translation. T2I was an inverted image captioning (IC). Thai image captioning (Thai-IC) was interesting (Mookdarsanit & Mookdarsanit, 2020a) before the birth of diffusion models (Ho, Jain & Abbeel, 2020). IC was to generate a textual caption from an input image, while T2I generated an image from a textual description. The image generation was done by a stable diffusion model (Rombach, Blattmann, Lorenz, Esser & Ommer, 2021). The presented algorithm was called Thai Text-to-Image prompt engineering (TH-T2I), which was the meeting between Thai-NLP and computer vision (Lee, Hoover, Strobelt, Wang, Peng, Wright, Li, Park, Yang, Chau, 2023). As motivated by Thai-dessert image synthesis (Mookdarsanit & Mookdarsanit, 2018d)., TH-T2I could be further enhanced in many semantic Image-Text relations in Thai culture and art (Mookdarsanit & Rattanasiriwongwut, 2017c). Available on large-scale social media: street surveillance (Sutthaluang & Prakancharoen, 2020)., plant recognition (Mookdarsanit & Mookdarsanit, 2019b)., image location estimation (Mookdarsanit & Rattanasiriwongwut, 2017b)., Buddhism and temples (Mookdarsanit & Rattanasiriwongwut, 2017a)., food image description (Soimart & Mookdarsanit, 2017a)., tourism classification (Mookdarsanit & Mookdarsanit, 2018c)., GPS place estimation (Soimart & Mookdarsanit, 2017b), pesticide analytics (Sutthaluang, 2019)., agricultural product quality (Mookdarsanit & Mookdarsanit, 2021a)., facial verification and recognition (Soimart & Mookdarsanit, 2016)., and cosmetic recommender systems (Mookdarsanit & Mookdarsanit, 2023).

The significant contribution of TH-T2I is abridged as:

- TH-T2I shifted a new paradigm of Thai-NLP research areas that would be one of Thai-NLP timelines (as well as Thai-IC).
- TH-T2I expanded from T2T to T2I translation that combined Thai-NLP and stable

diffusion model.

- TH-T2I would be a T2I baseline model in Thai, especially in AI competitions (e.g., BEST Hackathon).

- TH-T2I supported the preservation of Thai cultural language on digital heritage (that could be inherited by the next generation of Thai AGI researchers), comparable to other high-resource languages.

This paper is organized into five parts. Attention mechanism and pre-trained significant language with a stable diffusion model were described in parts 2 and 3. Part 4 discussed experimental evaluations and results. Finally, part 5 was the conclusion.

### *2. Attention mechanism*

Since statistical machine translation (SMT) was unsuitable for linguistic big data, Transformer architecture (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser & Polosukhin, 2017). was introduced by Google in 2017 to handle parallel translations of a large number of languages on Google. As to the vanishing gradient in processing long text by the recurrent neural network, the Transformer was a sequence-to-sequence (seq2seq) designed for large-scale machine translation. A transformer could be seen as an encoder-decoder model, like SMT. One of the essential mechanisms in the Transformer was attention. The presented TH-T2I used self-attention (as a multi-head parallel block within the encoder and decoder) and mask self-attention (as a multi-head parallel block within the decoder) in the T2T part and self-attention in the T2I part.

#### *2.1. Self-attention*

Prior to self-attention, the attention concept was to focus the tokens (terms or words/phrases) within the text because the main idea of a sentence depended on the weight of significant tokens. By the following steps, the attention should be computed by (1)

$$Attention(q,k,v) = soft\max\left(\frac{q \bullet k^T}{\sqrt{d}}\right) \bullet v \tag{1}$$

where $Attention(\bullet)$ such a Softmax function, $k$ as "Key matrix $\begin{bmatrix} k_0 \\ k_1 \\ k_2 \end{bmatrix}$", $q$ as "Query matrix $\begin{bmatrix} q_0 \\ q_1 \\ q_2 \end{bmatrix}$" and $v$ as "Value matrix $\begin{bmatrix} v_0 \\ v_1 \\ v_2 \end{bmatrix}$" ($\left(\frac{q \bullet k^T}{\sqrt{d}}\right)$ represented by $a_{i,j} \bullet v_i$)

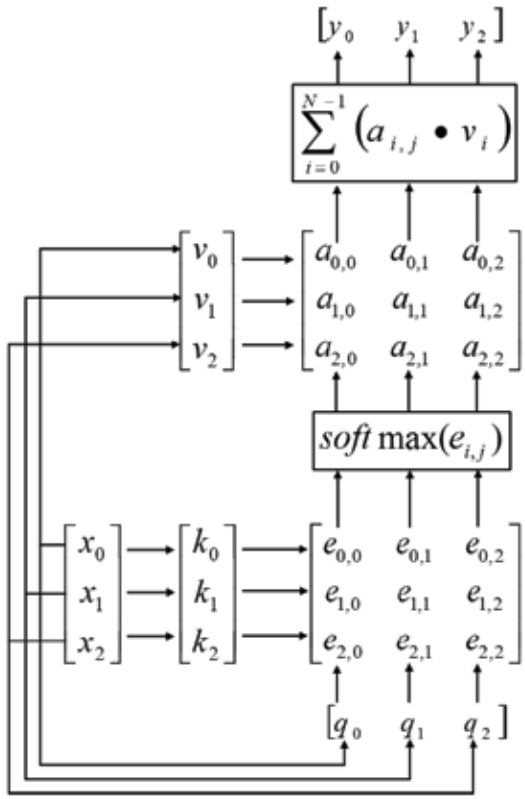


*Fig. 2: Self-attention mechanism in encoder and decoder (in form of multi-head attention); and generator within TH-T2I architecture.*

As to the self-attention in Figure 2, the input as "Input matrix $\begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix}$" was used to compute the "Key matrix $\begin{bmatrix} k_0 \\ k_1 \\ k_2 \end{bmatrix}$", "Query matrix $\begin{bmatrix} q_0 \\ q_1 \\ q_2 \end{bmatrix}$" and "Value matrix $\begin{bmatrix} v_0 \\ v_1 \\ v_2 \end{bmatrix}$", as (2)-(4)

$$k_i = W_k^T x_i \tag{2}$$

$$q_j = W_q^T x_i \tag{3}$$

$$v_i = W_v^T x_i \tag{4}$$

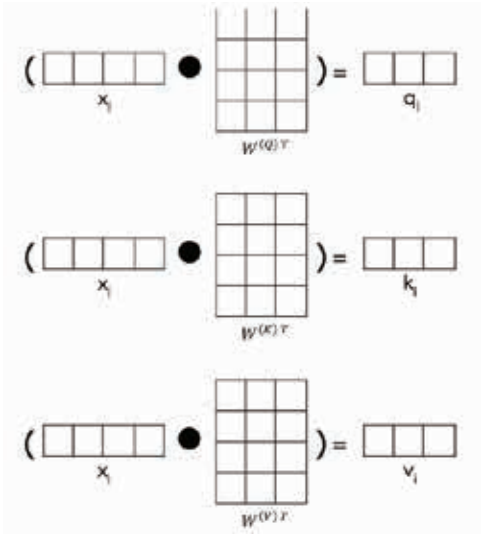

*Fig. 3: The illustration of Key matrix, Query matrix and Vector matrix by (2)-(4).*

First, suppose a Thai text had two tokens, token #0 and token #1; the computation explanation was in token #0. Both scores (by a2nd $q_0 k_1^T$ ) were assumed.

Then, in such an illustration of token #0, the score was divided by 8 (as called ) since the vectors had 64 dimensions.

Next, the Softmax function would be computed to make the range value between 0 and 1.

After that, the Softmax outputs were multiplied by $v_0$ (in the case of token #0)

Finally, the sum of the product was computed and called. $z_{00}$

In TH-T2I, self-attention was a composition in the encoder and decoder side of LLM (in the form of Multi-head self-attention) and in generator stable diffusion to build an image.

### *2.2. Masked Self-attention*

Masked self-attention was in the decoder side of LLM (in the form of Multi-head masked self-attention) in TH-T2I. The main difference between self-attention and masked self-attention was Alignment and Softmax computation.

Masked self-attention, alignment, and Softmax function computations could be defined by (5) and (6), respectively.

$$Alignment_{ij\ Masked} = \begin{cases} (Score/8)_{ij} & if\ i \geq j \\ -\infty & if\ i < j \end{cases} \quad (5)$$

$$Soft\max_{ij\ Masked} = \begin{cases} Soft\max_{ij} & if\ i \geq j \\ 0 & if\ i < j \end{cases} \quad (6)$$

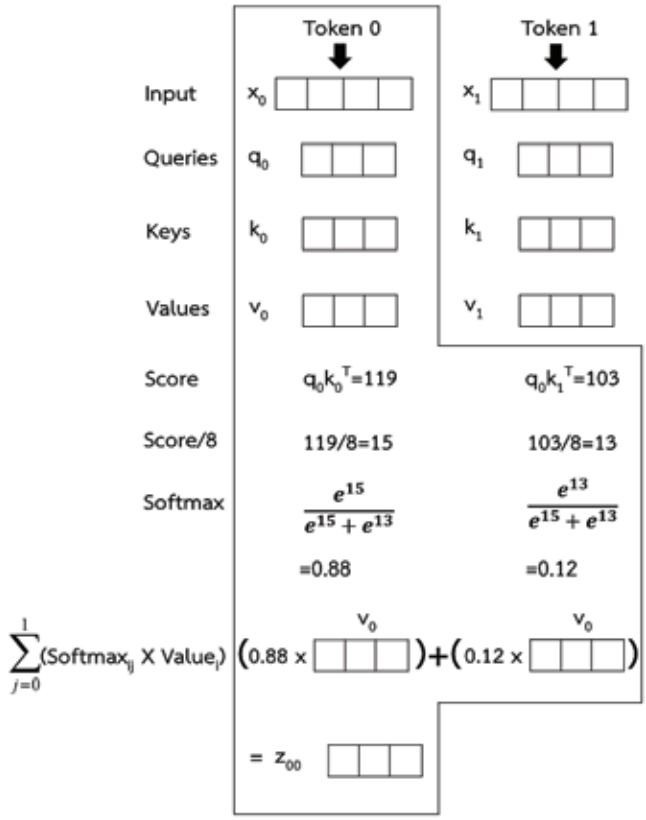


*Fig. 4: The illustration of a Thai word (e.g., "ส้มโอ") represented by a token #0 attention computation.*

The Masked self-attention architecture can be shown in Figure 5. Both self-attention and Masked self-attention could be in the form of a multi-head layer.

*3. Pre-trained Significant Language with a Stable Diffusion Model*

TH-T2I shifted a new paradigm of Thai-NLP areas by generating an image from a Thai text input; it was a combination of LLMs (Lowphansirikul, Polpanumas, Rutherford & Nutanong, 2022). and computer vision (Lee, Hoover, Strobelt, Wang, Peng, Wright, Li, Park, Yang, Chau, 2023). based on Transformer (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser & Polosukhin, 2017). In T2T translation, the SCB-MT-EN-TH pre-trained model was used. SCB-MT-EN-TH pre-trained model was a Transformer-based model for a large-scale TH-EN machine translator that was pre-trained by 1M TH-EN texts. This pre-trained model was the best TH-EN machine translator (compared to others, e.g., Google Bard and ChatGPT) constructed by VISTEC researchers. Transformers could be divided into encoder (for source language), decoder (for target language), respectively and generator (for image generation). And the stable diffusion generator (Rombach, Blattmann, Lorenz, Esser &

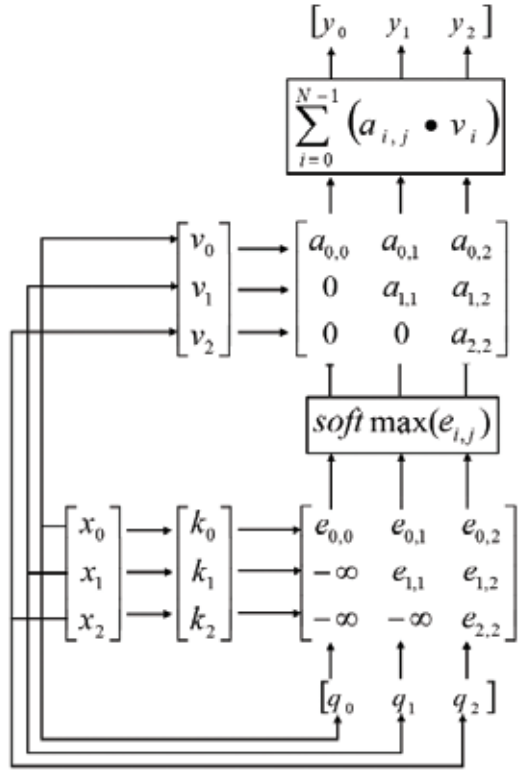


*Fig. 5: Masked self-attention mechanism in decoder (in form of multi-head attention) within TH-T2I architecture.*

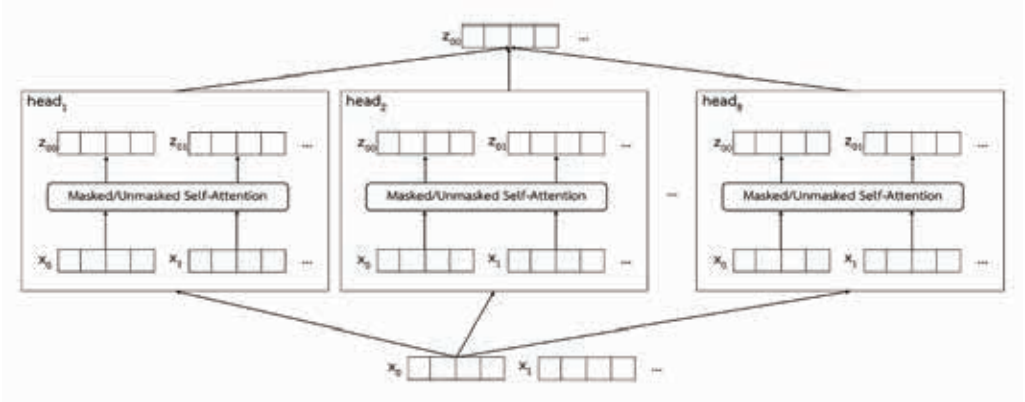


*Fig. 6: Multi-head attention of masked (or unmasked) wihin TH-T2I architecture.*

Ommer, 2021). was also based on a vision transformer (ViT) and used for T2I generation. The presented TH-T2I architecture was quickly shown in Figure 7. which could be further enhanced for human authentication by reCaptcha image generators (Mookdarsanit & Mookdarsanit, 2020a).

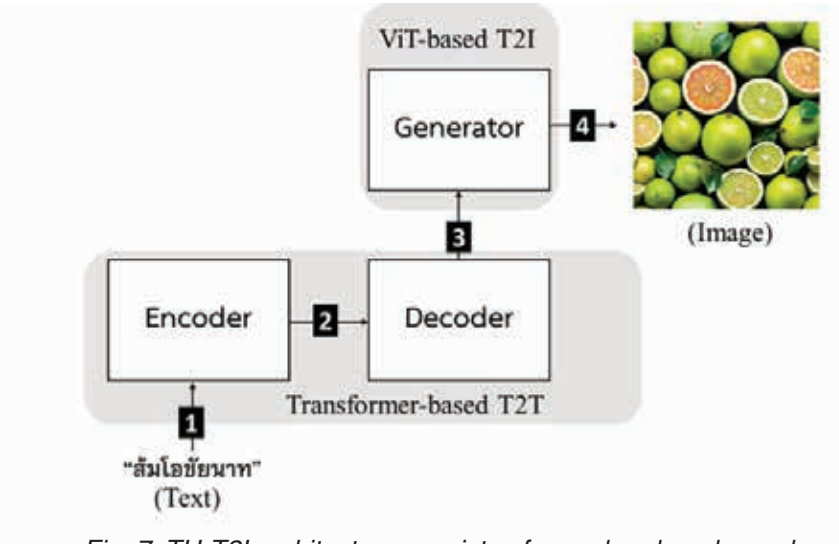


*Fig. 7: TH-T2I architecture consists of encoder, decoder and generator.architecture.*

### *3.1. Textual Transformer-based T2T Encoder*

Encoder (or translation model for inputting the source language as "Thai: TH") was a language input (Lowphansirikul, Polpanumas, Rutherford & Nutanong, 2020). that might have word and phrase alignment. Encoder was measured by adequacy in (7)

$$Score_{Adequacy}(x,y) = \max\left(\mathrm{P}(\mathrm{y} = \mathrm{w}_{\text{target } i} \mid \mathrm{x} = \mathrm{w}_{\text{source}})\right) \tag{7}$$

where $x$ (in SCB-MT-EN-TH model) as source language (Thai), $y$ as the target language (English), and $\mathrm{P}(\mathrm{y} = \mathrm{w}_{\text{target } i} \mid \mathrm{x} = \mathrm{w}_{\text{source}})$ defined by (8)

$$\mathrm{P}(\mathrm{y} = \mathrm{w}_{\text{target } i} \mid \mathrm{x} = \mathrm{w}_{\text{source}}) = \frac{n(\mathrm{w}_{\text{target } i})}{n(\forall w_{target} \equiv \mathrm{w}_{\text{source}})} \tag{8}$$

The Transformer encoder consisted of input embedding (in part 2.1), positional embedding, multi-head self-attention (in part 2.2), skip connection, and layer normalization.

Positional embedding was the sine and cosine representation of token sequence (as well as wave frequency) that could be defined by (9), where $\varpi_k = \frac{1}{10{,}000^{2k/d}}$.

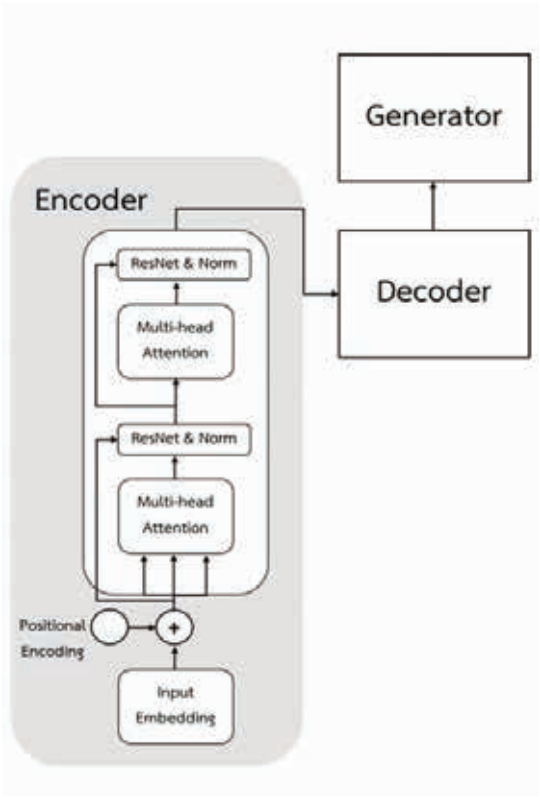


*Fig. 8: Transformer-based TH-T2I encoder.*

$$Positional\ \ Embedding(\mathrm{x} = \mathrm{w}_{\text{source}}) = \begin{bmatrix} \sin(\varpi_1 t) \\ \cos(\varpi_1 t) \\ \sin(\varpi_2 t) \\ \cos(\varpi_2 t) \\ \vdots \\ \sin(\varpi_{d/2} t) \\ \cos(\varpi_{d/2} t) \end{bmatrix}^T \tag{9}$$

Skip connection (or Res-block) was first proposed in the Residual network (ResNet) in 2015. ResNet won the ImageNet Large Scale Visual Recognition Challenge 2015 (ILSVRC 2015). As ResNet was the most popular pre-trained model in computer vision, it has become the baseline image classification model.

Skip connection coupled with element-wise addition and normalization with ReLU activation for pixels in image classification were proposed.

In Transformer, the layer normalization (LayerNorm) was used for tokens in the

Skip connection (or Res-block) was first proposed in the Residual network (ResNet) in 2015. ResNet won the ImageNet Large Scale Visual Recognition Challenge 2015 (ILSVRC 2015). As ResNet was the most popular pre-trained model in computer vision, it has become the baseline image classification model.

Skip connection coupled with element-wise addition and normalization with ReLU activation for pixels in image classification were proposed.

In Transformer, the layer normalization (LayerNorm) was used for tokens in the Seq2Seq model that could be defined in (10), where and $\sigma = \sqrt{\frac{1}{H}\sum_{i=1}^{H}(x-\mu)^2}$.

$$\hat{x} = \frac{x-\mu}{\sqrt{\sigma^2+\varepsilon}}\Theta\gamma + \beta \tag{10}$$

### *3.2. Textual transformer-based T2T decoder*

The Decoder (or language model for output of the target language as "English: EN") was translated into the language output (as well as the n-Gram model with Beam Search in SMT). Decoder was measured by fluency in (11)

$$Score_{Fluency}(y) = \max\left(\sum_{j=1}^{m}\log\left(\mathrm{P}(y_j \mid < start >, \mathrm{y}_1, \mathrm{y}_2, \mathrm{y}_3, ..., \mathrm{y}_n)\right)\right) \tag{11}$$

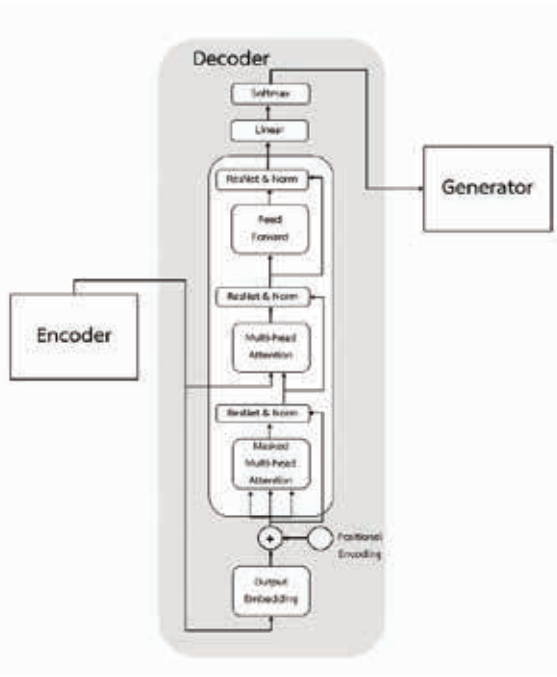


*Fig. 9: Transformer-based TH-T2I decoder.*

where $y$ (in SCB-MT-EN-TH model) as the target language (English) for output of translation and $\mathrm{P}(y_j |< start >, \mathrm{y}_1, \mathrm{y}_2, \mathrm{y}_3, ..., \mathrm{y}_n)$ defined by (12)

$$\mathrm{P}(y_j |< start >, \mathrm{y}_1, \mathrm{y}_2, \mathrm{y}_3, ..., \mathrm{y}_n) = \frac{\mathrm{n}(< start > \cap \mathrm{y}_1 \cap \mathrm{y}_2 \cap \mathrm{y}_3 \cap ... \cap \mathrm{y}_j)}{\mathrm{n}(\mathrm{y}_j)} \quad (12)$$

The Transformer decoder (Lowphansirikul, Polpanumas, Rutherford & Nutanong, 2022). It consisted of positional embedding, multi-head masked and unmasked self-attention (in part 2.2), skip connection, layer normalization, and a neural network feed-forward with linear and softmax functions.

### *3.3. Vision transformer (ViT)-based T2I generator*

The output from the previous T2T decoder (based on the SCB-MT-EN-TH model) was prompted as textual input. In TH-T2I, text-to-image generation could be done by a stable diffusion model (Lee, Hoover, Strobelt, Wang, Peng, Wright, Li, Park, Yang, Chau, 2023). The diffusion model was used to learn a data distribution by denoising a normal distribution in the UNet backbone from 2D convolution (2x2 CONV). The stable diffusion model applied the self-attention ($\tau_\theta(\bullet)$) mechanism as (1) (in part 2.1) for image generation ($z_t$) under the conditional textual language prompting ($y$), mathematically defined by (13).

$$L_{img\ gen} = \mathrm{E}_{\varepsilon(x),y,e\sim N(0,1),t}\left[\left\|\varepsilon - \varepsilon_\theta\left(z_t, t, \tau_\theta(y)\right)\right\|_2^2\right] \quad (13)$$

where = $\varepsilon_\theta\left(z_t, t, \bullet\right)$; $t$ = 1,…, $T$ to generate the suitable image ($z_t$)

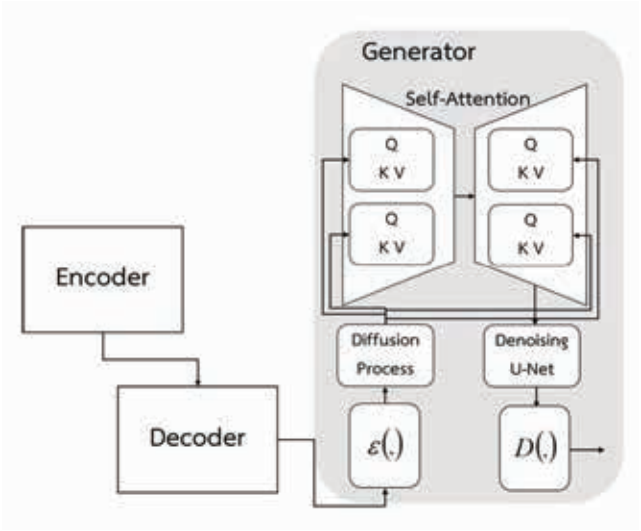


*Fig.u 10: ViT-based TH-T2I generator.*

Technically, a text representation generator encoded an input text prompt from the SCB-MT-EN-TH model's output to tokenize and weigh the primary significant tokens within a text for generating an image. Moreover, the image representation refiner was to generate an image in different scales and angles according to the dense vector representation until the final image representation.

### *4. Experimental evaluations and results*

We categorize the experimental results and evaluations into two sub-parts: evaluating quality metrics of text translation and image generation and experimental T2T and T2I results.

#### 4.*1. Evaluating metrics*

The T2T translation (as SCB-MT-EN-TH model) was evaluated by a Bi-lingual evaluation understudy (BLEU). At the same time, T2I generation (as stable diffusion) was done by Inception Score and Frechet Inception Distance (FID).

Bi-lingual evaluation understudy (BLEU) was a translation evaluation (both adequacy and fluency) by comparing machine to human translation. A higher BLEU value is better. BLEU applied precision metrics in the n-gram model (where $n$ =1, 2, 3, 4) that could be defined by (14).

$$BLEU = \min\left(1, \frac{length_{machine}}{length_{human}}\right) \bullet \left(\prod_{i=1}^{n} precision_{i-gram}\right) \quad (14)$$

Inception score (as the name from Inception v3) was used to evaluate the quality of the generated image from the stable diffusion model (A higher score is better), defined by (15)

$$Inception = \exp\left(\mathrm{E}_{y\sim z_t} D\left(P(z \mid y) \mid P\right)(z)\right) \quad (15)$$

Frechet Inception Distance (FID) measures the distance between the generated and authentic images. (The smaller FID referred to the better quality.) The FID could be computed by (16).

$$FID = \left\|\mu(z) - \mu(z_{generated})\right\|_2^2 + Tr\left(\sum_z + \sum_{z_{generated}} - 2 \times \sqrt{\sum_z \sum_{z_{generated}}}\right) \quad (16)$$

where $\mu(z)$ and $\mu(z_{generated})$ as the average of natural images and generated images, $\|\bullet\|_2^2$ as Euclidean L2 normalization, $\sum_z$ and $\sum_{z_{generated}}$ as covariance matrices of natural images and generated images, $Tr(\bullet)$ as the main diagonal of a matrix

#### *4.2 Experimental results*

Based on 100 text prompts, the BLEU evaluation on T2T translation in Table 1; and Inception and FID evaluation on T2I generation could be evaluated in Table2, respectively.

The averaged 100 TH-EN parallel corpora was evaluated by BLEU in n-Gram (where n=1,2,3,4) as shown in Table 1. We also compared to other TH-EN machine

translation, e.g., Google translate, AI for Thai. The results showed that SCB-MT-EN-TH model provided the better performance than Google translate and AI for Thai.

*Table 1: T2T translation comparison based on 100 text prompts*

| T2T metric | Google Translate | AI for Thai | SCB-MT-EN-TH |
|---|---|---|---|
| BLEU-1 | 0.57 | 0.61 | 0.67 |
| BLEU-2 | 0.45 | 0.48 | 0.54 |
| BLEU-3 | 0.39 | 0.32 | 0.41 |
| BLEU-4 | 0.22 | 0.17 | 0.28 |

In case of T2I evaluation, Inception and FID were used to compare stable diffusion to other T2I generation, e.g., diffusion probabilistic model (DPM), generative adversarial network (GAN), based on 100 text prompts in term of quality of generated images in Table 2. From the experiments, T2I by stable diffusion could generate images in the highest quality as the stable diffusion applied self-attention to weight the tokens.

*Table 2: T2I generation comparison based on 100 text prompts*

| T2I metric | DPM | GAN | Stable diffusion |
|---|---|---|---|
| Inception | 1.36 | 1.81 | 2.73 |
| FID | 236.81 | 193.64 | 124.23 |

Some examples of Thai Text-to-Image generation shown in Figure 11, the presented TH-T2I could be contributed to many local arts and cultures to preserve Thai as the digital heritage on the open-source world, e.g., Thai amulet (Mookdarsanit, 2020)., colorful Siamese fighting fishes (Mookdarsanit & Mookdarsanit, 2019a)., Thai dance gestures (Mookdarsanit & Mookdarsanit, 2018b)., Muay-Thai Folklores (Mookdarsanit & Mookdarsanit, 2018a). or nutrients and calories estimation in Thai-foods (Mookdarsanit & Mookdarsanit, 2020c). The next Thai generation of AGI researchers could inherit these resources and materials to discover new knowledge.

*Conclusion*

Unlike English, Thai was a low-resource language for NLP. Thai speech, handwriting, or comments over social media could be a fruitful material and resource for growing Thai-NLP among AGI era, as well as English. There were so many gaps in Thai-NLP for Thai AGI researchers to preserve Thai as the digital heritage on the open-source world. In this paper, we propose a novel TH-T2I as a new research area of Thai-NLP to generate an image according to Thai-text prompt. Previous Transformer-based LLM researches on Thai were only T2T. This paper firstly introduce T2I in Thai. The presented TH-T2I as planting a T2I model in digital forest for Thai preservation and could be inherited by next generation of Thai AGI researchers and students. Moreover, TH-T2I could be a T2I baseline model for any local competitions (e.g., BEST Hackathon


*Fig. 11: Some image generations based on Thai-text prompt engineering.*

organized by NECTEC). For the shortcoming, TH-T2I was such a two-stage model (divided into T2T by SCB-MT-EN-TH and T2I by stable diffusion) that could be further developed into a single stage one.

*Acknowledgement*

The paper "Thai Text-to-Image Prompt Engineering by Pre-trained Large Language with Stable Diffusion Model (TH-T2I)" was presented to integrate Thai-NLP for Thai liguistic heritage conservation as Rajabhat's mission by shifting a new paradigm of Thai-NLP research area and planting TH-T2I in digital forest. Furthermore, TH-T2I could be a Thai-NLP resources, local linguistic data and programming problems. The working resources were dedicated to Chandrakasem Rajabhat University, Bangkok, Thailand.

*References*

Arreerard, R., Mander, S. & Piao, S. (2022). Survey on Thai NLP language resources and tools. In *Proceedings of the 13th Conference on Language Resources and Evaluation* (6495-6505). ACL.

Boonkwan, P. & Supnithi, T. (2017). Bidirectional deep learning of context representation for joint word segmentation and POS tagging. In *Proceedings of the 5th In-*

*ternational Conference on Computer Science, Applied Mathematics and Applications* (184-196). Berlin, Germany: Springer.

Emsawas, T. & Kijsirikul, B. (2016). Thai Printed Character Recognition using Long Short-Term Memory and Vertical Component Shifting. In *Proceedings of 14th Pacific Rim International Conference on Artificial Intelligence* (106-115). Phuket, Thailand: Sprinker.

Haruechaiyasak, C., Kongthon, A., Palingoon, P., & Trakultaweekoon, K. (2013). S-Sense: A sentiment analysis framework for social media sensing. In *Proceedings of the 6th International Joint Conference on Natural Language Processing* (6-13). Nagoya, Japan: The Association for Computational Linguistic

Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models, *arXiv: 2006.11239.*

Inthajakra, L., Prachyapruit, A. & Chantavanich, S. (2016). The Emergence of communication intellectual history in Sukhothai and Ayutthaya kingdom of Thailand. *Social Science Asia, 2*(4), 32-41.

Ketui, N., Theeramunkong, T. & Onsuwan, C. (2013). Thai news text summarization and its application. In *Proceedings of the 2013 International Symposium on Natural Language Processing,* Phuket, Thailand : AIAT.

Klahan, A., Pannoi, S., Uewichitrapochana, P. & Wiangsripanawan, R. (2018). Thai word safe segmentation with bounding extension for data indexing in search engine. In *Proceedings of the 14th International Conference on Computing and Information Technology (83-92).* Chiang Mai, Thailand: Springer.

Koanantakool, T., Karoonboonyanan, T. & Wutiwiwatchai, C. (2009). Computers and the Thai Language. *IEEE Annals of the History of Computing, 31(1),* 46-61.

Kobchaisawat, T., Chalidabhongse, T. H. & Satoh, S. (2020). Scene text detection with polygon offsetting and border augmentation. *Electronics, 9*(1), 117.

Lapjaturapit, T., Viriyayudhakom, K. & Theeramunkong, T. (2018). Multi-Candidate word segmentation using bi-directional LSTM neural networks. In *Proceedings of the 2018 International Conference on Embedded Systems and Intelligent Technology & International Conference on Information and Communication Technology for Embedded Systems* (1-6). Khon Kaen, Thailand: IEEE

Lee, S., Hoover, B., Strobelt, H., Wang, Z. J., Peng, S. Y., Wright, A., Li, K., Park, H., Yang, H. & Chau, D. H. (2023). Diffusion explainer: visual explanation for text-to-image stable diffusion, *arXiv: 2305.03509.*

Lowphansirikul, L., Polpanumas, C., J Rutherford, A. T. & Nutanong, S. (2020). scb-mt-en-th-2020: A Large English-Thai Parallel Corpus, *arXiv: 2007.03541.*

Lowphansirikul, L., Polpanumas, C., J Rutherford, A. T. & Nutanong, S. (2022). A large English–Thai parallel corpus from the web and machine-generated text. *Language Resources and Evaluation. 56*(2), 477-499.

Lowphansirikul, L., Polpanumas, C., Jantrakulchai, N. & Nutanong, S. (2021). WangchanBERTa: Pretraining transformer-based Thai Language Models, *arXiv: 2101.09635.*

Mookdarsanit, L. & Mookdarsanit, P. (2019a). SiamFishNet: The deep investigation of Siamese fighting fishes. *International Journal of Applied Computer Technology and Information Systems, 8*(2), 40-46.

Mookdarsanit, L. & Mookdarsanit, P. (2019b). Thai herb identification with medicinal properties using convolutional neural network. *Suan Sunandha Science and Technology Journal, 6(2),* 34-40.

Mookdarsanit, L. & Mookdarsanit, P. (2020a). An adversarial perturbation technique against reCaptcha image attacks. *Journal of Science and Technology Buriram Rajabhat University, 4(1),* 33-45.

Mookdarsanit, L. & Mookdarsanit, P. (2020b). The insights in computer literacy toward HR intelligence: some associative patterns between IT subjects and job positions. *Journal of Science and Technology RMUTSB, 4(2),* 12-23 .

Mookdarsanit, L. & Mookdarsanit, P. (2021a). Combating the hate speech in Thai textual memes. *Indonesian Journal of Electrical Engineering and Computer Science, 21(3),* 1493-1502.

Mookdarsanit, L. & Mookdarsanit, P. (2021b). ThaiWritableGAN: Handwriting generation under given information. *International Journal of Computing and Digital Systems, 10(1),* 689-699.

Mookdarsanit, L. & Mookdarsanit, P. (2022). 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, 11(2),* 62-67.

Mookdarsanit, L. & Mookdarsanit, P. (2023). The cosmetic surgery recommendation: Facial acne localization and recognition. *International Journal of Applied Computer Technology and Information Systems, 12(2),* 1-6.

Mookdarsanit, L. (2020). The intelligent genuine validation beyond online Buddhist amulet market. *International Journal of Applied Computer Technology and Information Systems, 9(2),*7-11.

Mookdarsanit, P. & Mookdarsanit, L. (2018a). A content-based image retrieval of Muay-Thai folklores by salient region matching. *International Journal of Applied Computer Technology and Information Systems, 7(2),* 21-26.

Mookdarsanit, P. & Mookdarsanit, L. (2018b). An automatic image tagging of Thai dance's gestures. In *Proceedings of Joint Conference on ACTIS & NCOBA* (76-80). Ayutthaya, Thailand.

Mookdarsanit, P. & Mookdarsanit, L. (2018c). Contextual image classification towards metadata annotation of Thai-tourist attractions. *ITMSoc Transactions on Information Technology Management, 3(1),* 32-40.

Mookdarsanit, P. & Mookdarsanit, L. (2018d). Name and recipe estimation of Thai-desserts beyond image tagging. *Kasem Bundit Engineering Journal, 8(Special Issue),* 193-203.

Mookdarsanit, P. & Mookdarsanit, L. (2019). TGF-GRU: A cyber-bullying autonomous detector of lexical Thai across social media. *NKRAFA Journal of Science and*

*Technology, 15,* 50-58.

Mookdarsanit, P. & Mookdarsanit, L. (2020a). Thai-IC: Thai image captioning based on CNN-RNN architecture. *International Journal of Applied Computer Technology and Information Systems, 10(1),* 40-45.

Mookdarsanit, P. & Mookdarsanit, L. (2020b). ThaiWrittenNet: Thai handwritten script recognition using deep neural networks. *Azerbaijan Journal of High Performance Computing, 3(1),* 75-93.

Mookdarsanit, P. & Mookdarsanit, L. (2020c). The autonomous nutrient and calorie analytics from a Thai food image. *Journal of Faculty Home Economics Technology RMUTP, 2(1),* 1-12.

Mookdarsanit, P. & Mookdarsanit, L. (2021a). PhosopNet: An improved grain localization and classification by image augmentation. *TELKOMNIKA Telecommunication, Computing, Electronics and Control, 19(2),* 479-490.

Mookdarsanit, P. & Mookdarsanit, L. (2021b). The COVID-19 fake news detection in Thai social texts. *Bulletin of Electrical Engineering and Informatics, 10(2),* 988-998.

Mookdarsanit, P. & Rattanasiriwongwut, M. (2017a). GPS determination of Thai-temple arts from a aingle photo. In *Proceedings of 11th International Conference on Applied Computer Technology and Information Systems (42-47).* Bangkok, Thailand.

Mookdarsanit, P. & Rattanasiriwongwut, M. (2017b). Location estimation of a photo: a Geo-signature MapReduce workflow. *Engineering Journal, 21(3),* 295-308.

Mookdarsanit, P. & Rattanasiriwongwut, M. (2017c). MONTEAN Framework: a magnificent outstanding native-Thai and ecclesiastical art network. *International Journal of Applied Computer Technology and Information Systems, 6(2),* 17-22.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. (2021). High-resolution image synthesis with latent diffusion models, *arXiv: 2112.10752.*

Ruangrajitpakorn, T. (2006). *An example-based machine translation: a case study of translating stock reports from thai to english* [Master's thesis, Chulalongkorn University]. Graduate School, Chulalongkorn University.

Soimart, L. & Mookdarsanit, P. (2016). Gender estimation of a portrait: Asian facial-significance framework. In *Proceedings of the 6th International Conference on Sciences and Social Sciences.* Mahasarakham, Thailand.

Soimart, L. & Mookdarsanit, P. (2017a). Ingredients estimation and recommendation of Thai-foods. *SNRU Journal of Science and Technology, 9(2),* 509-520.

Soimart, L. & Mookdarsanit, P. (2017b). Name with GPS auto-tagging of Thai-tourist attractions from an image. In *Proceedings of the 2nd Technology Innovation Management and Engineering Science International Conference* (211-217). Nakhon Pathom, Thailand.

Sornlertlamvanich, V. (2019). Natural language processing research in Thai context - A 29-year journey of Thai NLP. Retrieved from: https://www.slideshare.net/virach/nlp-historythaivirach20191025

Sriwirote, P., Thapiang, J., Timtong, V. & Rutherford, A. T. (2023). PhayaThaiBERT:

Enhancing a Pretrained Thai Language Model with Unassimilated Loanwords, *arXiv: 2311.12475.*

Sutthaluang, N. & Prakancharoen, S. (2020). Prediction and protection of car driving accident in urban zone. *International Journal of Innovation, Creativity and Change, 14(8),* 308-336.

Sutthaluang, N. (2019). An open library development for pesticide residue analytics in vegetables. *International Journal of Applied Computer Technology and Information Systems, 8(2),* 31-36.

Taerungruang, S. & Aroonmanakun, W. (2018). Constructing an academic Thai plagiarism corpus for benchmarking plagiarism detection systems. *GEMA Online Journal of Language Studies, 18(3),* 186-202 .

Tapsai, C., Unger, H. & Meesad, P. (2020). The application of Thai natural language processing. *Thai Natural Language Processing, 1,* 131-159.

Theeramunkong, T., Sornlertlamvanich, V., Tanhermhong,T. & Chinnan, W. (2000). Character cluster based Thai information retrieval. In *Proceedings of the 2000 International Workshop on Information Retrieval with Asian Languages,* (75-80), Hong Kong, China : ACM.

Tirasaroj, N. (2016). *A study of word sense discrimination in Thai using latent semantic analysis* [Doctoral dissertation, Chulalongkorn University]. Graduate School, Chulalongkorn University.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). “Attention Is All You Need,” In *Proceedings of the 2017 International Conference on Neural Information Processing Systems,* (6000-6010). Long Beach, California : ACM.

***Submitted 27.09.2023***
***Accepted 16.11.2023***