Enhancing Entertainment Translation for Indian Languages Using Adaptive Context, Style and LLMs

Pratik Rakesh Singh, Mohammadi Zaki and Pankaj Wasnik

Media Analysis Group, Sony Research India, Bangalore
{pratik.singh, mohammadi.zaki, pankaj.wasnik}@sony.com

Abstract

We address the challenging task of neural machine translation (NMT) in the entertainment domain, where the objective is to automatically translate a given dialogue from a source language content to a target language. This task has various applications, particularly in automatic dubbing, subtitling, and other content localization tasks, enabling source content to reach a wider audience. Traditional NMT systems typically translate individual sentences in isolation, without facilitating knowledge transfer of crucial elements such as the context and style from previously encountered sentences. In this work, we emphasize the significance of these fundamental aspects in producing pertinent and captivating translations. We demonstrate their significance through several examples and propose a novel framework for entertainment translation, which, to our knowledge, is the first of its kind. Furthermore, we introduce an algorithm to estimate the context and style of the current session and use these estimations to generate a prompt that guides a Large Language Model (LLM) to generate high-quality translations. Our method is both language and LLM-agnostic, making it a general-purpose tool. We demonstrate the effectiveness of our algorithm through various numerical studies and observe significant improvement in the COMET scores over various state-of-the-art LLMs. Moreover, our proposed method consistently outperforms baseline LLMs in terms of win-ratio.

Introduction

Recent advancements in neural machine translation (NMT) have become increasingly important in the entertainment industry for automatic content localization. These advancements have addressed some limitations of entertainment translation by incorporating contextual understanding and cultural nuances into translations (Yao et al. 2024; Matusov, Wilken, and Georgakopoulou 2019; Vincent et al. 2024a).

In entertainment content, where dialogues often depend on prior interactions to convey a scene's meaning and emotion effectively, context-aware translation plays a vital role (Vu, Kamigaito, and Watanabe 2024; Maruf, Saleh, and Haffari 2021; Vincent et al. 2024b; Agrawal et al. 2023). Incorporating the broader dialogue or narrative context, rather than translating sentences in isolation, is crucial to ensure accurate and emotionally relevant translations (McClarty 2014).

Examples
Source: bro fruits belong to the one who has got the gun **Desired Translation:** भाई माल उसका होता है जिसके हाथ में गन होती है। **Chat GPT:** भाई, फल उसी के होते हैं जिसके पास बंदूक होती है। **Error:** Simile and Contextual
Source: will do it from your roof top I will shout in front of everyone. **Desired Translation:** कल हर घर जा कर तेरी बदनामी करूंगा! **Chat GPT:** तेरी छत से ही मैं चिल्लाऊँगा, सबके सामने चिल्लाऊँगा। **Error:** Contextual, Literal Translation
Source: okay okay. alright get ready. cat behind the mouse. mouse in front of the cat too much fan man too much fun. **Desired Translation:** ठीक है ठीक है। हाँ हो जा रेडी तू डेविल के पीछे डेविल तेरे पीछे टू मच फन टू मच फन हा हा **Chat GPT:** ठीक है, तैयार हो जाओ। बिल्ली माउस के पीछे। माउस बिल्ली के सामने। बहुत सारा पंखा, यार, बहुत मज़ा! **Error:** Literal Translation, Idiomatic

Table 1: Example Translations and Errors

On the other hand, entertainment translation also needs a culturally adaptable system to address the challenge of cultural unawareness (Etchegoyhen et al. 2014; Yao et al. 2024). Such systems should integrate cultural context for localization to ensure translations are suitable for the intended audience. They should go beyond literal translations, modifying idiomatic expressions, jokes, and cultural references to align with the audience's customs and values, thereby enhancing the relevance of the translated content(Gupta et al. 2019; Li et al. 2024). In Table 1 we show some examples of common mistakes made by NMT systems when translating entertainment content. In Example 1, 'fruits' idiomatically refers to 'reward,' but ChatGPT's literal translation misses this. In Example 2, the desired translation is culturally more creative, aligning with native Hindi speakers by conveying "I will badmouth you by knocking door to door." In Example 3, the desired translation uses idiomatic language effectively, unlike ChatGPT's literal approach.

In this paper, we address the challenging task of entertainment translation, where we are given a sequence of source sentences from the entertainment domain without any additional information about the timestamp, speaker ID, or context, and our task is to translate these sentences into dialogues in the target language. The challenge lies in preserving the context, mood, and style of the original content

while also incorporating creativity and considering regional dialects, idioms, and other linguistic nuances (Gupta et al. 2019). The importance of our study is underscored by the need to produce translations that are not only accurate but also engaging for the target audience.

In particular, we treat the entertainment translation task as a sequential process to extract time-dependent contextual information by dividing the input text into a series of sessions. We primarily employ context-retrieval and domain adaptation to facilitate in-context learning of Large Language Models to extract both the style, representing the cultural nuances and temporal context from these *sessions*. We can then use this characteristic information to generate culturally enriched translations. In addition, our proposed methodology does not need auxiliary information such as speaker information, timestamps, and conversation mood, making it generalized and applicable in a wide range of applications. Our key contributions can be summarized as follows:

- We proposed an algorithm (Alg. 2), which we call Context And Style Aware Translation (CASAT). It incorporates context and style awareness, enhancing the input prompt and enabling LLM to produce culturally relevant translations.
- Proposed methodology is language and LLM-agnostic. further, it does not rely on dialogue timestamps, speaker identification, etc., making it a versatile approach.
- We proposed Context retrieval–Advanced RAG module to extract a precise and relevant context from entertainment content such as a movie or series episode.
- We proposed a Domain Adaptation Module to provide a cultural understanding of input to LLMs.

Background and Motivation

In this section, we provide a review of some of the major research works in the field of machine translation as well as applications of LLMs in NMT.

NMT was introduced in the seminal works of (Bahdanau, Cho, and Bengio 2015; Cho et al. 2014), who used basic encoder-decoder architectures and RNNs, respectively, for the NMT task. These techniques were superseded by attention-based mechanisms introduced in (Luong, Pham, and Manning 2015; Wu et al. 2016). With the advent of Transformers in (Vaswani et al. 2017), the attention computation became massively parallelized, increasing the speed and efficiency of modern NMT systems.

LLMs for NMT: In the last couple of years, LLMs have caused a major shift in the way AI research is carried out (Brown et al. 2020). The translation task has become a goto application of the LLMs since their advent (Lyu et al. 2024). A comprehensive review of machine translation using LLMs can be found in (Cai et al. 2024).

Entertainment Translation: Most of the previously presented research on entertainment domain translation focuses primarily on subtitling and segmentation (Vincent et al. 2024b; Karakanta et al. 2022; Vincent et al. 2024a; Matusov, Wilken, and Georgakopoulou 2019; Etchegoyhen et al. 2014). These works depend on additional information like timestamps and speaker details from the input text. However, timestamp information may not always be present or could be incorrect, leading to ambiguity or distortions in the temporal context, making entertainment translation more challenging (Gaido et al. 2024).

Use of contextual information for NMT: In recent years, the importance of (correct) context in the translation task has been studied and highlighted (Voita, Sennrich, and Titov 2019) for document-level translations (Maruf, Saleh, and Haffari 2021). However, these approaches do not perform consistently while dealing with overly large contexts or complicated scenarios (Vu, Kamigaito, and Watanabe 2024), as is usually the case in the entertainment domain.

LLMs for Creative Translations and Style Transfer: Use of LLMs to induce creativity can be accomplished to a certain extent using prompt engineering techniques (Zhang, Haddow, and Birch 2023). In addition, advanced retrieval-based techniques (Agrawal et al. 2023; Reheman et al. 2023; Glass et al. 2022) can be used to generate context from a given text and be used to provide necessary information for the desired translations. On the other hand, recent work on style transfer (Tao et al. 2024) introduces a Domain Adaptation Module to copy the style of the input text to be used for modifying the LLM-based translations. However, all these methods are static; that is, they do not change with respect to the variation in the mood, genre, or context, which is an inherent property of the entertainment content. Similarly, Li et al. (2024) tries to induce cultural nuances of the target language by introducing a knowledge base (KB) for idioms, which are difficult to translate in general. However, these models do not cover Indian languages, which have their own structural and lexical nuances (Leong et al. 2023).

LLMs for Entertainment Translation: Machine Translation using LLMs has started to gain popularity in recent times (Brown et al. 2020; Zhang, Haddow, and Birch 2023; Tao et al. 2024). Broadly, this can be classified into two categories: (i) prompt-based guiding and (ii) translation memory/RAG-based translation aiding. Below, we point out the issues with these techniques when applied to entertainment translation.

(i) **Prompt-based Guiding:** Prompt-based guiding of LLMs to perform translation can be treated as providing a conditioning parameter p, *viz*., the *prompt*, to the translation model:

$$P_\theta\left(y|x,p\right) = \prod_{i=1}^{L} P_\theta\left(y_i|p,x,y_1,\ldots,y_{i-1}\right).$$

where L is the length of the output sentence y. However, when working in the automatic dubbing application for movies and OTT content, the prompt needs to be time-dependent, i.e. $p \rightarrow p_t$, in order to deal with the dynamic context c_t. In particular, the prompt can be formulated

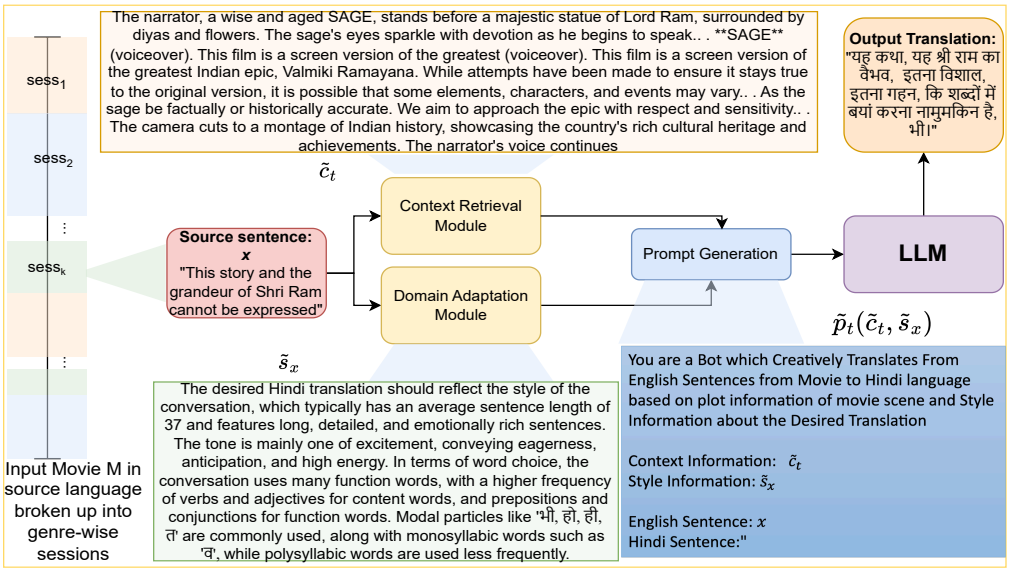

Figure 1: A high-level overview of our proposed methodology.

as $p_t = h(p, c_t)$, where h is a linking and weight function in a latent space. The adaptive nature of the prompt p_t induced by the time-varying context c_t is vital in generating context-relevant translations for dubbing applications. However, it has received limited attention from researchers (Gao et al. 2023).

(ii) **Translation memory-based approach:** Traditional retrieval-aided translation systems have two primary components: (i) a retriever $p_\eta(.|x)$ which gives a probability distribution over a set of hidden context vectors stored in a vector database, and (ii) a generator $p_w(.|x, z)$ which gives a probability distribution over the output tokens given the source sentence x and context z. The retriever aims at providing additional information to the generator, which is an LLM performing translation, by retrieving context z by Maximum Inner-Product Search (MIPS) (Lewis et al. 2021). However, the retrieved context vectors z are semantically similar to the query sentence x and do not take into account the style s_x of the source sentence, for example, politeness, (in-)formality, regional dialect, etc. (Tao et al. 2024)

Potential Resolution

The above-mentioned limitations reflect the need for a machine translation system that takes into account the context c_t and preserves the style s_x of a given source input sentence x. To this extent, a potential solution is to segment the (sequential) text into *sessions*, where the 'genre' of the sentences in a session remains constant. These 'constant mood' sessions can be used to estimate the context and style, i.e., $\tilde{c}_t$ and $\tilde{s}_x$. By incorporating this additional information, a time-varying prompt $p_t(\tilde{c}_t, \tilde{s}_x)$ can be obtained to leverage LLM's reasoning and understanding capabilities for generating context and style-aware translations.

Algorithm 1: Genre Classification and Segmentation

Input: $\mathcal{M}$, clusters of the three classes, minimum number of sentences for a new session (α), maximum number of sentences in a session (β)
Extract embeddings for each $x \in \mathcal{M}$ and use k-NN to assign its class label and store in a array g.
current-session $\leftarrow \{g[0]\}$
session-list $\leftarrow \emptyset$
while $i <$ length(g) **do**
  if length(current-session) == β **then**
    session-list $\leftarrow$ current-session
    current-session $\leftarrow \emptyset$
  end if
  if $g[i] \neq$ current-session[0] $\wedge$ length(current-session) $\geqslant \alpha$ **then**
    majority-label $\leftarrow$ MAJORITY$(g[i] : g[i + \alpha])$
    if majority-label $\neq$ current-session[0] **then**
      session-list $\leftarrow$ current-session
      current-session $\leftarrow \emptyset$
    end if
  end if
  current-session $\leftarrow g[i]$
  $i \leftarrow i + 1$
end while
Output: session-list

Methodology

In this section, we describe our methodology beginning with stating the problem statement formally. Next, we explain the necessity of segmenting the input text and how to obtain it. We then describe the method for extracting the $\tilde{c}_t$ and $\tilde{s}_x$ for a particular dialogue x to generate the context and style-aware prompt p_t.

Problem Formulation

We consider the entertainment translation as an extension of neural machine translation task, where we primarily try to translate sentences from a source language ($\mathcal{S}$) to a target language ($\mathcal{T}$). These sentences can be dialogues from movies, web series, novels, etc. Formally, let $\mathcal{D}^{\mathcal{S}}$ be defined as the set of all sentences in a $\mathcal{S}$ and $\mathcal{D}^{\mathcal{T}}$ the corresponding set in $\mathcal{T}$. The goal of a translation system is to find a mapping $g : D^{\mathcal{S}} \mapsto \mathcal{D}^{\mathcal{T}}$. However, translating movie dialogues from one language to another requires additional knowledge of the running *context* c_t as well as the style s_x of the source sentence x. Hence, we define a mapping g_E, which is specific to translation in the entertainment domain, as a function that outputs the translated text y as $y = g_E(x; c_t, s_x)$. In other words, the aim of an entertainment translation system is to find a mapping g_E which not only translates any $x \in \mathcal{D}^{\mathcal{S}}$, into a sentence $y \in \mathcal{D}^{\mathcal{T}}$ but also preserves the context c_t and the style s_x of the input source sentence. Further, the mapping learned should be such that it induces creativity in the translation, which can increase the target audience's interest and engagement. These additional factors make the task of entertainment translation unique and challenging.

Adaptive Session Classification and Segmentation

For this section, we will take a concrete example of a movie $\mathcal{M}$ to explain the key concepts (note: we only consider the sequence of text dialogues as $\mathcal{M}$). In entertainment content, each movie or web series is characterized by a sequence of scenes, each belonging to a specific *genre* or tone, such as action, horror, comedy, and so forth. Therefore, a movie $\mathcal{M}$ can be represented as $\mathcal{M} = (sess_1, sess_2, \ldots, sess_M)$, where M is the total number of scenes/sessions in the movie. Suppose a dialogue $x \in sess_k$, the style of translation of x is expected to be more likely dependent on the current and K neighboring sessions than much older sessions, which necessitates the segmentation of the text to ensure translation quality.

We provide an offline algorithm for achieving adaptive segmentation of $\mathcal{M}$ in Alg. 1, which classifies each session into one of the three three primary tonal categories: Serious (Intense genres: action, mystery, thriller, horror), Casual (Light genres: comedy, romance, fantasy) and Neutral (Dialogues with low emotional intensity). While not all the input texts may fit perfectly into these three categories, this approach provides a foundation for simple yet consistent classification by grouping genres with similar tones. We pretrain a k-NN classifier and generate clusters using example dialogues from the three categories. We refer the reader to the Appendix for details on the segmentation process. We also remark that Alg. 1 only provides a rough estimate of the session boundaries of $\mathcal{M}$. Next, we demonstrate how we extract the context and style information from the available sessions.

Session Information Generation

This section provides a thorough insight to the crux of our method. Let the current input dialogue be x. As depicted in Figure 1, x passes through two separate pipelines for $\tilde{c}_t$ and $\tilde{s}_x$ extraction. Subsequent paragraphs provide detailed description of these blocks.

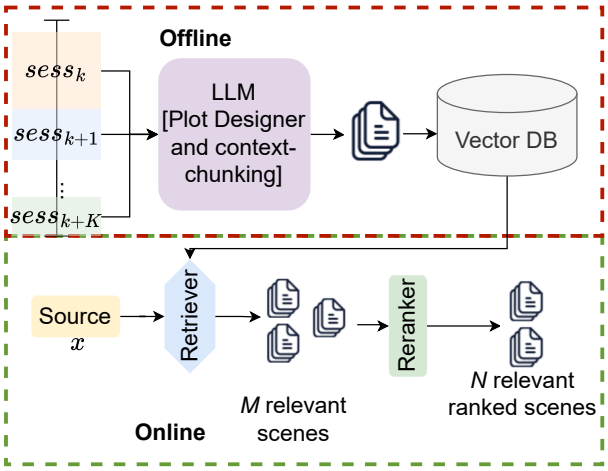

Figure 2: A block diagram of the Context retriever block.

Context retrieval–Advanced RAG: Using Large Language Models (LLMs) to translate dialogues from one language to another without any prior context can lead to *disconnected* translations, especially in a conversation. In order to induce interest among the target audience, LLMs can generate creative translations, which may lead to hallucinations (Zhang et al. 2023). Hence, providing the current session information can guide the LLM in translating the source sentence creatively with respect to the context of the movie, reducing hallucinations.

As depicted in Figure 2, we consider an offline process to extract the *plots*, *i.e.*, a summary of movie scenes, from K consecutive sessions via an LLM. This extracted context is then subdivided into small chunks and stored in a vector database. This chunking helps our methodology two-fold. Firstly, the generated prompt might be too large for the LLM to comprehend. Secondly, the most relevant chunk/scene for the source sentence could well be from a different session (in the past or in the future). During the translation phase, a retriever uses the source sentence x to retrieve M most relevant chunks from the vector database (Lewis et al. 2021). This is then passed through a re-ranker (Glass et al. 2022), to generate N most relevant chunks in a ranked fashion, which we denote as $\tilde{c}_t$ for sentence x.

Style extraction-Domain Adaptation Module: By using the above pipeline for context information extraction, we can generate creative translation that aligns with the current context and mood of the scene. However, this does not help in extracting the style or tone of the dialogue. In particular, $\tilde{c}_t$ does not include the most used words, idioms, and emotional state of the current scene, which define the overall language register. To tackle this, we designed a Domain Adaptation Module (DAM), which is a collection of various information-extracting NLP subroutines. These subroutines help in constructing $\tilde{s}_x$, which acts as a clear and comprehensive style-determining prompt to be fed to the LLM. We

note that we get inspired from (Tao et al. 2024) with changes in the DAM module owing to our specific application and the change in language family. In particular, we pay special attention to dialogue and session-level information, respectively, which is in contrast with their approach, which dealt with style transfer as a one-shot method for the entire text at once. Subsequently, we explain these modules in detail.

Dialogue Level Module: This module provides the structural information of the dialogues, giving us the overall conversational style of speakers. It consists of three parts as described in brief below.

- **Content and Function words:** Here, we take the output translations of the past K sessions as input and pass it through a PoS Tagger trained on Indic languages. We categorize these tagged words into *content words* and *function words*(Carnap 1967), which we then convert to the respective prompts f_c and f_f. For the explicit prompts, we refer the reader to the Appendix.
- **Frequent Syllabic Words:** Every speaker may have a different style of speaking for instance, depending on the regional dialect, pronouns like "I" or "myself" can be termed in Hindi as "*apun*", (spoken in Mumbai region) or "*hum*", (spoken in northern India), etc. Identifying this will provide the model with information on the frequent use of *monosyllabic* and *polysyllabic* words. Similar to the above case, we convert them into prompts as f_m and f_p respectively.
- **Modal Words and Idioms:** Modal words and idioms contribute to the tone, politeness, and effectiveness of the conversation (f_{modal}, f_{idioms} respectively).

Algorithm 2: Context and Style Aware Translation (CASAT)

Input: Source Sentences ($\mathcal{M}$), M, N, session-list (See Alg. 1)
for $x \in \mathcal{M}$ **do**
  $\tilde{c}_t \leftarrow$ Extract M relevant scenes from the vector DB and choose N best through the Context Retriever Module.
  $\tilde{s}_x \leftarrow$ Extract dialogue level and session level information through DAM.
  $p_t \leftarrow$ Generate prompt using $\tilde{c}_t$ and $\tilde{s}_x$
  Translation $y_x \leftarrow$ LLM(p_t, x)
end for

Session Level Module: In contrast with the dialogue-level information extraction, the session-level module allows an understanding of the global intent of the ongoing and past sessions.

- **Sentence Intent and Emotion:** Intent of a session can be derived from the use of punctuation marks. For instance, excessive use of question marks in a particular scene can indicate the scene to be interrogatory. Hence, we count all the punctuation in session, then define intent based on thresholds (f_{intent}). Further, to extract the emotion, we pass the current session through an LLM to generate $f_{emotion}$. Furthermore, we provide a concrete example of the DAM in the Appendix.

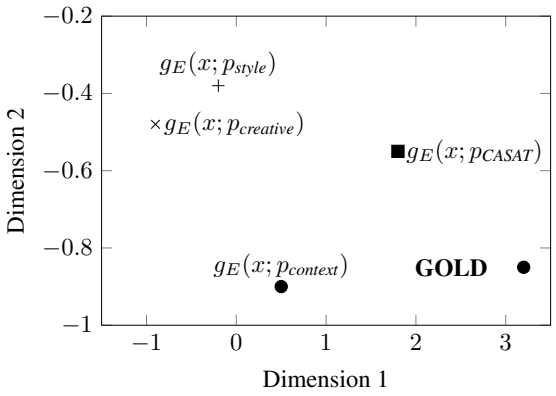

Figure 3: Visualization of embedding projections

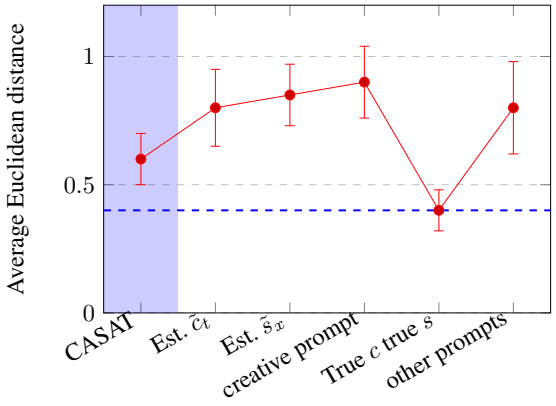

Figure 4: Embedding distances from the reference output

Finally, we obtain the Context and Style Aware prompt p_t, by concatenating the outputs from the context retrieval module ($\tilde{c}_t$) and the DAM module ($\tilde{s}_x$). We refer the reader to the Appendix, where we illustrate detailed examples of prompt p_t for enhanced clarity.

Experiments

In this section, we present the experimental evaluation of our proposed approach. We will also describe the effect caused by the individual components of CASAT through ablation studies. All experiments were carried out on 1x1H100 80 GB GPU.

Experimental Settings

Evaluation Dataset: This section provides details of the datasets used to evaluate our approach, focusing on the data employed for testing Alg. 2 since our method does not have an explicit training phase. In addition, due to the unavailability of Indian language entertainment domain public datasets, we use web-scraped data for our simulations, which will be explained later.

We scrapped parallel text data of popular movies from the popular subtitle website *OpenSubtitles.org*. To see the effect on text data, which requires even more human-induced

LLMs Sizes	Models	en-hi					en-bn					en-te				
		Base		CASAT			Base		CASAT			Base		CASAT		
		B.	C.	B.	C.	Δ	B.	C.	B.	C.	Δ	B.	C.	B.	C.	Δ
Small-Sized LLMs	Mistral 7B	1.33	0.41	1.55	0.42	0.56	0.2	0.43	0.38	0.48	**0.62**	0.1	0.42	0.07	0.41	**0.8**
	LLaMa-3 8B	3.42	0.51	4.51	0.56	0.56	0.75	0.61	1.12	0.67	0.52	0.85	0.51	0.60	0.58	0.69
	Aya23 8B	6.67	0.61	**7.21**	**0.64**	**0.67**	0.30	0.48	0.4	0.53	0.56	0.16	0.42	0.3	0.46	0.76
	Gemma2 9B	6.68	0.56	6.88	0.62	0.62	1.55	0.68	**2.53**	**0.75**	**0.62**	1.43	0.65	**1.75**	**0.69**	0.79
Mid-Sized LLMs	Gemma2 27B	4.71	0.62	8.07	0.67	0.69	1.49	0.70	**3.08**	**0.77**	**0.67**	1.7	0.67	**2.07**	**0.71**	0.77
	Aya23 35B	9.25	0.63	**9.59**	**0.68**	**0.70**	0.80	0.62	0.82	0.65	0.59	0.17	0.44	0.23	0.48	**0.8**
Large-Sized LLMs	LLaMa3 70B	7.96	0.63	9.84	0.70	**0.73**	2.46	0.66	2.09	0.75	0.74	0.92	0.60	1.11	0.65	**0.85**
	GPT-3.5 Turbo	11.84	0.69	**14.44**	**0.72**	**0.73**	12.88	0.79	**14.91**	**0.82**	**0.75**	7.41	0.66	**10.9**	**0.78**	**0.85**

Table 2: Performance comparison of CASAT with various SOTA LLMs fed with prompts to generate creative translations. Here B.:BLEU, C.:COMET score in range [0,1] and Δ is the win-ratio of CASAT-assisted models *vs.* base models.

creativity, we further used parallel text data from a popular children 's cartoon series. All our experiments are conducted on the set of three language directions, *viz.*, English-to-Hindi (en-hi), English-to-Bengali (en-bn) and English-to-Telugu (en-te). Next, we mention the specific details of the text content used for our numerical studies. We label our text data into three categories: (i) literal, (ii) semi-creative, and (iii) creative, owing to the increasing levels of creativity in the reference gold data.

- **English-to-Hindi:** We choose subtitles scrapped from the *opensubtitles* website for the following movies: (i) *Adipurush* (creative), (ii) *Pushpa* (semi-creative), and (iii) *Interstellar* (literal). In addition, we use episodes from a popular cartoon series (creative) for evaluation. The total number of sentence pairs for en-hi was 5238.
- **English-to-Bengali:** Similarly, we scrapped subtitles of two movies, namely *Wolves* and *Maharaja* from *opensubtitles* website was used for evaluation for this language pair, amounting to a total of 3259 sentence pairs.
- **English-to-Telugu:** For English to Telugu translation, we scrapped subtitles of two movies, namely *Without Remorse* and *Bumblebee* from *opensubtitles* for evaluation, comprising of 1698 sentence/dialogue pairs.

LLMs used for comparison: We randomly select 800 data samples from each language source and translate them to the target language utilizing 5 distinct Large Language Models with varying sizes, categorizing them into three sections:

- **Small Sized LLMs:** We focused on three multi-lingual small sized models which perform well in Indic Languages i.e Mistral 7B (Jiang et al. 2023), Gemma2 9B (Gemma-Team 2024), Aya23 8B (Aryabumi et al. 2024), Llama3 8B (Meta-Team 2024).
- **Mid-sized LLMs:** We consider two LLMs, namely, Gemma2 27B (Gemma-Team 2024) and Aya23 35B (Aryabumi et al. 2024), for mid-sized category. Both of these LLMs have performed consistently well in Indic languages.
- **Large-sized LLMs:** Likewise we considered two large-sized LLMs that are Llama3 70B (Meta-Team 2024) and GPT-3.5 Turbo, both having excellent reasoning and translation qualities.

Evaluation Metrics: We adopt three metrics for the evaluation task. SacreBLEU (Post 2018) represents n-gram matching while COMET (wmt22-cometkiwi-da) (Rei et al. 2022) represents the reference-free neural-based evaluation. Third, we use GPT-4o for evaluation of the translated text, which is well-known to replicate human-level judgment (Fu et al. 2023) by calculating the win-ratio (Δ) of our approach over the baseline models as follows:

$$\Delta = \frac{\begin{pmatrix}\#\text{times GPT-4o chooses CASAT based} \\ \text{translation over baseline LLM translation}\end{pmatrix}}{\#\text{total translations}}.$$

Can CASAT provide audience-engaging translations?

Main Result and Analysis: The outcomes presented in Table 2 illustrate that our method demonstrates superior performance by consistently incorporating plot and style information compared to directly prompting creativity in LLMs (see the exact prompt used for baseline LLMs in the Appendix). For the en-hi direction, we observe an increase of 21. 95% in the BLEU score and 7. 4% in the COMET score across all models. In en-bn direction, BLEU score improved by 43.2%, while COMET by 9.3%. For the en-te direction, the improvements were 21. 9% in BLEU and 8.57% in COMET, further demonstrating the effectiveness of CASAT. Secondly, irrespective of the LLM chosen to produce the translation, CASAT significantly improves its quality across the evaluation metrics. Interestingly, Mistral 7B shows minimal enhancement for the en-hi and en-te directions, yet exhibits a commendable win ratio for en-bn and en-te directions. Third, both performance in win ratio *and* COMET scores improve with larger model sizes, suggesting that increasing the model size enhances LLM's capability of plot development and comprehension of the in-context information. However, surprisingly, we observe that the 9B and 27B versions of Gemma2 either perform similarly to or even outperform models such as Aya23 35B and Llama3 70B for en-bn and en-te language directions in terms of COMET scores.

How does the inclusion of context and style impact the resulting output? We plot the multi-dimensional scaling (MDS) representation of the generated text from Llama 3-8B, with varying prompts in Figure 3. We observe that prompting the LLM differently affects the output translation in a significant manner, as also reported in (Salinas and Morstatter 2024). The plot indicates that solely incorporating the style has minimal impact on the translation quality, whereas solely providing the plot information (context) enhances the quality, evident by the reduced distance

Model Size	Models	Base		Context Only			DAM Only			CASAT		
		BLEU	COMET	BLEU	COMET	Δ	BLEU	COMET	Δ	BLEU	COMET	Δ
Small-Sized LLMs	Mistral 7B	1.33	0.41	1.16	0.41	**0.67**	1.89	0.41	0.67	1.55	0.42	0.56
	LLaMa3 8B	3.42	0.50	3.82	0.55	0.58	3.43	0.51	0.55	4.51	0.56	0.60
	Aya23 8B	6.67	0.61	6.77	0.64	0.66	7.01	0.62	0.64	7.21	0.64	**0.67**
	Gemma2 9B	6.68	0.56	**7.80**	**0.67**	0.61	7.14	0.59	0.60	6.88	0.62	0.62
Mid-Sized LLMs	Gemma2 27B	4.71	0.62	7.46	0.66	0.62	5.17	0.64	0.66	8.07	0.67	0.69
	Aya23 35B	9.25	0.63	6.95	0.67	0.67	8.32	0.66	**0.70**	**9.59**	**0.68**	**0.70**
Large-Sized LLMs	LLaMa3 70B	7.96	0.62	7.12	0.66	0.71	8.99	0.67	0.64	**9.84**	**0.70**	**0.73**

Table 3: Analysis of the effect of the individual components of CASAT.

between the context and reference in comparison to style alone. CASAT, i.e., the simultaneous provision of context and style, significantly enhances the quality of the translation. Figure 4 plots the average Euclidean distance of the generated text from the reference translations for a range of prompts. The plot shows that CASAT is closest to the reference translation.

Models	K=1	K=2	K=3	K=4
Mistral 7B	0.367	**0.424**	0.402	0.371
Llama3 8B	0.487	**0.562**	0.534	0.507
Gemma2 9B	0.644	0.62	0.647	**0.658**
Aya23 8B	0.637	0.64	**0.65**	0.644
Gemma2 27B	0.66	**0.67**	0.64	0.63
Aya23 35B	0.67	0.68	**0.69**	0.66
Llama3 70B	0.68	**0.70**	0.67	0.66

Table 4: COMET scores showing the effect of varying the value of number of sessions K .

How does CASAT fare against traditional MT Systems? We evaluate the Win-Ratio (Δ) of CASAT-augmented models against traditional machine translation (MT) systems across en-hi, en-bn, and en-tl translation directions. Specifically, we compare the performance of Gemma2 9B (CG9) and Gemma2 27B (CG27) models, enhanced with the CASAT approach, against traditional systems such as IndicTrans2 (ITv2) (Gala et al. 2023) and NLLB (Team et al. 2022). The results, summarized in Table 5, demonstrate that CASAT-augmented models are consistently preferred in the entertainment domain, underscoring the effectiveness of the CASAT approach in improving translation quality, particularly in domain-specific contexts.

How many sessions K to consider? The performance of all models on en-hi language pair datasets are compared for various values of K in Table 4. Since K is utilized in plot design and DAM, it is a crucial parameter to consider. Generally, it has been observed that $K = 2$ and $K = 3$ exhibit good performance. The results indicate that using $K = 1$ yields insufficient contextual information, while K=4 results in less specificity.

Models	en-hi	en-bn	en-te
CG9 vs ITv2	58%	51%	53%
CG27 vs ITv2	65%	58%	53%
CG9 vs NLLB	66%	51%	54%
CG27 vs NLLB	64%	61%	68%

Table 5: CASAT *vs.* traditional MT systems win-ratio Δ.

Ablation Studies

We conduct ablation studies on the effect of the domain adaptation module for style transfer and the context retriever block and compare the results with the respective baseline LLMs. Table 3 presents BLEU scores, COMET scores and win-ratios for all the considered open-source LLMs. In the Table, the highlighted numbers represent the best performing model in the corresponding model category as per the model size. We observe that providing 'context only' improves the relevancy of output translation, which is reflected in COMET and win ratio scores. On the other hand, 'DAM only' helps to navigate the output to copy the style of text and hence a larger value for the metric BLEU. Finally, combining the two, i.e., for CASAT, we obtain better BLEU score, COMET, and win-ratios across LLMs, which we conjecture that the LLM is able to gain complementary information from each of the two blocks.

Conclusion

We explored the challenging task of entertainment translation, where we identified two key aspects, context, and style, which make this problem unique. We proposed a methodology to estimate these factors and use them to generate context and style-aware translations from an LLM. We showcased the efficacy of our algorithm via numerous experiments using three Indian language entertainment text datasets and various LLMs. Further, our approach has an offline component for partitioning of sessions and generation of contextual information, which we intend to eliminate to develop a completely online algorithm.

References

Agrawal, S.; Zhou, C.; Lewis, M.; Zettlemoyer, L.; and Ghazvininejad, M. 2023. In-context Examples Selection for Machine Translation. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 8857–8873. Toronto, Canada: Association for Computational Linguistics.

Aryabumi, V.; Dang, J.; Talupuru, D.; Dash, S.; Cairuz, D.; Lin, H.; Venkitesh, B.; Smith, M.; Campos, J. A.; Tan, Y. C.; Marchisio, K.; Bartolo, M.; Ruder, S.; Locatelli, A.; Kreutzer, J.; Frosst, N.; Gomez, A.; Blunsom,

P.; Fadaee, M.; Üstün, A.; and Hooker, S. 2024. Aya 23: Open Weight Releases to Further Multilingual Progress. arXiv:2405.15032.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Cai, W.; Jiang, J.; Wang, F.; Tang, J.; Kim, S.; and Huang, J. 2024. A Survey on Mixture of Experts. arXiv:2407.06204.

Carnap, R. 1967. *The Logical Syntax of Language*. International library of psychology, philosophy and scientific method. Routledge & Kegan Paul.

Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In Moschitti, A.; Pang, B.; and Daelemans, W., eds., *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734. Doha, Qatar: Association for Computational Linguistics.

Etchegoyhen, T.; Bywood, L.; Fishel, M.; Georgakopoulou, P.; Jiang, J.; van Loenhout, G.; del Pozo, A.; Maučec, M. S.; Turner, A.; and Volk, M. 2014. Machine Translation for Subtitling: A Large-Scale Evaluation. In Calzolari, N.; Choukri, K.; Declerck, T.; Loftsson, H.; Maegaard, B.; Mariani, J.; Moreno, A.; Odijk, J.; and Piperidis, S., eds., *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 46–53. Reykjavik, Iceland: European Language Resources Association (ELRA).

Fu, J.; Ng, S.-K.; Jiang, Z.; and Liu, P. 2023. GPTScore: Evaluate as You Desire. arXiv:2302.04166.

Gaido, M.; Papi, S.; Negri, M.; Cettolo, M.; and Bentivogli, L. 2024. SBAAM! Eliminating Transcript Dependency in Automatic Subtitling. arXiv:2405.10741.

Gala, J.; Chitale, P. A.; Raghavan, A. K.; Gumma, V.; Doddapaneni, S.; M, A. K.; Nawale, J. A.; Sujatha, A.; Puduppully, R.; Raghavan, V.; Kumar, P.; Khapra, M. M.; Dabre, R.; and Kunchukuttan, A. 2023. IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages. *Transactions on Machine Learning Research*.

Gao, J.; Xiang, L.; Wu, H.; Zhao, H.; Tong, Y.; and He, Z. 2023. An Adaptive Prompt Generation Framework for Task-oriented Dialogue System. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 1078–1089. Singapore: Association for Computational Linguistics.

Gemma-Team. 2024. Gemma 2: Improving Open Language Models at a Practical Size. arXiv:2408.00118.

Glass, M.; Rossiello, G.; Chowdhury, M. F. M.; Naik, A. R.; Cai, P.; and Gliozzo, A. 2022. Re2G: Retrieve, Rerank, Generate. arXiv:2207.06300.

Gupta, P.; Sharma, M.; Pitale, K.; and Kumar, K. 2019. Problems with automating translation of movie/TV show subtitles. arXiv:1909.05362.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. arXiv:2310.06825.

Karakanta, A.; Bentivogli, L.; Cettolo, M.; Negri, M.; and Turchi, M. 2022. Post-editing in Automatic Subtitling: A Subtitlers' perspective. In Moniz, H.; Macken, L.; Rufener, A.; Barrault, L.; Costa-jussà, M. R.; Declercq, C.; Koponen, M.; Kemp, E.; Pilos, S.; Forcada, M. L.; Scarton, C.; Van den Bogaert, J.; Daems, J.; Tezcan, A.; Vanroy, B.; and Fonteyne, M., eds., *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, 261–270. Ghent, Belgium: European Association for Machine Translation.

Leong, W. Q.; Ngui, J. G.; Susanto, Y.; Rengarajan, H.; Sarveswaran, K.; and Tjhi, W. C. 2023. BHASA: A Holistic Southeast Asian Linguistic and Cultural Evaluation Suite for Large Language Models. arXiv:2309.06085.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; tau Yih, W.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401.

Li, S.; Chen, J.; Yuan, S.; Wu, X.; Yang, H.; Tao, S.; and Xiao, Y. 2024. Translate Meanings, Not Just Words: IdiomKB's Role in Optimizing Idiomatic Translation with Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17): 18554–18563.

Luong, T.; Pham, H.; and Manning, C. D. 2015. Effective Approaches to Attention-based Neural Machine Translation. In Màrquez, L.; Callison-Burch, C.; and Su, J., eds., *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1412–1421. Lisbon, Portugal: Association for Computational Linguistics.

Lyu, C.; Du, Z.; Xu, J.; Duan, Y.; Wu, M.; Lynn, T.; Aji, A. F.; Wong, D. F.; and Wang, L. 2024. A Paradigm Shift: The Future of Machine Translation Lies with Large Language Models. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 1339–1352. Torino, Italia: ELRA and ICCL.

Maruf, S.; Saleh, F.; and Haffari, G. 2021. A Survey on Document-level Neural Machine Translation: Methods and Evaluation. *ACM Comput. Surv.*, 54(2).

Matusov, E.; Wilken, P.; and Georgakopoulou, Y. 2019. Customizing Neural Machine Translation for Subtitling. In

Bojar, O.; Chatterjee, R.; Federmann, C.; Fishel, M.; Graham, Y.; Haddow, B.; Huck, M.; Yepes, A. J.; Koehn, P.; Martins, A.; Monz, C.; Negri, M.; Névéol, A.; Neves, M.; Post, M.; Turchi, M.; and Verspoor, K., eds., *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, 82–93. Florence, Italy: Association for Computational Linguistics.

McClarty, R. 2014. In support of creative subtitling: contemporary context and theoretical framework. *Perspectives*, 22: 592 – 606.

Meta-Team. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Post, M. 2018. A Call for Clarity in Reporting BLEU Scores. In Bojar, O.; Chatterjee, R.; Federmann, C.; Fishel, M.; Graham, Y.; Haddow, B.; Huck, M.; Yepes, A. J.; Koehn, P.; Monz, C.; Negri, M.; Névéol, A.; Neves, M.; Post, M.; Specia, L.; Turchi, M.; and Verspoor, K., eds., *Proceedings of the Third Conference on Machine Translation: Research Papers*, 186–191. Brussels, Belgium: Association for Computational Linguistics.

Reheman, A.; Zhou, T.; Luo, Y.; Yang, D.; Xiao, T.; and Zhu, J. 2023. Prompting Neural Machine Translation with Translation Memories. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11): 13519–13527.

Rei, R.; Treviso, M.; Guerreiro, N. M.; Zerva, C.; Farinha, A. C.; Maroti, C.; de Souza, J. G. C.; Glushkova, T.; Alves, D. M.; Lavie, A.; Coheur, L.; and Martins, A. F. T. 2022. CometKiwi: IST-Unbabel 2022 Submission for the Quality Estimation Shared Task. arXiv:2209.06243.

Salinas, A.; and Morstatter, F. 2024. The Butterfly Effect of Altering Prompts: How Small Changes and Jailbreaks Affect Large Language Model Performance. arXiv:2401.03729.

Tao, Z.; Xi, D.; Li, Z.; Tang, L.; and Xu, W. 2024. CAT-LLM: Prompting Large Language Models with Text Style Definition for Chinese Article-style Transfer. arXiv:2401.05707.

Team, N.; Costa-jussà, M. R.; Cross, J.; Çelebi, O.; Elbayad, M.; Heafield, K.; Heffernan, K.; Kalbassi, E.; Lam, J.; Licht, D.; Maillard, J.; Sun, A.; Wang, S.; Wenzek, G.; Youngblood, A.; Akula, B.; Barrault, L.; Gonzalez, G. M.; Hansanti, P.; Hoffman, J.; Jarrett, S.; Sadagopan, K. R.; Rowe, D.; Spruit, S.; Tran, C.; Andrews, P.; Ayan, N. F.; Bhosale, S.; Edunov, S.; Fan, A.; Gao, C.; Goswami, V.; Guzmán, F.; Koehn, P.; Mourachko, A.; Ropers, C.; Saleem, S.; Schwenk, H.; and Wang, J. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. arXiv:2207.04672.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vincent, S.; Prescott, C.; Bayliss, C.; Oakley, C.; and Scarton, C. 2024a. A Case Study on Contextual Machine Translation in a Professional Scenario of Subtitling. arXiv:2407.00108.

Vincent, S.; Sumner, R.; Dowek, A.; Prescott, C.; Preston, E.; Bayliss, C.; Oakley, C.; and Scarton, C. 2024b. Reference-less Analysis of Context Specificity in Translation with Personalised Language Models. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 13769–13784. Torino, Italia: ELRA and ICCL.

Voita, E.; Sennrich, R.; and Titov, I. 2019. When a Good Translation is Wrong in Context: Context-Aware Machine Translation Improves on Deixis, Ellipsis, and Lexical Cohesion. In Korhonen, A.; Traum, D.; and Màrquez, L., eds., *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1198–1212. Florence, Italy: Association for Computational Linguistics.

Vu, H. H.; Kamigaito, H.; and Watanabe, T. 2024. Context-Aware Machine Translation with Source Coreference Explanation. *Transactions of the Association for Computational Linguistics*, 12: 856–874.

Wu, Y.; Schuster, M.; Chen, Z.; Le, Q. V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; Klingner, J.; Shah, A.; Johnson, M.; Liu, X.; Kaiser, L.; Gouws, S.; Kato, Y.; Kudo, T.; Kazawa, H.; Stevens, K.; Kurian, G.; Patil, N.; Wang, W.; Young, C.; Smith, J.; Riesa, J.; Rudnick, A.; Vinyals, O.; Corrado, G.; Hughes, M.; and Dean, J. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *CoRR*, abs/1609.08144.

Yao, B.; Jiang, M.; Yang, D.; and Hu, J. 2024. Benchmarking LLM-based Machine Translation on Cultural Awareness. arXiv:2305.14328.

Zhang, B.; Haddow, B.; and Birch, A. 2023. Prompting large language model for machine translation: a case study. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Zhang, Y.; Li, Y.; Cui, L.; Cai, D.; Liu, L.; Fu, T.; Huang, X.; Zhao, E.; Zhang, Y.; Chen, Y.; Wang, L.; Luu, A. T.; Bi, W.; Shi, F.; and Shi, S. 2023. Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. arXiv:2309.01219.