This is a preprint of the following article: Siu, S.C., "Revolutionizing Translation with AI: Unravelling Neural Machine Translation and Generative Pre-trained Large Language Models" (2023).

Revolutionizing Translation with AI: Unravelling Neural Machine Translation and Generative Pre-trained Large Language Models

Sai Cheong Siu

Abstract This work explores the technical advancements in artificial intelligence (AI) for translation, focusing on neural machine translation (NMT) and large language models (LLMs) as primary drivers of progress. We analyze foundational deep learning techniques, the influential Transformer model, and the strengths and weaknesses of NMT and LLMs for translation. Moreover, we offer practical recommendations for translation educators and developers on how to effectively harness AI's power while addressing potential challenges. Our goal is to enhance understanding of AI-driven translation technologies and equip language professionals with the knowledge needed to make informed decisions when using these cutting-edge tools.

Keywords neural machine translation, large language models, translation technology, artificial intelligence, deep learning, ChatGPT

1 Introduction

Artificial intelligence (AI) has advanced rapidly in recent years, enabling numerous applications transforming various fields (e.g., LeCun et al. 2015; Zhang et al. 2021). Translation technology has also made significant progress, catalyzed by innovative developments such as neural machine translation (NMT) (e.g., Bahdanau et al. 2014; Vaswani et al. 2017) and pre-trained large language models (LLMs) (e.g., Devlin et al. 2018; Brown et al. 2020). These AI-driven techniques have spurred a new wave of tools applicable to translation and language services, including conversational chatbots capable of real-time language translation and communication (e.g., ChatGPT (OpenAI 2022)). In this work, we delve into the latest AI developments in translation technology. We focus specifically on artificial neural networks that underpin neural machine translation and the generative language models employed in the development of advanced chatbot applications. This discussion aims to enhance our understanding of how AI facilitates translation and assists language professionals in determining the most effective ways to leverage cutting-edge, AI-powered applications.

The present work is structured as follows: Section 2 provides an overview of deep learning, an AI approach featuring the design and use of multilayer artificial neural networks. Section 3

S. C. Siu
School of Translation and Foreign Languages
The Hang Seng University of Hong Kong, Hong Kong, China
e-mail: scsiu@hsu.edu.hk

explains the Transformer model, an architecture that has enabled significant advances in natural language processing (NLP). Section 4 explores two types of AI for translation: NMT, which is designed for translation between languages, and generative LLMs, which generate human-like text and power chatbots applicable to translation. Section 5 analyzes the strengths and weaknesses of these applications, highlighting their potential and limitations. Sections 6 and 7 offer recommendations for translation educators and developers to train language professionals with necessary skills and improve AI tools. Section 8 presents concluding remarks. It is hoped that this work provides insights for language professionals to make informed decisions about incorporating these technologies.

2 An Overview of Deep Learning

Deep learning is a branch of machine learning that uses multilayer artificial neural networks to analyze and process data, with applications in various domains like NLP and computer vision (LeCun et al. 2015). These networks are loosely inspired by biological neural networks and consist of simple units called artificial neurons. The objective is to learn how to extract relevant features from large amounts of data for a specific task, without needing to handcraft rules like in symbolic AI or feature engineering as in other machine learning methods. In the context of NLP, deep learning typically involves four key tasks, as explained in Sections 2.1-2.4 below. For more information, refer to Goodfellow et al. (2016), Zhang et al. (2021), and Siu (2023b).

2.1 Model Design

Model design refers to the creation of a model that maps input data to output data based on specific parameters. In language translation, for example, the input and output are sequences of words or tokens. Token sequences are often represented as high-dimensional vectors using techniques like one-hot encoding or word embeddings, such as Word2Vec or GloVe (e.g., Mikolov et al. 2013; Pennington, Socher, and Manning 2014; Jurafsky and Martin 2023). The number of parameters in a model can range from a few hundred million (Radford et al. 2019) to over a hundred billion for large models like GPT-3 (Brown et al. 2020).

2.2 Data Collection

Data collection involves gathering a dataset for training the model. In machine translation, this usually means assembling large datasets of input-output pairs to teach the model how to generate correct translations. These datasets often consist of parallel corpora containing aligned sentences in the source and target languages. In addition to bilingual data, monolingual data may also be utilized, which can be back-translated to form bilingual sentence pairs or used directly for training language or translation models (e.g., Sennrich, Haddow, and Birch 2016; Edunov et al. 2018).

2.3 Model Training

Model training involves finding the best values of parameters that minimize the error. During training, the dataset is divided into smaller batches for efficient computation. Each batch contains a fixed number of input-output pairs, which are processed through the model.

There are two essential processes during training: forward propagation and backpropagation. In the forward propagation, the input is passed through the model, generating an output prediction. The loss function then calculates the error between the predicted output and the true output for each input-output pair in the batch. In NLP and language translation, the cross-entropy loss can be used as the loss function (for details, see Jurafsky and Martin (2023)). It measures the difference between the predicted probability distribution of tokens in the output sequence and the true distribution.

Backpropagation is an algorithm used to compute the gradients of the loss function with respect to the model's parameters. It works by applying the chain rule of calculus recursively, starting from the output layer and going back through the model's layers. The gradients provide information about how the model's parameters should be adjusted to minimize the loss function. Optimization algorithms are used to minimize the cross-entropy loss during model training. These algorithms update the parameters based on the gradients of the average loss with respect to the parameters. Examples of optimization algorithms include stochastic gradient descent (SGD) with random initialization and momentum (Sutskever et al. 2013) and Adam (Kingma and Ba 2014).

2.4 Model Evaluation

To assess the performance of a trained model, a separate test dataset that was not used during the training process is utilized. This approach helps to estimate the model's ability to generalize to previously unseen data and detect potential overfitting. Examples of evaluation metrics for machine translation include BLEU (Papineni et al. 2002) and CHRF (Popović 2015), which quantify the similarity between the model-generated output sequences and their corresponding reference sequences. There are also embedding-based metrics that calculate the similarity between the output and reference sentence embeddings. Notable examples of these metrics are BERTScore (Zhang et al. 2019), BLEURT (Sellam, Das, and Parikh 2020), and COMET (Rei et al. 2020). For further details on evaluation metrics for machine translation, refer to Siu (2022).

3 The Transformer Model Explained

The Transformer model, introduced by Vaswani et al. (2017), has become a widely used architecture in the field of NLP (Tay et al. 2020). Given the model's importance in NMT and generative LLMs, this section explains its key features.

3.1 Input and Output of the Transformer Model

The Transformer model processes a sequence of tokens, such as words, subwords, or characters, based on the chosen tokenization strategy (Kudo and Richardson, 2018; Tay et al., 2021). Each token is represented by a high-dimensional embedding vector.

The model's output depends on the specific task and implementation. Generally, it produces a sequence of continuous vectors that represent the learned information for each position. These output vectors can be used as features for other models or fed into subsequent layers within the Transformer model. For sequence-to-sequence tasks, the model generates tokens one at a time in an auto-regressive manner, decoding them into probabilities for all possible tokens, normalized using the softmax function. The model predicts the next token in the output sequence until it generates an end-of-sequence token or reaches the maximum output sequence length.

3.2 Key Components of the Transformer Model

The Transformer model comprises several essential components: encoder and decoder stacks, positional encodings, multi-head self-attention, and position-wise feed-forward networks with layer normalization and residual connections.

Encoder and Decoder Stacks The original Transformer model features an encoder-decoder architecture. The encoder maps input tokens to hidden states, while the decoder generates the output sequence based on these hidden states. Both encoder and decoder stacks consist of identical layers, with the encoder containing a multi-head self-attention mechanism followed by a position-wise feed-forward network. The decoder stack also has identical layers but includes an additional multi-head attention component attending to the encoder stack's output, making the model suitable for sequence-to-sequence tasks. Decoder-only transformers, as seen in some LLMs, omit this extra attention component, focusing solely on the decoder's input.

Positional Encodings The Transformer model processes input tokens in parallel, requiring explicit information about each token's position within a sequence. Positional encodings provide this information by adding position-specific information to token embeddings. A common method for generating positional encodings uses sinusoidal functions.

Multi-Head Self-Attention Mechanism This is central to the Transformer model, capturing relationships between tokens in an input sequence by calculating multiple attention scores for each token pair using distinct learned projections. These scores are used to calculate a weighted sum of token embeddings, allowing the model to consider various aspects of the input simultaneously.

Position-wise Feed-Forward Networks (FFNs) Position-wise FFNs process the output matrix of the multi-head self-attention mechanism. These FFNs consist of two linear layers with a ReLU (Rectified Linear Unit) activation function in between.

Layer Normalization and Residual Connections Layer normalization is applied to the input before summing with the output, creating a residual connection. This takes place after the multi-head self-attention mechanism and the position-wise FFN. For more information, refer to Xiong et al. (2020).

3.3 Variants of the Transformer Model

Over time, the Transformer model has been adapted for various tasks and domains, resulting in popular variants and applications like BERT (Devlin et al. 2018), GPT (Radford et al. 2018; 2019), T5 (Raffel et al. 2020), and Vision Transformers (Dosovitskiy et al. 2020; Khan et al. 2022). BERT has revolutionized natural language understanding by leveraging bidirectional context in pre-training and fine-tuning for a wide range of tasks. GPT has become known for its strong generative capabilities, particularly in text generation and completion tasks. T5 has unified various natural language processing tasks under a single text-to-text framework for both pre-training and fine-tuning. Vision Transformers have applied the Transformer architecture to computer vision tasks, showcasing the model's versatility beyond natural language processing. These variants have significantly advanced artificial intelligence across multiple domains and continue to inspire new developments in the field.

4 AI for Translation: Neural Machine Translation and Generative Pre-trained Large Language Models

In this section, we discuss two prominent examples of AI for translation: (1) neural machine translation; and (2) generative pre-trained large language models. Both leverage deep learning techniques discussed in previous sections, with the Transformer being a popular architecture.

4.1 Neural Machine Translation (NMT)

NMT often uses an encoder-decoder architecture, which can consist of Transformer or alternative networks like recurrent neural networks (RNNs) and convolutional neural networks (CNNs) (Stahlberg 2020; Siu 2023b). The encoder maps the source text to an internal representation, which the decoder then transforms to predict the next token in the target text. Mathematically, we can define the probability distribution of the next target token y_t as follows:

$$P(y_t | \boldsymbol{y}_{<t}; \boldsymbol{x}; \theta) = F(y_t, \boldsymbol{y}_{<t}, \boldsymbol{x}; \theta).$$

Here, F represents the NMT model (e.g., a Transformer) that maps the input tokens $\boldsymbol{x}$, previous target tokens $\boldsymbol{y}_{<t}$, and the next token y_t to logit values, which are then transformed into probability values using the softmax function. The model parameters are denoted by θ.

4.1.1 The Transformer Model for NMT

As discussed in Section 3, the Transformer model is a popular architecture for NLP. It is designed to overcome the limitations of RNNs and CNNs in handling long-range dependencies and parallelization (Vaswani et al. 2017).

For NMT, the Transformer encoder comprises multiple encoding layers, with each layer consisting of a multi-head self-attention mechanism and position-wise feed-forward networks with layer normalization and residual connections. The first layer accepts positional encoding, which provides information about the relative positions of tokens in the input sequence. The multi-head self-attention mechanism enables the model to determine the importance of different tokens in the input sequence by capturing dependencies among words. The position-wise feed-forward networks apply a non-linear transformation to each token independently. The output is an internal representation of the source text used by the decoder.

The decoder consists of multiple layers that resemble the encoder layers but with two key differences. First, it employs a self-attention mechanism that masks future target tokens during training, preventing the decoder from attending to them. Second, it has an additional self-attention component that combines the encoder's output representation of the source text. More specifically, the decoder initially uses a self-attention mechanism like the encoder's but attends only to target tokens predicted so far while masking future tokens, and the second self-attention mechanism attends to the encoder output. The output is a sequence of logit values for the next token in the target text, which are then converted into probabilities using the softmax function.

4.1.2 Decoding Methods

Decoding methods in NMT models, such as those used in Bahdanau, Cho and Bengio (2014) and Vaswani et al. (2017), predict the next word in a sequence based on the probabilities of subsequent tokens. Various decoding methods help strike a balance between translation quality and diversity, enhancing the performance of these models.

A simple method is greedy search, which selects the token with the highest probability at each step. However, it may generate sub-optimal translations due to early errors that accumulate over time. An alternative to greedy search is beam search, which maintains multiple candidate translations at each step. Beam search keeps *n* partial candidate sequences, called beams, where *n* is the beam width. The beams are expanded by selecting the top-*n* most probable tokens at each step. Beam search continues until the end-of-sequence token is reached or a maximum sequence length is achieved. The beam with the highest final probability is chosen as the output translation. By exploring more candidates, beam search (using narrow beams) produces better translations compared to greedy search (Koehn and Knowles 2017). However, its computational cost also increases with the beam width.

We may also consider sampling to introduce randomness and explore more diverse translations. Sampling with temperature scaling, for example, selects tokens based on their probability distribution (Hinton, Vinyals, and Dean 2015). This method involves a temperature parameter that adjusts the output probability distribution. Another decoding approach is top-*k* sampling (Fan, Lewis, and Dauphin 2018), which selects a token from the top-k most probable tokens at each step, where *k* is an integer representing the number of tokens to consider. Alternatively, top-*p* sampling, also known as nucleus sampling (Holtzman et al. 2019), chooses tokens from the top-*p* most probable tokens, where *p* represents a cumulative probability threshold. This method first sorts the tokens by their probabilities and then selects the smallest set of tokens whose cumulative probability surpasses *p*. Finally, it samples a token from the chosen set.

4.2 Generative Pre-trained Large Language Models (LLMs)

Generative pre-trained LLMs have shown remarkable versatility across a wide range of NLP tasks beyond translation. LLMs feature the use of a massive number of parameters and can be classified as encoder-based, decoder-based, or encoder-decoder-based. For example, GPT-3, a decoder-based LLM, has 96 layers with 128 attention heads and an input token dimension exceeding 12,288 (Brown et al., 2020), whereas the "big" Transformer model for NMT, which is an encoder-decoder-based model, comprises 6 encoding layers and 6 decoding layers, with 8 attention heads and an input token dimension of 1024 (Vaswani et al., 2017). Formally, for decoder-based language models like ChatGPT, the probability distribution over the next token given the previous tokens can be denoted as follows:

$$P(y_t|\boldsymbol{y}_{<t};\theta) = \mathrm{softmax}(F(y_t, \boldsymbol{y}_{<t};\theta)),$$

where F is the decoder-based model, and θ denotes the model parameters. In this equation, $\boldsymbol{y}_{<t}$ represents the sequence of tokens generated so far by the decoder, which takes these previous tokens as input to predict the next token y_t in the target sequence. Decoding methods for LLMs share similarities with those for NMT models.

LLMs leverage transfer learning through pre-training, fine-tuning, zero-shot, and few-shot learning (e.g., Radford et al. 2018, 2019; and Brown et al. 2020). They are first pre-trained on a vast corpus of text in an unsupervised manner to "learn" language in general, capturing syntactic, semantic, and contextual information embedded in the training data. Then, the models can be fine-tuned for specific tasks using smaller, task-specific datasets. Unlike NMT, which is task-specific and often trained end-to-end on parallel corpora, LLMs learn translation indirectly. Sophisticated models can effectively tackle new tasks without having observed any examples during fine-tuning (zero-shot learning) or with only a limited number of examples (few-shot learning).

GPT-3, for instance, as a decoder-based transformer model, is similar to the decoder for machine translation presented in the Transformer paper, with two differences: (1) it lacks the second self-attention mechanism (i.e., the one taking encoder output as input), and (2) it employs sparse attention in Transformer layers (Brown et al., 2020). For ChatGPT, pre-training is followed by instruction fine-tuning and reinforcement learning with human feedback based on GPT-3 model (see OpenAI (2022) and Ouyang et al. (2022)), enabling the model to follow human instructions and generate more accurate and contextually relevant responses.

When using LLMs as translation tools, the input is a prompt that includes an instruction specifying the task and other relevant information, along with the source text to be translated. The model then predicts the next token iteratively, in a manner similar to NMT, computing probabilities and decoding using the aforementioned methods. Effective prompt engineering plays a crucial role in achieving high-quality translations, as LLMs rely on prompts to understand the context and specific task. The prompt should clearly specify the desired output format and include any additional information that might assist the model in generating an accurate translation. For example, to translate English text into Chinese, the prompt can be framed as “Translate the following text from English into Chinese: ‘<source_text>’.” See Gao et al. (2023), Peng et al. (2023) and Siu (2023a) for more discussion on prompting for translation.

Although both NMT and LLMs can be employed for translation, they exhibit different features. First, LLMs have been pre-trained on a diverse range of text, making them potentially more adaptable to different domains and styles, while NMT models may require additional fine-tuning on domain-specific parallel corpora for optimal performance in specialized domains (see Koehn & Knowles (2017) for more discussion on this “domain mismatch” issue). Second, LLMs can perform zero-shot or few-shot learning, making them more data-efficient (when they need to be developed for translation) compared to NMT models, which typically require large parallel corpora for training. Third, LLMs are capable of handling multiple tasks without the need for dedicated models for each task, unlike NMT models, which are specifically designed for translation and require separate models for other language tasks. Lastly, while NMT models are optimized specifically for translation tasks, LLMs, though capable of generating high-quality translations in some cases, may not always match the performance of dedicated NMT models (e.g., Brown et al. 2020).

5 Strengths and Limitations of NMT and LLMs

In this section, we explore the strengths and weaknesses of NMT and generative pre-trained LLMs in the context of translation tasks.

5.1 Strengths

5.1.1 Superior Performance over Conventional Machine Translation Approaches

NMT systems have demonstrated improvements over their predecessors, such as Statistical Machine Translation (SMT), when sufficient training data is available (Stahlberg 2020). One reason for NMT’s success lies in its ability to capture complex linguistic relationships. Unlike SMT, which relies heavily on word alignment and phrase-based methods, NMT learns directly from parallel data to understand context and semantics. The use of deep neural networks and attention mechanisms further equips NMT systems with the ability to handle long-range dependencies and ambiguities in language.

5.1.2 Enhanced Context Awareness through Transformer Models

The adoption of Transformer models facilitates improved context awareness by utilizing self-attention mechanisms, as discussed in Sections 3 and 4. This capability allows NMT systems and LLMs to weigh the importance of words in a sentence based on their relevance to the output text. Each token in the input sequence has a dynamic internal representation that changes according to the context and other tokens present, enabling the model to adapt to different linguistic situations more effectively. Consequently, the model can dynamically adjust to various contexts, resulting in higher quality translations. This is particularly beneficial in cases where context plays a crucial role, such as preserving the meaning of idiomatic expressions and handling pronoun disambiguation.

5.1.3 Impressive Performance and Customizability of LLMs

LLMs, such as ChatGPT, demonstrate promising performance in translation tasks and could potentially surpass NMT systems (Jiao et al. 2023). One of the key strengths of LLMs lies in their customizability, as users can provide more information in the form of prompts to supply additional details about a translation job (Siu 2023a). This feature enables tailored translations that accurately capture the intended meaning and adhere to specific preferences in terms of style or tone. Moreover, LLMs can be fine-tuned further on domain-specific data, which may result in enhanced translation quality and adaptability across various subject areas.

5.1.4 Versatility of LLMs as Advanced Assistive Tools for Translators

LLMs can handle a wide range of NLP tasks beyond producing translations, making them valuable tools for translators. As discussed in Siu (2023a), they can function as interactive editors, suggesting possible edits and modifications to translators' drafts, and facilitate human-in-the-loop collaboration. This approach allows translators to leverage the language fluency and grammatical knowledge embedded in large LLMs while maintaining full control over the final translation. LLMs can also generate rough draft translations from scratch, provide explanations for expressions in the source and/or target language, and offer multiple alternative translations for consideration. These features contribute to a more efficient and effective translation process, allowing translators to focus on complex, nuanced translations and delegate repetitive tasks to their AI collaborators. This human-computer partnership paves the way for a mutually enriching translation experience and showcases the potential of LLMs as advanced assistive tools in the translation industry.

5.2 Limitations

While NMT and LLMs have achieved remarkable success in translation, they also present some shortcomings that necessitate further research and development. In this section, we discuss five major areas, including the processing of long sentences, support for specialized translation domains, support for low-resource languages, contextual and background analysis, and hallucination.

5.2.1 Translation of Long Input Sentences

NMT systems may struggle to produce complete and accurate translations when presented with very long sentences (Koehn and Knowles 2017). The following example illustrates how two NMT systems fail to fully translate a long sentence in Chinese recovery (Hong Kong and Macao Affairs Office of the State Council 2023), comprised of over 110 characters, concerning the Central Government's support for Hong Kong's post-pandemic.

- **Source text** 中央始終十分關心香港的疫情形勢，多措並舉支持香港統籌做好疫情防控和經濟社會發展工作，中央有關部門高度重視在疫情常態化防控下恢復內地與香港『通關』的問題，一直與特區政府保持密切溝通，深入交換意見，推動兩地防控措施的有效銜接。

- **Reference** The Central Government has consistently paid close attention to the epidemic situation in Hong Kong and has implemented a variety of measures to support Hong Kong in effectively coordinating its epidemic prevention and control efforts, as well as its economic and social development. Relevant central government departments place great importance on the "resumption of normal cross-border travel" between the Mainland and Hong Kong under normalized epidemic prevention and control measures. They have been maintaining close communication and in-depth exchanges of views with the SAR government to promote effective coordination of prevention and control measures between the two regions.

- **Google Translate** The central government has always been very concerned about the epidemic situation in Hong Kong, and has taken multiple measures to support Hong Kong in coordinating the work of epidemic prevention and control and economic and social development. The SAR government maintains close communication, exchanges views in depth, and promotes the effective convergence of prevention and control measures between the two places.

Although the translation appears fluent at first glance, a closer examination reveals that the middle part of the source sentence is omitted (i.e. "中央有關部門高度重視在疫情常態化防控下恢復內地與香港『通關』的問題"). For NMT in general, problems relating to the translation of long sentences could result from several factors. The training data used to train NMT models may have predominantly contained short sentences, causing the models to struggle with properly translating longer sequences; more data with long, complex sentences are needed. In addition, the encoder and decoder of NMT models might have insufficient capacity (e.g., the number of layers or dimensionality of embeddings) to fully capture the information in the long input sentence, so more advanced model architectures with increased capacity can be explored. The models may also have a hard constraint on the maximum length of input/output sequences, leading them to generate partial translations for sentences that exceed the maximum length. Finally, the attention components of the models that align tokens may have difficulty maintaining alignment over long distances, and developing more sophisticated attention mechanisms that can capture long-range dependencies could lead to better translation results.

5.2.2 Support for Specialized Translation

Neural machine translation engines, especially general-purpose ones, may encounter difficulties when translating specialized documents due to inconsistencies in translation style and the accuracy of domain-specific terminology. Consequently, there has been extensive research on domain adaptation – the process of adapting translation systems to specialized domains (see, for instance, Chu and Wang (2018)). Similar challenges have been observed when using LLMs, such as ChatGPT, for translation tasks (Siu 2023a).

The following short example illustrates stylistic and terminological issues in translating a sentence from a listing document in Hong Kong. The source sentence in English is from an "Important Notice to Investors" (Landrich Holding Limited 2020), advising investors to rely only on information in the prospectus when making investment decisions.

- **Source text** You should rely only on the information contained in this prospectus to make your investment decision.

- **Reference** 閣下應僅依賴本招股章程所載資料作出投資決定。

- **Google Translate** 您應該僅依賴本招股說明書中包含的信息來做出投資決定。

- **ChatGPT** 您應該只依賴於本招股說明書所包含的資訊，以做出您的投資決定。

In both Google Translate and ChatGPT, the pronoun "you" is rendered as "您", which is inconsistent with the use of "閣下" in Hong Kong IPO prospectuses. The word "prospectus" is translated as "招股說明書", more common in Mainland China than "招股章程" used in Hong Kong prospectuses. Similarly, in Google Translate, the word "information" is translated as "信息" rather than "資料" in the official translation, with the former more typical in Mainland China and the latter in Hong Kong. These issues in translation style and terminology may stem from the use of Mainland Chinese training data and different usage patterns in general and specialized documents. This could be addressed by training models on specialized texts. For example, we developed IPOTranslate (Siu 2019), a neural machine translation system trained on sentences from IPO prospectuses. Exposure to the writing style and terminology of these documents allows IPOTranslate to translate the above sentence consistently with the official translation. This suggests that training on in-domain data and tailoring models to domain-specific language usage and conventions can help address these issues in specialized translation.

5.2.2 Limited Support for Non-English Languages

LLM-driven applications, depending on the composition of the languages in the training data, tend to perform better in English if they are trained on data primarily in English, with limited support for other languages such as Chinese. The following example from Siu (2023a), which discussed ChatGPT's limitations in generating Chinese sentences, illustrates its lack of support for Chinese, leading to suboptimal English-Chinese translation. This is given that ChatGPT was largely trained on English data despite demonstrating an ability to handle multilingual data. The example is a short sentence describing Hong Kong's features and population statistics in 2019:

- **Source text** Translate the following into Traditional Chinese: Hong Kong, with a population of about 7.51 million in mid-2019, is a small but dynamic city which has earned an international reputation as a leading commercial and financial centre as well as a highly efficient entrepot.

- **Reference** 在2019年年中，香港人口約751萬。香港雖然面積小，但朝氣 蓬勃、幹勁十足。香港被譽為全球首屈一指的商業及金融中心，亦是一個效率超卓的轉口港。

- **ChatGPT** 香港是一座人口约在2019年中期达到750万的小而有活力的城市，不仅因为是一个高效的贸易港口而享有国际声誉，同时也是一个领先的商业和金融中心。

The results indicate that ChatGPT fails to follow the instructions for producing a translation in Traditional Chinese, instead generating a translation in simplified Chinese. The Chinese translation "在2019年中期达到750万的小而有活力的城市" is also largely unclear, inaccurately translating the number "7.51 million" as "750万" (7.5 million). In addition to inadequate support for Chinese, the literature suggests that there is potential for ChatGPT to enhance its non-English translations (refer to Khoshafah 2023, for instance), and increasing the proportion of non-English data might offer a solution (e.g., Anil et al. 2023).

5.2.4 In-depth Analysis of Context and Background

Although NMT and LLMs enable translation systems to consider neighboring words within the input, translating texts that require a nuanced understanding of context beyond the input itself remains a challenge. Tasks that depend on paratextual context, background knowledge, or historical context strain current AI technologies, particularly if the necessary contextual information is not present in the training data.

The following example illustrates the significant room for improvement in translation tasks that require an in-depth analysis of context and background. The source text (HKSAR Government 2023) is a speech given by Mr. John Lee, Chief Executive of Hong Kong, a few days before Chinese New Year 2023. The speech announces the resumption of cross-border travel between mainland China and Hong Kong, following nearly two years of COVID-19 border closures.

- **Source text** 今日是年廿七，再過幾日就是大年初一，我們將會送虎迎兔。我先在這裏預祝大家新一年事事如意、身體健康。新一年香港的第一件大事必然是與內地逐步、有序、全面通關。由一月八日實施首階段通關到昨日，剛好十天，已經有差不多40萬人次經陸路口岸免檢疫往返兩地，由今日到年三十晚，更會是高峰期。見到兩地人民家庭團聚，在春節前達成重聚的心願，笑容滿面，團團圓圓，溫馨感人，我非常欣慰。

- **Reference** Today marks the 27th day of the 12th lunar month, with the Chinese New Year just a few days away. We will soon bid farewell to the Year of the Tiger and welcome the Year of the Rabbit. I would like to take this opportunity to extend my warmest wishes to everyone for a prosperous and healthy New Year in advance. The first significant event for Hong Kong in the New Year is the progressive, orderly, and comprehensive resumption of normal travel with Mainland China. Since the implementation of the first phase of travel resumption on January 8 up until yesterday, a span of ten days, approximately 400,000 passengers have traveled quarantine-free between Hong Kong and Mainland China through boundary control points. The peak travel period will continue from today until Chinese New Year's Eve. Witnessing the reunions of cross-boundary families as they fulfill their long-held desire to gather before the Spring Festival, I am deeply touched and filled with immense joy, seeing their beaming smiles and heart-warming celebrations.

- **ChatGPT** Today is the 27th day of the year. In a few days, it will be the first day of the Lunar New Year, and we will be sending off the Year of the Tiger and welcoming the Year of the Rabbit. I would like to take this opportunity to wish everyone a happy and healthy new year.
  The first major event in Hong Kong in the new year will undoubtedly be the gradual, orderly, and comprehensive clearance with mainland China. From the implementation of the first phase of clearance on January 8th to yesterday, which was exactly ten days, almost 400,000 people have crossed the border without quarantine checks by land. From today until New Year's Eve, it will be the peak period. It is heartwarming to see families from both sides reunite before the Spring Festival, with smiles on their faces and warmth in their hearts, and I am very pleased about this.

The ChatGPT translation presents two primary issues. Firstly, the model inaccurately translates the expression "年廿七" as "the 27th day of the year." This error is likely due to insufficient exposure to the Chinese cultural context, which leads to a flawed interpretation of expressions related to dates preceding the Chinese New Year. Secondly, the model translates "通關" as "clearance," failing to fully convey the intended meaning within the given context. Clues such as "quarantine inspection" and "Spring Festival" suggest that the term refers to the reopening of borders after the COVID-19 pandemic, rather than clearance in a general sense. However, the model lacks the contextual understanding to infer the correct translation. From this example, we can see that while neural models have enabled translation systems to consider local context, adequately handling background knowledge and context remains challenging. The translation of complex texts requires in-depth analyses of textual and para-textual context and proper integration of external knowledge, and given this, for now, human input still plays a critical role in the most nuanced, specialized translation tasks.

5.2.5 Hallucination

Hallucination, or generating fictional details, is a well-known shortcoming of large language models, as discussed in the literature, including reports by OpenAI (2023) and Bubeck et al. (2023). This issue can significantly impact various use cases, such as the recent incident where a lawyer citing fabricated cases generated by ChatGPT, which falsely claims to have verified that they real, as reported by The New York Times (Weiser 2023). Hallucination could be a key disadvantage of large language models compared to neural machine translation systems, which are developed specifically for translation and trained on bilingual data pairs, thus tending to follow the source text more closely and exhibiting less hallucination.

The problem of hallucination may become more apparent when a large language model processes input beyond its training data. In these cases, the model may generate fictional details in the target text based on outdated knowledge acquired during pre-training, rather than faithfully translating the source text. For example, consider the following example from Siu (2023a). The source text was published by the HKSAR Government on March 15, 2023, describing a visit by Mr. John Lee, the Chief Executive, to Beijing, accompanied by major government officials:

- **Source Text** 行政長官李家超今日（三月十五日）繼續在北京的訪問行程。他先後拜訪國家科學技術部、文化和旅遊部、海關總署和商務部，與各部門領導就共同關注的事宜交換意見。文化體育及旅遊局局長楊潤雄、政制及內地事務局局長曾國衞、保安局局長鄧炳強、商務及經濟發展局局長丘應樺、醫務衞生局局長盧寵茂教授和創新科技及工業局局長孫東教授亦有出席相關會面。

- **Reference** The Chief Executive, Mr John Lee, today (March 15) continued his visit in Beijing, where he called on the Ministry of Science and Technology, the Ministry of Culture and Tourism, the General Administration of Customs of the People's Republic of China (GACC), and the Ministry of Commerce to exchange views with the respective leaders on issues of common interest. The Secretary for Culture, Sports and Tourism, Mr Kevin Yeung; the Secretary for Constitutional and Mainland Affairs, Mr Erick Tsang Kwok-wai; the Secretary for Security, Mr Tang Ping-keung; the Secretary for Commerce and Economic Development, Mr Algernon Yau; the Secretary for Health, Professor Lo Chung-mau; and the Secretary for Innovation, Technology and Industry, Professor Sun Dong, also joined relevant meetings.

The translation from ChatGPT contains fictional details not present in the source text:

- **ChatGPT** On March 15th, the Chief Executive, Mr. Carrie Lam, continued her visit to Beijing. She visited the Ministry of Science and Technology, the Ministry of Culture and Tourism, the General Administration of Customs, and the Ministry of Commerce, and exchanged views with leaders of these departments on matters of common concern. The Director of the Tourism Bureau, Mr. Yang Runxiong, the Director of the Constitutional and Mainland Affairs Bureau, Mr. Patrick Nip, the Director of the Security Bureau, Mr. John Lee Ka-chiu, the Director of the Commerce and Economic Development Bureau, Mr. Edward Yau Tang-wah, the Director of Health, Professor Sophia Chan Siu-chee, and the Director of the Innovation and Technology Bureau, Professor Alfred Sit Wing-hang, also attended the relevant meetings.

The platform mistakenly insists that the Chief Executive mentioned in the text is Mrs. Carrie Lam, who was in office at the time of the 2021 training data cut-off, before the new government took office in July 2022 and the current Chief Executive, Mr. John Lee, assumed duties. The tool struggles with the conflict between the presence of the male third-person pronoun “他” in the source text and “Carrie Lam,” who is female, learned from the training data, leading to the interesting translation of “Mr. Carrie Lam” and inconsistent usage of personal pronouns in the article (e.g., using “She” to refer to “Mr. Carrie Lam”). The job titles and names of government officials in the translation are largely incorrect; instead of translating according to the source text, the system attempts to translate the names of government officials into the responsible officials of bureaus with a similar nature in 2021. Notably, Mr. John Lee, the current Chief Executive, appears as “the Director of the Security Bureau” (the correct translation of the post should be “Secretary for Security”), which he held from 2017 to June 2021, with the post currently held by Chris Tang Ping-keung.

In contrast, the output of Google Translate adheres more closely to the factual details provided in the source text, with some inaccuracies but less hallucination:

- **Google Translate** The Chief Executive, Mr Lee Kar-chao, continued his visit to Beijing today (March 15). He successively visited the Ministry of Science and Technology, the

> Ministry of Culture and Tourism, the General Administration of Customs and the Ministry of Commerce, and exchanged views with leaders of various departments on issues of common concern. Yang Runxiong, Secretary for Culture, Sports and Tourism, Tsang Kwok-wai, Secretary for Constitutional and Mainland Affairs, Tang Ping-keung, Secretary for Security, Yau Ying-wah, Secretary for Commerce and Economic Development, Professor Lo Chung-mao, Secretary for Medical and Health Bureau, and Professor Sun Dong, Secretary for Innovation, Technology and Industry Attend relevant meetings.

The output is generally based on the source text, without attempting to compare the current officials' names with those prior to 2022. However, there are still discrepancies that need to be revised in the post-editing stage, including the correct names of the officials (such as "Mr. Lee Kar-chao" and "Yang Runxiong," the latter using pinyin transliteration for Mr. Kevin Yeung), along with the mistranslation of the post "Secretary for Health," which came into effect in July 2022, becoming "Secretary for Medical and Health Bureau" in the target text.

6 Recommendations for Translation Educators

As AI-driven translation tools continue to advance rapidly, it is crucial for translation practitioners to adapt their practices to maximize the benefits of these technologies while mitigating their limitations. When leveraged properly with a deep understanding of their capabilities and shortcomings, AI can significantly enhance human-AI collaboration and cooperation in translation. This section provides recommendations for translation educators to help prepare language professionals to work effectively with cutting-edge technologies. For further discussion, please refer to Siu (2023a).

6.1 Focus on specialized translation domains

With AI systems becoming increasingly adept at handling standard, generic translation tasks, human expertise may prove more valuable in specialized domains such as legal, medical, and technical translation. It would be beneficial for educators to develop curricula and design teaching and learning activities concentrating on these areas of specialization, helping students develop expertise in them.

6.2 Provide in-depth exploration of AI-based translation technologies

Students can benefit from gaining a deeper understanding of the fundamental principles behind AI technologies that power translation, moving beyond conventional applications like translation memories and termbases. By learning how neural networks, attention mechanisms and transfer learning work, students can gain valuable insights into the inner workings of AI translation tools. Teaching students how to leverage NMT models through effective prompting and temperature tuning methods, as well as encouraging experimentation to optimize AI-based translation tools, can further enhance their skills in this area.

6.3 Enhance programming and computational skills

By equipping students with basic programming and computational skills, they can customize, enhance, and even develop their own AI solutions. For example, we can introduce students to programming languages such as Python and popular machine learning libraries and frameworks. It is also beneficial to familiarize students with data preprocessing and automatic evaluation, helping them appreciate the important role of quality data and evaluation in the development of AI-driven translation applications. The enhancement of programming education enables students to better understand the limitations of existing tools and build new solutions tailored to their needs.

6.4 Focus on reviewing, editing and post-editing translations

Given improved fluency in AI-driven translation tools, identifying and correcting issues in their output may become more challenging. It is important to emphasize the critical assessment of AI-based translation output and train students to identify common issues such as omissions, problematic structures in NMT, or hallucinations in LLMs. In addition, developing post-editing techniques targeting typical AI errors and strategies for verifying and improving AI translations using official sources and parallel texts can further enhance students' capabilities.

6.5 Design more sophisticated assessment tasks

As AI becomes more prevalent, simple translation tasks might not adequately assess student abilities, especially when they have access to automatic translation tools that produce moderately accurate drafts. It is, therefore, necessary to incorporate more diverse or challenging assessment tasks into the teaching and learning of practical translation. These tasks should require complex contextual analyses, and students need to demonstrate their ability to verify and correct AI output, as well as explain the corrections made. This approach can better assess students' actual performance and encourage them to pay attention to the development of these essential skills for AI-assisted translation. Real-world simulation tasks, in which students collaborate with AI tools to manage complex projects utilizing various resources and technologies, can also provide a more accurate evaluation of students' performance.

7 Recommendations for Developers

To better equip language professionals to work effectively with cutting-edge technologies and identify potential areas for new application development, this section offers recommendations for developers, based on Siu (2023a).

7.1 Enhance Domain and Multilingual Support for Models

It would be desirable to create models that deliver improved domain-specific translations by utilizing domain data or adaptation techniques. To enhance multilingual support, developers can incorporate more non-English data during training and exploring zero-shot translation for low-resource language pairs.

7.2 Improve Integration of NMT, LLMs, and Other Tools

Integrating AI technologies such as translation memory, termbases, and quality estimation models can enhance translation accuracy while addressing issues like hallucination. Merging outputs from multiple sources allows for the incorporation of accurate information in translations, and the development of hybrid systems can leverage the strengths of various technologies.

7.3 Develop AI for Cross-Modal Translation

With the exponential growth of digital multimedia, cross-modal translation is becoming increasingly important. It would be advantageous to develop AI solutions that support multimodal, multilingual translation going beyond text and speech, encompassing images, videos, and other media forms. Investigating the potential of cross-modal attention and multimodal embeddings can enhance multimedia translation quality and enable seamless communication across modalities.

7.4 Improve Interpretability and User Interaction

Working on techniques that enhance the interpretability of AI translation tools would be beneficial, as it allows users to better understand translation choices and pinpoint areas for improvement. The implementation of user-friendly interfaces promotes interaction between translators and AI tools, offering opportunities for feedback, clarification requests, or alternative translations as needed.

8 Concluding Remarks

In this work, we have presented a comprehensive analysis of neural machine translation (NMT) and large language models (LLMs), two major drivers of innovation in translation technology. We have explained how these AI-based systems are built using deep learning techniques and the Transformer model architecture, highlighting their strengths while addressing their limitations. Based on our analysis and insights, we have offered recommendations for educators and developers.

AI has immense potential to augment human translation through NMT and LLMs, but only if developed and applied strategically with human guidance and oversight. By understanding how these technologies work, including their strengths and limitations, translation educators, developers, and translators can forge strong partnerships to maximize the benefits of AI. This can broaden the reach of translation, fostering connections among people across diverse languages and communication channels. As human translators continue advancing the field, AI serves as a powerful tool to unlock human potential and transforms translation by complementing human intelligence.

References

Anil, R., A. M. Dai, O. Firat, M. Johnson, D. Lepikhin, A. Passos, S. Shakeri, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Bahdanau, D., K. Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bubeck, S., V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, et al. 2023. Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

Chu, C., and R. Wang. 2018. A survey of domain adaptation for neural machine translation. *arXiv preprint arXiv:1806.00258*.

Devlin, J., M. W. Chang, K. Lee, and K. Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Edunov, S., M. Ott, M. Auli, and D. Grangier. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 489-500. Association for Computational Linguistics.

Fan, A., M. Lewis, and Y. Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 889–898. Association for Computational Linguistics.

Gao, Y., R. Wang, and F. Hou. 2023. How to design translation prompts for ChatGPT: An empirical study." *arXiv preprint arXiv:2304.02182*.

Goodfellow, I., Y. Bengio, A. Courville, and Y. Bengio. (2016). *Deep learning*. Cambridge: MIT Press.

Hinton, G., O. Vinyals, and J. Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

HKSAR Government. 2023a. 行政長官出席 2022 年特區政府施政十件大事評選頒獎典禮致辭. https://www.info.gov.hk/gia/general/202301/18/P2023011800598.htm. Accessed 3 July 2023.

Holtzman, A., J. Buys, L. Du, M. Forbes, and Y. Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Hong Kong and Macao Affairs Office of the State Council. 2023. 國務院港澳辦發言人：內地與香港『通關』工作正有序推進. https://big5.hmo.gov.cn/gate/big5/www.hmo.gov.cn/xwzx/xwfb/xwfb_child/202111/t20211104_22957.html.

Jiao, W., W. Wang, J. Huang, X. Wang, and Z. Tu. 2023b. Is ChatGPT a good translator? Yes with GPT-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Jurafsky, D., and J. H. Martin. 2023. *Speech and language processing* (3rd ed. draft). https://web.stanford.edu/~jurafsky/slp3/ed3book_jan72023.pdf. Accessed 2 July 2023.

Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781.*

Khan, S., M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah. 2022. Transformers in vision: A survey. *ACM computing surveys* 54(10): 1-41.

Khoshafah, F. 2023. ChatGPT for Arabic-English translation: Evaluating the accuracy. https://doi.org/10.21203/rs.3.rs-2814154/v2. Accessed 3 July 2023.

Kingma, D. P., and J. Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koehn, P. and R. Knowles. 2017. Six challenges for neural machine translation. In *First Workshop on Neural Machine Translation*: 28-39. Association for Computational Linguistics.

Kudo, T., and J. Richardson, J. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 66-71. Association for Computational Linguistics.

Landrich Holding Limited. 2020. Share Offer. https://www1.hkexnews.hk/listedco/listconews/sehk/2020/0930/2020093000019.pdf. Accessed 3 July 2023.

LeCun, Y., Y. Bengio, and G. Hinton. 2015. Deep learning. *Nature* 521: 436-444.

OpenAI. 2022. Introducing ChatGPT. https://openai.com/blog/chatgpt. Accessed 3 July 2023.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35: 27730-27744.

Papineni, K., S. Roukos, T. Ward, T., and W. J. Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311-318. Association for Computational Linguistics.

Peng, K., D. Liang, Q. Zhong, L. Shen, X. Liu, M. Zhang, Y. Ouyang, et al. 2023. Towards making the most of ChatGPT for machine translation. *arXiv preprint arXiv:2303.13780*.

Pennington, J., R. Socher, and C. D. Manning. 2014. GloVe: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 1532-1543. Association for Computational Linguistics.

Popović, M. 2015. CHRF: Character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, 392-395. Association for Computational Linguistics.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. 2019. Language models are unsupervised multitask learners. https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf. Accessed 3 July 2023.

Radford, A., K. Narasimhan, T. Salimans, and I. Sutskever. 2018. Improving language understanding by generative pre-training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. Accessed 3 July 2023.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21(1): 5485-5551.

Rei, R., C. Stewart, A. C. Farinha, and A. Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2685-2702. Association for Computational Linguistics.

Sellam, T., D. Das, and A. Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7881-7892. Association for Computational Linguistics.

Sennrich, R., B. Haddow, and A. Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 86-96. Association for Computational Linguistics.

Siu, S. C. 2019. IPOTranslate: An Online System for English-Chinese Translation of IPO Prospectuses. https://ipotranslate.com/static/IPOTranslate.pdf. Accessed 3 July 2023.

Siu, S. C. 2022. COVID-19 MT Evaluator: A platform for the evaluation of machine translation of public health information related to COVID-19. In *Translation and Interpreting in the Age of COVID-19*, ed. K. Liu and A.K.F. Cheung, 85-109. Singapore: Springer.

Siu, S. C. 2023a. ChatGPT and GPT-4 for professional translators: Exploring the potential of large language models in translation. Available at SSRN: http://dx.doi.org/10.2139/ssrn.4448091.

Siu, S.C. 2023b. Deep learning and translation technology. In *Routledge Encyclopedia of Translation Technology* (2nd ed.), ed. S. W. Chan, 797-817. London: Routledge.

Stahlberg, F. 2020. Neural machine translation: A review. *Journal of Artificial Intelligence Research* 69: 343–418.

Sutskever, I., J. Martens, G. Dahl, and G. Hinton. 2013. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*: 1139-1147. PMLR.

Tay, Y., M. Dehghani, D. Bahri, and D. Metzler. 2020. Efficient Transformers: A survey. *arXiv preprint arXiv:2009.06732*.

Tay, Y., V.Q. Tran, S. Ruder, J. Gupta, H.W. Chung, D. Bahri, Z. Qin, et al. 2021. Charformer: Fast character transformers via gradient-based subword tokenization. *arXiv preprint arXiv:2106.12672*.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 6000-6010. NIPS Foundation.

Weiser, B. 2023, May 27. Here's what happens when your lawyer uses ChatGPT. *The New York Times*.

Xiong, R., Y. Yang, D. He, K. Zheng, S. Zheng, C. Xing, H. Zhang, et al. 2020. On layer normalization in the transformer architecture. In *International Conference on Machine Learning*, 10524-10533. PMLR.

Zhang, A., Z. C. Lipton, M. Li, and A. J. Smola. 2021. Dive into deep learning. *arXiv preprint arXiv:2106.11342*.

Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. 2019. BERTScore: Evaluating text generation with BERT. *arXiv preprint arXiv:1904.09675*.