Testing the Effectiveness of Prompt Engineering Technique in the Translation of CSIs using LLMs: Naguib Mahfouz's *Midaq Alley (1947),* A Case in a Point

Noura Wael Aly Kamel

1. Introduction

Throughout the past two years, advancements in translation technology have transformed the way different languages are bridged paving the way for more researches about the quality of automatically translated texts. As a subfield of Artificial Intelligence (AI), Natural Language Processing (NLP) has advanced rapidly in the field of text generation and understanding context. Machine translation (MT) is one of the applications of NLP that has benefited from such advancements as it focuses on producing accurate fully automated translated texts (Wang et al., 2021, p.143). Although MT has witnessed significant technological progress recently, the improvement in the quality of the translated texts is obvious in scientific and technical texts, rather than the literary texts that are embedded with rhetorical devices and ambiguity (Cespedosa & Mitkov, 2023, p.48). Translation of CSIs is a highly challenging mission because they are deeply embedded in the historical, religious, and social contexts of the source language which they often lack direct equivalents in the target language. This is reflected in the inability of LLMs to efficiently translate literary texts. Cultural translation can be described as a process depending on the concept of cultural fax, where the translator's aim is to convey the meaning, style, and form of the source text along with its cultural references, so that the target audiences receive the same effect and experience of the cultural essence in a way that resembles the response of the source text audiences (Zhang, 2020, p.1). According to Nida and Taber (1969), translation of culture is the change in the content of the message to suit the receptor's culture along with adding some information to make the text clearer for the target audiences in some cases (p.199).

TANWĪR: A Journal of Arts and Humanities
Online ISSN: 3062-4789
Print ISSN: 3062-4797
https://tanwir.journals.ekb.eg/
November 2025, Issue (3)

Many technological advanced techniques are presented nowadays to address this issue. Prompt Engineering has recently emerged as a crucial technique to guide the LLMs to accurately perform the users' tasks. Through designing the prompts, researchers can influence the LLMs output to better align with translation goals such as preserving cultural nuances and ensuring contextual understanding. This paper aims to optimize the quality of the translation of CSIs from Arabic into English in Mahfouz's *Midaq Alley* (1947) using five prompt engineering methods. It compares the effectiveness of this advanced technique on both models; Falcon-H1-7B-Instruct and Aya-expanse-8B in light of Davies' (2003) strategies of translating culture-specific items and categorizing these items using Newmark's (1988) taxonomy.

1.1 Significance of the study

This paper bridges the gap between translation studies and the field of AI through its contribution in optimizing the translation of CSIs by automated systems without human intervention. Furthermore, it provides detailed analysis of two LLMs translations; Falcon-H1-7B-Instruct and Aya-expanse-8B. This helps in identifying the MT drawbacks in two different machines and how it can be addressed through the advanced technique, prompt engineering. The experiment attempts to both theoretical and technical domains by engaging translation theories into AI-driven perspectives trying to leverage the LLMs.

1.2 Objective of the study

The paper focuses on optimizing the LLMs performance in translating literary texts including CSIs, and reducing the hallucinations. It aims at determining the challenges faced by the automated systems to translate complex Arabic language in a literary context. Application of five prompt engineering methods takes place along with evaluating the efficiency of each in improving the accuracy, cultural adequacy, and the contextual relevance of the CSIs. Moreover, it identifies the translation strategies used by the LLM in each prompt engineering method, and which of them is the most accurate one to render the meaning.

1.3 Statement of Research Problem

The translation of CSIs has posed several challenges especially those presented in literary works by Naguib Mahfouz due to their embedded connotations and lack of direct equivalents across different languages. While human translators are completely aware of these cultural nuances to render such items effectively, automated systems are facing difficulty in the translation process resulting in inaccurate, literal, or culturally inappropriate renderings. An accurate translation of a CSI is essential to keep the integrity and flavor of the source language.

1.4 Research Questions

The paper hence attempts to answer the following questions:

1. What challenges do the Large Language Models (LLMs) encounter when translating CSIs from Arabic into English?
2. How do different prompt engineering methods contribute to the accurate translation of culture-specific items in literary texts?
3. Which prompt engineering method help yield a more accurate and culturally appropriate target product?
4. What are the translation strategies used by each of the two models before and after prompt engineering to render the CSIs?

2. Theoretical Frameworks

2.1 Large Language Models

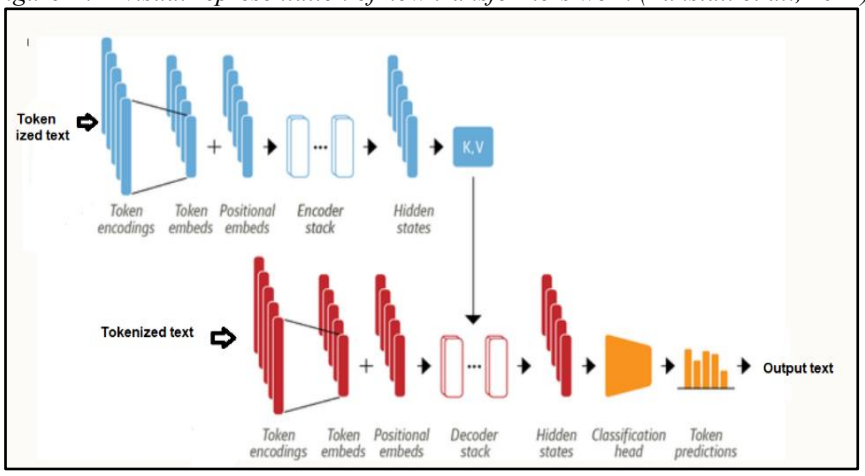
Large language models (LLMs) are deep learning systems trained on huge amount of data and developed to receive an input by a user, then produce human-like outputs to answer users' inquiries (Blank, 2023, p. 987). In their technique, LLMs depend on transformers, a type of neural network architecture that can process and generate meaningful texts through bypassing recurrence and depending on the attention mechanism as it do not read the words one by one, but it looks at all the words together and focus on the important ones to understand the contextual meaning of the provided text (Vaswani et al., 2017, p.2). LLMs are defined as generative decoder-only transformer models because of the autoregressive nature of language generation (Alammar & Grootendorst, 2024). They generate text sequentially through predicting each token based on the preceding one using a simpler architecture compared to encoder-decoder models, that have to process the entire input text (encoder) first and then generate the output (decoder) (Vaswani et al., 2017, p.5).

Beforehand, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) were the dominant approaches for processing both image and text data, respectively. However, RNN-based models struggled with many challenges, such as long sequential processing, and vanishing gradients (Vaswani et al., 2017, p.2). Bahdanau et al. (2015) have deployed the attention mechanism for neural machine translation that have benefited transformers to address such challenges. It depends on modelling dependencies among input tokens with no need for recurrence making it computationally efficient (Vaswani et al., 2017,

p.2). Transformers divide data into small units to process each separately allowing multiple calculations at the same time across different GPUs. This parallelization noticeably reduces training time and several batches can be processed at the same time (Vaswani et al., 2017, p.3).

As shown in Figure A, embeddings are adopted in sequence models to convert both input and output tokens into vectors. These embeddings are combined with positional encodings to capture the semantic meaning and the order of a single token which allow the model to accelerate the training compared to the traditional recurrent models. (Tunstall et al., 2022, p.2). The attention mechanism enables each token to focus on relevant parts of the sequence regardless the distance and capture contextual relevance (Vaswani et al., 2017, p.5-6). Positional encoding and attention mechanism allow transformers to process sequences efficiently which makes them effective for translation tasks.

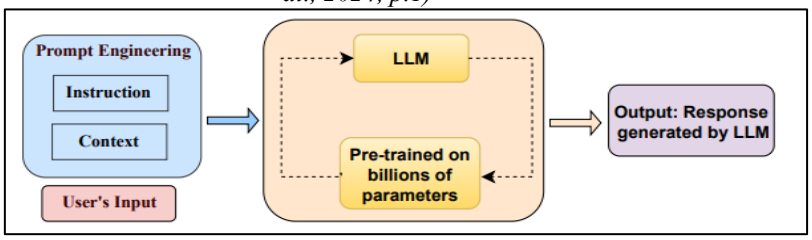
Figure A: A visual representation of how transformers work (Tunstall et al., 2022)



2.2 Prompt Engineering

Prompt engineering is a recent technique aims at enhancing the performance of LLMs through different methods. It focuses on designing the prompts to the LLM for more relevant and suitable results as shown in Figure B (El Amri, 2024). The significance of prompt engineering lies in its ability to enhance the adaptability of the LLMs without changing their parameters. Prompt engineering methods are divided according to their application areas; for example, performing new tasks, reasoning, and reducing hallucinations. (Sahoo et al., 2024, p.3)

Figure B: A visual representation of the components of the prompt engineering process in the LLMs (Sahoo et al., 2024, p.1)



Sahoo et al. (2024) list twelve methods of prompt engineering for different tasks (p.3). The main difference among prompt engineering methods is in task complexity depending on the pre-trained model knowledge, reasoning processes to generate reasoning paths in specific tasks and external dependencies such as python programs. Zero-shot prompting is used when no examples are provided allowing the LLM to predict the new class based on its prior knowledge, while one-shot prompting provides the LLM with a single example and few-shots is used by providing more than one example for the desired topic (Chen et al., 2023, p.6).

Chain-of-thought (CoT) refers to a series of natural language reasoning steps provided to the LLM to think step by step and arrive at the desirable information (Wei et al., 2022, p.1). It depends on some attractive properties to enhance reasoning in LLMs. As shown in Figure C, first, it let the model divides complex processes into intermediate steps making it easier for the model to get the accurate information. Second, it provides an interpretable window to explore the behavior of the model and identify how it can reach a particular output, providing insights into the error origin (Wei et al., 2022, p.2-3). Third, chain-of-thought reasoning techniques can be used in many tasks, such as arithmetic problems or tasks that require reasoning and understanding (Wei et al., 2022, p.2-3).

Figure C: An example of chain of thought prompt (Wei et al., 2022)

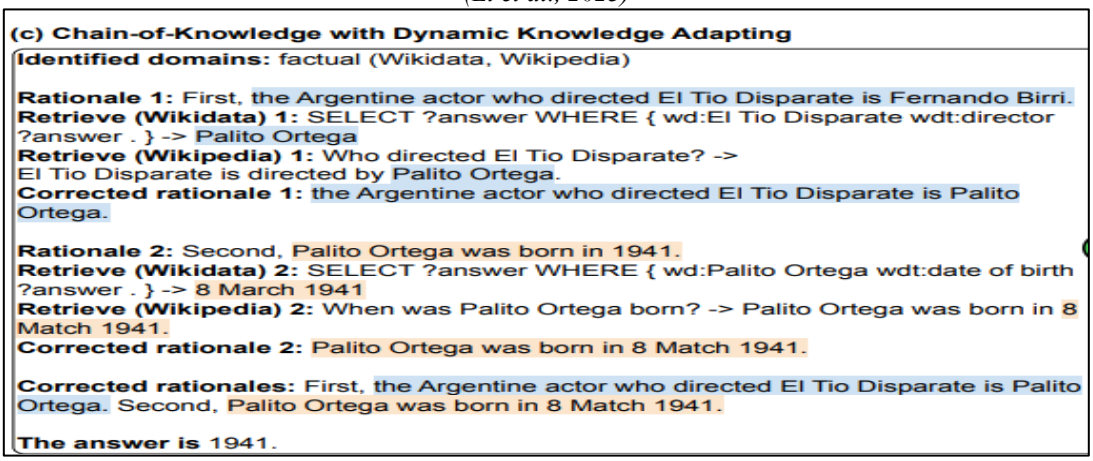
PROMPT FOR MATH WORD PROBLEMS

Q: There are 15 trees in the grove. Grove workers will plant trees in the grove today. After they are done, there will be 21 trees. How many trees did the grove workers plant today?

A: There are 15 trees originally. Then there were 21 trees after some more were planted. So there must have been 21 - 15 = 6. The answer is 6.

Chain-of-Knowledge (CoK) prompting method mitigates the limitations of Chain-of-Thought (CoT) by dividing complex tasks into structured knowledge and explanations to the LLM. This method, inspired by human cognitive process, allows language models to explicitly generate knowledge-based evidence as part of the reasoning process (Wang et al., 2023, p.1). As shown in Figure D, the method uses external knowledge and sources to facilitate the output generation process. This structured approach reduces the tendency of the model to hallucinate and provides clear, knowledge-based responses (Wang et al., 2023, p.2).

Figure D: An example of how CoK integrates different sources for knowledge such as Wikipedia and Wikidata (Li et al., 2023)



2.3 Newmark's (1988) Classification Model of Culture-specific Items

Newmark's (1988) Classification Model is adopted to classify the CSIs in Mahfouz's *Midaq Alley*. Peter Newmark (1988) classifies culture-specific items into five categories, "ecology", "material culture", "social culture", "organizations, customs, and ideas", and "gestures and habits". First, *ecology* refers to the environmental features in specific culture such as mountains, plants, and animals (p.96). Second, *material culture* refers to the tangible objects, artefacts, food, transportation, and types of houses (p. 97). Third, *social culture* refers to the activities in the communities in specific countries such as terms associated with work and leisure (p. 98-99).

The political, legal, historical, religious aspects fall under the category of organizations, customs and ideas, regarding the institutional names and titles, the names of administrative buildings, in addition to the religious and historical terms (Newmark, 1988, pp. 99-101). Finally, gestures and habits category reflects the culturally specific behavioral norms (Newmark, 1988, p. 102). Some items can be familiar in one culture and unfamiliar in another, they can pose challenges in translation.

2.4 Davies' (2003) strategies of translating CSIs

Translating culture-specific items pose various constrains due to their nuances and connotations which may differ from one language to another. Eirlys E. Davies (2003) introduces seven strategies to translate culture-specific items. The first is *preservation*; which refers to maintaining the source text in the translation when there is no equivalent in the TT (p. 75). For example, 'سمبوسة' is preserved as 'Sambusa' in English due to unique features in the Egyptian cuisine (Tenaijy & Al-Batineh, 2024, p.5). There are two types of preservation;

formal and *semantic*. The former signifies the transliteration of the CSI in the TT, while the latter refers to the literal translation of the item (Davies, 2003, p. 76).

The second strategy is *addition;* it involves providing additional explanatory information to mitigate the cultural gap, allowing the audience to understand the text (Davies, 2003, p. 77). For example, 'الصفصافة' is 'groves of willow' as a way of clarification for the target readers (Ali, 2024, p.87).

The third strategy is *omission;* it means that the translator omits a problematic CSI in the TT to avoid confusion (Davies, 2003, p. 79). For example, omitting titles as in ' التقى جلالة الاردن عبد الله الثاني بأخيه جلالة ملك المملكة العربية السعودية عبد الله بن عبد العزيزخادم الحرمين الشريفين' is translated into 'The King of Jordan met with his Saudi counterpart' as western audiences ignore such extra information in the ST (Alrumayh, 2021, p.2).

The fourth strategy is *globalization,* it indicates replacing the CSIs with more neutral and general ones to address readers of different backgrounds (Davies, 2003, p. 82). For example, "حصيرة", a specific type of rugs in the Arab world, becomes "mat" in English (Tenaijy & Al-Batineh, 2024, p.5).

The fifth strategy is *localization*; it focuses on using more familiar items in the TT, unlike globalization, allowing audiences to understand CSIs as closely as possible (Davies, 2003, p. 84). For example, a British CSI such as *Boiled and Roast Potatoes* becomes *Gratin* in French (Davies, 2003, p. 84).

The sixth strategy is *transformation*; it distorts the ST CSI resulting in significant changes in the context (Davies, 2003, p. 86). For example, the British title "*Harry Potter and the Philosopher's Stone*" becomes "*Harry Potter and the Sorcerer's Stone*" where the concept of philosopher's stone that may be obscure is replaced with a clearer title (Davies, 2003, p. 87).

The last strategy is *creation* which involves introducing a new CSI in the target language that does not exist in the source text (Davies, 2003, p. 88). This is clear in the creation of new names such as "*Anastasia*" becomes" نفيسة" in the Arabic dubbed version of *Cinderella* (R.M CARTOON TV, 2018).

2.5 Multidimensional Quality Metrics (MQM)

The Multidimensional Quality Metrics (MQM) framework is emerged in the first place to address the limitations of earlier models such as LISA QA, and SAE J2450 (Lommel et al., 2014, p.1). All of them are quality evaluation automation tools used to assess the translation

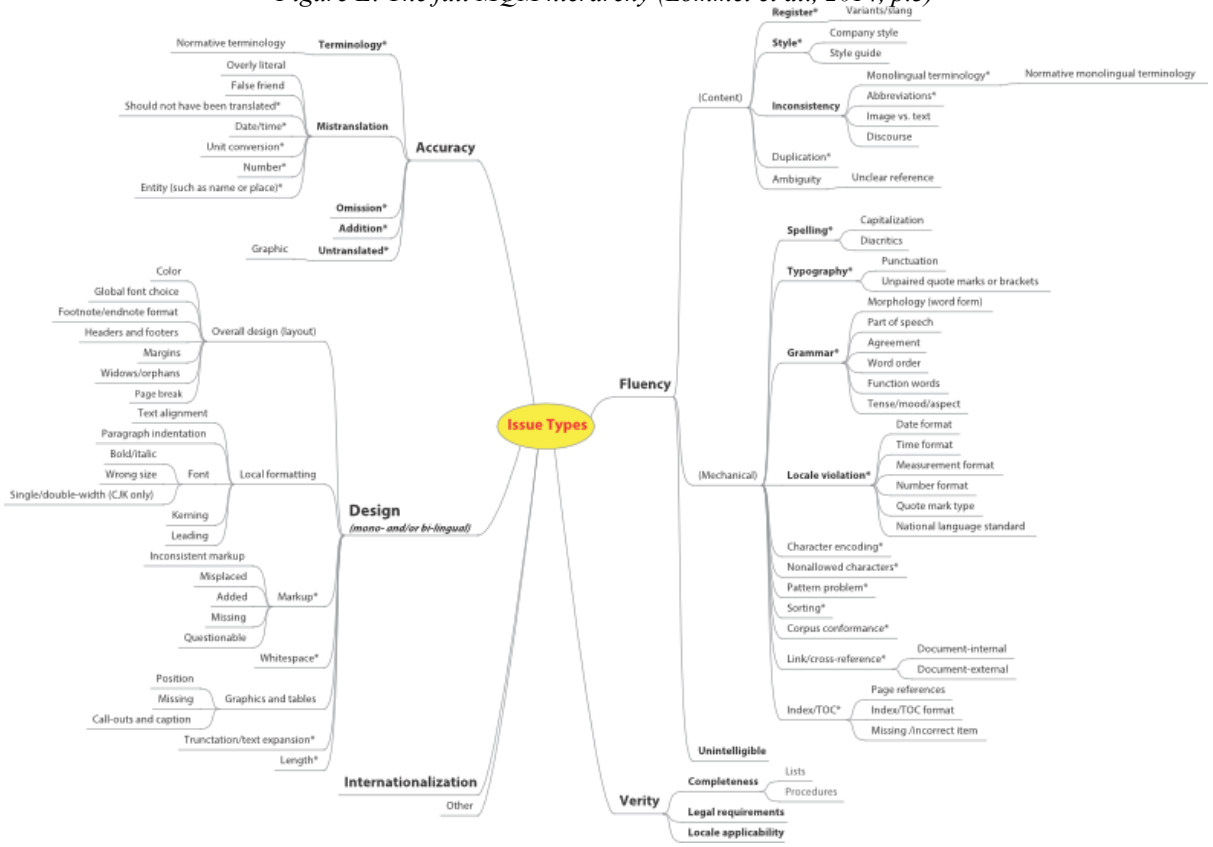
quality, however LISA QA, and SAE J2450 have some limitations making them less popular compared to the MQM. They have five main drawbacks. First, the one-size-fits-all limitation faces LISA QA. While it provides a general-purpose error list, it is not tailored for different translation contexts resulting in some categories that do not match the project needs. Second, the limited applicability of SAE J2450 as it is designed for evaluating automation systems making it unsuitable for other types of translation including the human one (Lommel et al., 2014, p.1-2). Third, these models are often customized by some organizations to meet their needs, which means that scores couldn't be compared to the ones from another customized version (Lommel et al., 2014, p.1-2). Fourth, the customization of models lead to the inconsistencies in evaluation methods among the providers. Finally, once a customized model is adopted in an organization, it persisted even if it has flaws because of being a standard practice (Lommel et al., 2014, p.1-2).

According to the MQM organization, this error typology provides a structured framework for classifying translation errors by both MT and Human Translation. It is divided into seven high-level error type dimensions with more specific error subtypes arranged hierarchically in each. This hierarchical structure allows the researchers to assess the quality of translation in terms of different details. The seven high-level MQM error dimensions are divided as follows:

Table 1: Seven High-level MQM Error Dimensions

Categories	Subtypes
Terminology	- Inconsistent use of terms - wrong terms
Accuracy	- Mistranslation - Over-translation - Under-translation - Addition - Omission - Untranslated
Linguistic Conventions	- Grammatical, spelling, punctuation errors - Other mechanical inaccuracies
Style	- inappropriate tone, or register
Locale Conventions	- Wrong date, time, or currency
Audience appropriateness	- Inappropriate terms for the TL audiences
Design and Markup	Wrong visual or technical presentation of the translation - Character - Paragraph formatting - Mark-up inconsistencies - Graphical elements - Page or interface layout issues

Figure E: The full MQM hierarchy (Lommel et al., 2014, p.3)



2.6 Automatic Evaluation Metrics

Automatic evaluation metrics are used in the automatic evaluation process along with the human analysis that is based on the MQM. These metrics are computational methods used to evaluate the quality of machine-translated text without human intervention. They are cost-effective and user-friendly, but they may lack the reliability of human evaluation (Lee et al., 2023, p. 2).

Many metrics have been introduced to assess translation outcomes; five of which are used in this paper; BLEU, METEOR, ROUGE, CHRF, and TER. BLEU stands for Bilingual Evaluation Understudy, and it measures the overlap of n-grams between the machine-translated text and the reference focusing on the precision (Lee et al., 2023, p. 4). It has some limitations as it does not consider the meaning penalizing very short translations and does not incorporate sentence structure (Lee et al., 2023, p. 4). METEOR stands for Metric for Evaluation of Translation with Explicit Ordering; it aims at addressing BLEU limitations by considering stems and synonyms to capture word variations (Lee et al., 2023, p. 6). Third, ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation; it measures the overlap of n-grams, word sequences, and pairs between the candidate and reference (Lin, 2004, p.1). It ranges between 0 and 1 with higher scores indicating the high similarity between the candidate and

the reference. Fourth, ChrF stands for Character n-gram F-score; it Calculates F-score (harmonic mean of precision and recall) at the character level rather than word level, so it captures partial matches (Lee et al., 2023, p.7). Finally, TER stands for Translation Edit Rate which measures the number of edits needed to change the candidate translation into the reference (Lee et al., 2023, p.5).

3. Review of Literature

In recent years, LLMs have significantly elevated the field of NLP by improving the efficiency of several tasks especially translation. There has been extensive research on advanced techniques over the past two years investigating their effectiveness in reducing LLMs hallucinations and improving the quality of their outputs.

Many experiments are applied to large language models testing the efficiency of prompt engineering in adjusting prompts to enhance translation quality. He (2024) investigates whether integrating translation-paper concepts such as "translation brief" and the personas of "translator" versus "author"—can improve ChatGPT translation quality. Through a lot of experiments on GPT-4, the paper finds that providing ChatGPT with translation-brief-style context (e.g., intended audience, publication medium) does not improve the translation. However, instructing the model to adopt the "translator" persona led to the best performance and it surpasses both the basic prompt and the "author" persona one according to both automatic metrics (BLEU and COMET-22) and human analyses. This suggests that traditional tools designed for translation projects may not be effective in human-machine prompting strategies, and that the more nuanced the prompt, the better enhancements are achieved for LLM-based translation tasks.

Researchers use different prompt engineering methods comparing various machine translation systems to evaluate the efficiency of each in improving the quality of translation. Wu and Hu (2023) test different methods for prompt engineering to enhance the performance of GPT- 3.5 in document-level machine translation especially in English-Chinese and Chinese-English texts in the WMT 2023 Competition. The experiment compares different prompt designs, multi-turn dialogues, temperature settings, and model variants (GPT-3.5-4k vs. GPT-3.5-16k) (p.166). Results and findings prove that multi-turn dialogue prompts, where multiple turns occur between the user and the model keeping track of context to provide desirable outputs, are successful in leveraging the context, improving the quality, and providing accurate outputs. Moreover, GPT 3.5 outperforms the 4 version, however, it struggles with the

ambiguous prompts (p.167-168). This underscores the importance of precise and accurate prompt design to optimize the translation outputs.

Lately, linguists and translation researchers have started to use other prompt engineering methods in an attempt to improve accuracy and adequacy. Tan et al. (2024) tackle the different performance of both zero-shot and few-shots learning in the machine translation task using LLMs. They attribute the significant differences to the writing styles rather than the semantic meaning. The researchers use two datasets: WMT'14 English-German and WMT'16 Romanian-English to apply three methods namely; zero-shot, one-shot, and style-learning (p.490). A shot refers to the number of examples provided to the LLM by the user, therefore zero-shot prompting depends on the LLMs knowledge with no external examples, while the one-shot method provides the model with only one example. The results suggest that the performance gap between zero-shot and few-shots can be simply narrowed by using the style-learning method. This highlights that the model performance significantly improves by adapting to the target text writing style in the few-shot (p.493-494).

Recent advances in prompt engineering techniques have gained prominence in various natural language processing tasks; however, their application to machine translation remains largely unexplored. Nguyen and Xu (2025) conduct a significant comparative analysis for using advanced prompt engineering methods. The study aims at evaluating how reasoning-based techniques affect the translation quality compared to the zero-shot prompting. Researchers use a diverse collection of datasets, including FLORES-200 for multilingual translation, WMT for domain adaptation, and MTNT for noisy texts on the commercial model (GPT-40 Mini) and the open-source model (Qwen 2.5 72B Turbo) for testing their translation. Then, automatic evaluation metrics such as SacreBleu and COMET are used to assess the translation quality. Findings suggest that using the ToT method excelled with complex and cultural items after the automatic and human evaluation. The paper suggests that advanced prompt engineering methods succeed in offering a perfect alternative for a model fine-tuning to improving translation systems.

While the previous studies have explored the application of prompt engineering on various LLMs, none have specifically investigated the improvement of LLMs performance in Arabic-English translation, specifically literary texts. This paper addresses the gap identified by the previous authors and contributes to the ongoing discussion of improving the Arabic-English translation of culture-specific items in two models; Falcon-H1-7B-Instruct and Aya-expanse-8B.

4. Methodology

This is an interdisciplinary comparative study that combines both fields; technology and translation. It goes through four main steps. The first is extracting suitable excerpts from *As-Sukkarīyah [Sugar Street]* (1957) to be used in the one-shot and few-shots methods as examples. The second is identifying a text full of CSIs from *Midaq Alley* (1947) to test the methods on the two LLMs. The third is the application of five methods on Falcon-H1-7B-Instruct and Aya-expanse-8B identifying the main drawbacks in translation. Finally, the five automatic evaluation metrics are used to calculate the efficiency of each method in both LLMs before and after prompt engineering along with using the MQM to identify the error typology of mistranslated items.

5. Data

5.1 Dataset

The paper uses Naguib Mahfouz's *al-Sukkarīyah [Sugar Street]* (1957) and its English translation (2006) by William Maynard Hutchins and Angele Botros Samaan (1992) in the one-shot and few-shots methods in prompt engineering. *Zuqāq al-Maddaq* [*Midaq Alley*], written by Naguib Mahfouz (1974), is used in the testing process to analyze all the CSIs translation based on Davies' (2003) strategies classifying them according to Newmark's (1988) Model. Mahfouz's works are considered a rich cultural representation describing the Egyptian society during the British occupation. This novel includes terms tied to the Islamic and social practices which pose challenges in translation to maintain authenticity.

5.2 Models

The experiment is applied to both Falcon-H1-7B-Instruct and Aya-expanse-8B. Falcon is a large language model produced by the Technology Innovative Institute (TII) center in UAE and trained on various Arabic data. Aya-expanse-8B is a pre-trained large language model based on transformer released in October 2024 supporting 23 languages, including English and Arabic. The number of parameters measure the complexity and capacity of the model, so the more the parameters, the more complex and sophisticated tasks the model can perform. The two LLMs are chosen based on two reasons. The first is the model architecture as both of them are in the 7-8B parameters range. The second reason is the multilingual specialization as both of them support several languages including the English and Arabic ones.

6. Analysis

This section presents the application of five prompt engineering methods; zero-shot, one-shot, few-shots, chain of though (CoT), and chain of knowledge (CoK) in Falcon-H1-7B-Instruct and Aya-expanse-8B comparing their effectiveness in optimizing CSIs translation in Naguib Mahfouz's *Midaq Alley (1947).* Excerpts from *As-Sukkarīyah [Sugar Street]* (1957) are used in both methods; one-shot, and few-shots to provide the LLMs with some examples needed. Finally, the outputs are evaluated using Davies' (2003) strategies of translating CSIs classifying them using Newmark's (1988) model 1988. The researcher then employs five automatic evaluation metrics. The errors are then classified according to the MQM error typology.

6.1 Prompt Engineering to Falcon-H1-7B-Instruct

The results in Table 2 display the automatic evaluation for the five prompt engineering methods using five different metrics.

Table 2: Evaluation of Translation across Five Prompt Engineering methods in Falcon

Prompting technique	SacreBleu	ROUGE (Fmeasure)	METEOR	CHRF	TER
Zero-shot	15.72	0.60	0.43	42.61	79.1
One-shot	16.85	0.59	**0.46**	49.7	81.01
Few-shots	16.80	**0.61**	0.44	**50.6**	78.4
CoT	14.37	0.59	0.44	38.5	82.28
CoK	**18.43**	0.55	0.41	39.15	**76.58**

The results in Table 2 show the improvement in the performance of Falcon by the CoK, Few-shots, and One-shot respectively. Moreover, it highlights the inefficiency of the CoT method in improving the translation quality comparing to the zero-shot (standard prompting). The inconsistency in the results is attributed to the different evaluation perspectives of each metric as they do not measure the translation quality in the same way. Each automatic metric captures different aspect of quality, semantic meaning, fluency, and lexical overlap. Metrics interpretation comes as follows;

SacreBleu

(Lexical overlap)

It measures how many words and phrases in the candidate text match those in the reference text. Therefore, it works on both word and sentence levels.

- Highest: CoK (18.43) scores the highest reflecting that it produces the closest words and phrases to the reference text. In terms of CSIs, it is the **best** by using almost the same items as the reference text.
- Lowest: The low score of CoT (14.37) indicates its lexical deviation from the reference text.

ROUGE

(Balance between precision and recall)

It focuses on meaning checking how much of the content in the reference appears in the candidate.

- Highest: Few-shots method (0.61) scores the highest f-score highlighting its success in covering the meaning regardless of the exact word choice. However, it might face difficulty in delivering the cultural meaning using specified well-known English equivalents as the reference.
- Lowest: The low score of the CoK (0.55) suggests content coverage loss, despite using closest items to the reference.

METEOR

(Semantic meaning)

It captures both lexical and semantic similarity considering synonyms and paraphrases.

- Highest: The one-shot (50.6) is the best in terms of lexical and semantic meaning compared to the zero-shot (0.43) followed by few-shots and CoT (0.44).
- Lowest: Despite scoring the best in SacreBLEU, CoK is not successful in using synonyms. It prioritizes the lexical overlap than the semantic one.

CHRF

(*Character n-gram F-score*)

It captures the lexical and morphological matches focusing on the word forms making it good for morphologically complex languages such as Arabic.

- Highest: Few-shots method is the best in getting partial matches followed by the one-shot.
- Lowest: The CoT low score suggest that it uses structural rewording resulting in morphological mismatches.

TER

(Edits)

It calculates the number of edits in the candidate to be like the reference. Such edits include; omissions, additions, substitutions, shifts, and word order.

- Lowest: The low score of CoK (76.58) refers to a **well-translated** text requiring the least number of edits to match the reference which matches its SacreBleu result followed by the few-shots.
- Highest: CoT (82.28) high score indicates the large number of edits required to match the reference text.

The absence of a leading method across the different metrics indicates that each method improves certain dimension of translation quality. All of these metrics measure the style as a whole; while regarding the translation of CSIs, researchers may look for the ones that consider word level evaluation to check whether the candidate uses the same or close items as the reference text or not. First, CoK is strong in SacreBleu and TER, but weak in METEOR, ROUGE, and CHRF implying that it succeeds in providing some similar and close items to the reference on the word level with fewer edits, but it loses some semantic richness in its literary style. Second, few-shots excels in ROUGE and CHRF referring to its ability to capture more content and semantic meaning resulting in morphological matches, though it is slightly below the CoK in SacreBleu, and TER, and below the one-shot in METEOR. This suggests the success of few-shots in almost all the metrics with very slight differences.

Third, one-shot ranks the first in METEOR suggesting the best semantic alignment and synonyms offering, though its lexical overlap in SacreBleu is lower than the CoK. Fourth, the chain of thought (CoT) is the consistent low score in every metric suggesting that Falcon does not perform well to the instructions and reasoning processes. Such results confirm that prompt engineering methods optimize different aspects in CSIs translation, metrics such as TER and SacreBleu refer to items closeness, METEOR and CHRF reward meaningful synonyms and paraphrases for the whole sentences and style, and ROUGE reward the complete content coverage. Figure F demonstrates the CoK prompt.

Figure F: CoK Prompt (Retrieved on August 12, 2025)

Follow these steps:

Step 1: CSI Detection (Rationale 1)

Carefully read the Arabic source text. Translate the text into English considering all **culture-specific items (CSIs)** that are potentially unfamiliar or untranslatable. These can include:

- Islamic phrases or invocations
- Names of people, places, neighborhoods
- Colloquial idioms
- Cultural practices or institutions
- Food, clothing, architecture, etc.

Step 2: Contextual Retrieval (Dynamic Knowledge Adapting)

For each detected CSI, consult external **cultural or literary sources** to understand its historical, cultural, or religious significance. Use:

- Quranic Arabic Corpus
- English translations of *Zuqaq al-Midaq* by Naguib Mahfouz
- Egyptian Arabic idiom dictionaries
- Arabic literature by Taha Hussein, Reem Bassiouney, Abbas El Akkad
- COCA, BNC

Step 3: Rationale Correction and Cultural Mapping

Decide the best **translation strategy** for each CSI using one of the following:

- **Retention with transliteration** (e.g., *Zuqaq al-Midaq*)
- **Literal translation** (if concept exists)
- **Cultural substitution** (if a close equivalent exists)
- **Footnoting or explicitation** (to preserve meaning)
- **Creative equivalence** (for idioms or expressions)

Step 4: Final Translation Output

Now, translate the entire Arabic text into fluent, literary English. Maintain:

- The **emotional and literary tone**
- The **symbolism, imagery, and narrative voice**
- The **cultural significance** of expressions and context

-Use transliteration for place names and proper nouns.

- If needed, include footnotes or brief explanations for CSIs that cannot be naturally translated.

Step 5: Translation Review & Refinement

- Check if the final translation feels **natural for English-speaking audiences**.
- Revise any awkward constructions while preserving original meaning.
- Ensure that **CSIs retain their cultural impact**.

Arabic source text:

آذنت الشمس بالمغيب، والتفَّ زقاق المدقِّ في غلالة سمراء من شفق الغروب، زاد من سُمرتها عمقًا أنَّه مُنحصرٌ بين جدران ثلاثة كالمصيدة، له باب على الصنادقيَّة، ثُمَّ يصعد صعودًا في غير انتظام، تحفُّ بجانبٍ منه دكَّان وقهوة وفرن، وتحفُّ بالجانب الآخر دُكَّان ووكالة، ثمَّ ينتهي سريعًا — كما انتهى مجده الغابر — ببيتَين مُتلاصِقين، يتكوَّن كلاهما من طوابق ثلاثة.

سكنت حياة النهار، وسرى دبيب حياة المساء. همسة هنا وهَمْهَمة هناك: يا ربُّ يا معين. يا رزَّاق يا كريم. حُسْن الختام يا ربُّ. كل شيء بأمره. مساء الخير يا جماعة. تَفَضَّلوا جاء وقت السمر. اصْحَ يا عم كامل وأغلق الدكَّان. غيِّرْ يا سنقر ماء الجوز. أطفئ الفرن يا جعدة. الفص كبَس على قلبي. إذا كنَّا نذوق أهوال الظلام والغارات منذ سنواتٍ خمس، فهذا من شرِّ أنفسنا

Table 3: Translation of Falcon-H1-7B-Instruct before and after prompt engineering and application of Davies' 2003 strategies of translating CSIs and Newmark (1988) taxonomy

The Term	Source Text	Target Text (Standard prompting)	Target Text (After CoK prompting)	Target Text (After few-shots prompting)	Type of CSI (Newmark 1988)	Standard Prompting Davies Strategy	CoK Prompting Davies Strategy	Few-shots Prompting Davies Strategy
CSI-1	زقاق المدق	***alley of Al-Madf***	***Zuqaq al-Midaq****	***Midaq Alley****	Social culture	Mistranslation	Formal preservation	Semantic preservation +

								Formal preservation
CSI-2	الصنادقية	***ground floor***	***ground floor***	***Sidewalk***	Social Culture	Mistranslation	Mistranslation	Mistranslation
CSI-3	دكان	***Shop****	***Shop****	***Shop****	Material Culture	Localization	Localization	Localization
CSI-4	وكالة	***agency***	***agency***	***agency***	Material Culture	Mistranslation	Mistranslation	Mistranslation
CSI-5	قهوة	***café***	***café***	***Coffeehouse****	Material Culture	Localization	Localization	Localization
CSI-6	فرن	***Bakery****	***Bakery****	***Bakery****	Material Culture	Localization	Localization	Localization
CSI-7	يارب يا معين	***Oh Lord, oh sustainer****	***O Lord, O sustainer****	***O Lord, O sustainer****	Customs and ideas	Semantic preservation	Semantic preservation	Semantic preservation
CSI-8	يا رزاق يا كريم	***Oh provider, oh generous one****	***O provider, O generous****	***O provider, O generous one****	Customs and ideas	Semantic preservation +Addition	Semantic preservation	Semantic preservation +Addition
CSI-9	يا جماعة	***Everyone****	***Everyone****	***everyone****	Social culture	Localization	Localization	Localization
CSI-10	كامل	***Khairallah***	***Kamil****	***Kamil****	Social culture	Mistranslation	Formal preservation	Formal preservation
CSI-11	عم	***----***	***Uncle****	***----***	Social culture	Omission	Localization	Omission
CSI-12	سنقر	***Sana'a***	***Senker****	***---***	Social culture	Mistranslation	Formal preservation	Omission
CSI-13	جعدة	***Ghawada***	***Jaada****	***Jaada****	Social culture	Mistranslation	Formal preservation	Formal preservation
CSI-14	حُسْن الختام يا ربُّ	***A good end, oh Lord****	***A good end, O Lord****	***A good end, O Lord.****	Customs and ideas	Semantic preservation	Semantic preservation	Semantic preservation
CSI-15	. كل شيء بأمره	***Everything is by Your will****	***Everything is by Your will****	***Everything according to His will***	Customs and ideas	Semantic preservation	Semantic preservation	Semantic preservation
CSI-16	الفص	***The cold***	***The news***	***The weight of the evening***	Customs and ideas	Mistranslation	Mistranslation	Mistranslation

"" refers to the accurate translations in the standard prompting and to the improved translations in the post prompting methods*

According to Table 3, the LLM provides more reliable and accurate translations for five CSIs after using the CoK prompting and three after using the few-shots method. First, CSI-1 "زقاق المدق" is rendered into "Zuqaq al-Midaq" using formal preservation by the CoK and into "Midaq Alley" using formal preservation in transliterating المدق into Midaq and semantic preservation in translating زقاق as Alley. According to *Cambridge Dictionary*, "Alley" refers to a narrow path located between the buildings. Moreover, since it is a name of a street, transliterating the name is acceptable. Therefore, both of translations are accurate and better than the mistranslation of standard prompting "alley of Al-Madf" due to the spelling mistake in Madq. This highlights the model confusion in recognizing the Arabic word المدق.

Consequently, Madf is unsuccessful on the mechanical level because of the MT hallucination according to the MQM.

Second, the LLM translated CSI-5 "قهوة" using localization as "café" through zero-shot, and CoK, and as "coffee house" by the few-shots methods. "Coffee house" is better than "café" as according to *Oxford Learner Dictionary*, "café", in English, refers to a small restaurant that sells light meals, desserts, and drinks. On the other hand, "coffee shop" refers to a place sells only drinks. In old Egyptian places, it is preferably translated as "coffee shop" where it is a small place in which people can drink tea, coffee, or other local drinks such as sahlab and it suitable for gatherings in local neighborhoods which suits the text and novel atmosphere. Third, CSI-10 and CSI-13 are rendered as "Kamil" and "Jaada" using formal preservation through both CoK and few-shots which is a successful translation compared to the standard prompting that rendered them as "Khairallah" and "Ghawada" because of MT hallucinations according to the MQM. Fourth, CSI-12 "سنقر" is transliterated as "senker" by the CoK using formal preservation, omitted by the few-shots method, and mistranslated as "Sana'a" by the standard prompting. The use of formal preservation to translate people's names is the best, rather than omitting it as the few-shots and misspelling it as the zero-shot as a mechanical error because of MT hallucination according to the MQM which results in losing the contextual meaning. This highlights the zero-shot difficulty in identifying and translating names.

Sixth, CSI-11 "عم" is omitted by both zero-shot and few-shots, however it is translated as "uncle" through CoK using localization. According to *Merriam Webster Dictionary*, "uncle" refers to someone who encourages and supports not only a family member so there is an improvement in rendering it as uncle rather than omitting them. Therefore, the translation quality improves after prompt engineering in some CSIs as it reduces the LLMs hallucinations and accurately renders a bigger number of the CSIs in the extract to the target-language audience.

On the other hand, the LLM hallucinates in translating some CSIs due to overly literal translation according to the MQM. First, CSI-4 "وكالة", the LLM translates it as "agency" and according to English version by Trevor Le Gassick, it is rendered as "office". According to *Cambridge Dictionary*, "office" means a room or a small part of a certain building used by some employees as a part of business, while "agency" refers to a government organization that includes a number of persons responsible for a huge business. In such context, "وكالة" is located in a local neighbourhood including a few persons of the working class; therefore, using office is a better option for translating "وكالة".

As per the MQM typology, CSI-2 "الصنداقية" translation as "ground floor" and "sidewalk" is wrong because of the ambiguous target content. According to *Al Ahram Gate*, this neighbourhood is called "حي الصنداقية" as it refers to the street of al-Muʿizz li-Dīn Allāh al-Fāṭimī in Cairo where there are a lot of shops selling bridal kits in boxes. Previously, Egyptian brides used to buy their personal supplies from these shops before the wedding. Therefore, such translations result in totally losing the meaning and confusing the target audiences. Furthermore, the LLM hallucinates in translating CSI-16 "القص" as "the cold", "the news", and "the weight of the evening" as they are totally irrelevant to the context. In Arabic, the term refers to the hashish people smoked to produce psychoactive effects such as relaxation.

6.2 Prompt Engineering to Aya-expanse-8B

The results in Table 4 display the automatic evaluation for the five prompt engineering methods to Aya-expanse-8B using five different metrics.

Table 4: Evaluation of Translation across Five Prompt Engineering methods in Aya

Prompting technique	SacreBleu	ROUGE (Fmeasure)	METEOR	CHRF	TER
Zero-shot	11.73	0.57	0.41	46.15	**74.7**
One-shot	**20.2**	0.60	**0.48**	**50.6**	**74.7**
Few-shots	15.17	0.58	0.41	45.75	77.8
CoT	16.09	0.56	0.42	46.51	77.2
CoK	13.95	**0.61**	0.43	47.8	75.9

The results in Table 4 presents the LLM success in providing more accurate translation according to four evaluation metrics out of the five despite measuring different aspects of translation quality. The one-shot scores the best results in all metrics except ROUGE, which suggest that it does not use different synonyms or paraphrases than the reference text, aligns with its low TER. Moreover, the zero-shot scores low numbers in all metrics, except in TER. Since the lowest the better in TER, zero-shot scores 74.7 like the one-shot which refers to some contradiction. As evaluation metrics work on the whole style, TER focuses on literal edits ignoring the meaning or the readability. Therefore, this suggest that the zero-shot is close to the reference in terms of literal edits, however, it is not right in semantic correctness, fluency, wording quality, or structure aligning with its low SacreBleu, METEOR, ROUGE, and CHRF. Figure G demonstrates the One-shot prompting and the translation process.

Figure G: One-Shot Prompt (Retrieved on August 13, 2025)

The translation of the following source text from Naguib Mahfouz Sugar Street (1957) is **Source text:**

ارتفعت عينا عائشة عن المجمرة إلى وجه أم حنفى لحظة ولكنها لم تعلّق بكلمة، قد علموا في حينه بهدم البيت الذي كان يومًا بيت السيد محمد رضوان ثم إعادة بنائه عمارة مكونة من أربعة أدوار باسم عم بيومي الشرياتلى، تلك الذكريات القديمة، مريم وياسين، ولكن تُرى أين مريم، وأم مريم وبيومى الشرياتلى الذي استولى على البيت بالوراثة والشراء، أيام كانت الحياة حياة، والقلب ناعم البال! وعادت أم حنفى تقول: أجمل ما فيها يا ستِ دكان عم بيومي الجديدة؛ تُريات ودندرمة وحلوى، كلها مرايا وكهرباء، والراديو ليل نهار، يا عيني على حسنين الحلَّاق ودرويش بائع الفول والفولى اللبَّان وأبو سريع صاحب المقلى وهم ينظرون من دكاكينهم البالية إلى دكان زميلهم القديم وعمارته.. فقالت أمينة، وهى تشبك الشال حول منكبيها: سبحان ربك الوهاب.. فعادت نعيمة تقول وهى تحيط عنق أمها بذراعيها: سدَّ جدار العمارة سطحنا من هذه الناحية، وإذا عمرت بالسكان فكيف نستطيع أن نُمضى الوقت فوق السطح؟ لم يكن فى وسع أمينة أن تتجاهل سؤالًا توجِّهه حفيدتها الجميلة مراعاةً لخاطر عائشة قبل كل شيء فقالت: لا يهمك السكان، امرحي كيف شِئت..

English Translation:

Aisha raised her eyes from the brazier to look at Umm Hanafi for a moment but made no comment. They had previously learned that the house once belonging to Mr. Muhammad Rid-wan would be torn down to allow construction of a four-story building for Uncle Bayumi the drinks vendor. This project had stirred up many old memories about Maryam and her divorce from Yasin what had become of Maryam? - and about Maryam's mother and her marriage to the drinks vendor Bayumi, who had gained possession of the house half by inheritance and half by purchase. Back then life had been worth living, and hearts had been carefree. Umm Hanafi continued: "The most beautiful part of it, my lady, is Uncle Bayumi's new place for soft drinks, ice cream, and sweets, tt has lots of mirrors and electric lights, with a radio playing day and night. I feel sorry for Hasanayn the barber, Darwish the bean seller, al-Fuli the milkman, and Abu Sari' with his snack shop. They have to look out of their dilapidated premises at the store and apartments of their former comrade." Pulling her shawl tighter around her shoulders, Amina said, "Glory to God who gives blessings…." With her arms around her mother's neck, Na'ima commented, "The building blocks off our roof on that side. Once it's inhabited, how can we spend any time up there?" Amina could not ignore the question raised by her beautiful granddaughter, if only out of concern for Aisha. She answered, "Pay no attention to the tenants. Do as you like."

Now, translate this text following the basic translation rules of literary texts

آذنت الشمس بالمغيب، والتفَّ زقاق المدقِّ في غلالة سمراء من شفق الغروب، زاد من سُمرتها عمقًا أنَّه مُنحصرٌ بين جدران ثلاثة كالمصيدة، له باب على الصنادقيَّة، ثُمَّ يصعد صعودًا في غير انتظام، تحفُّ بجانبٍ منه دكَّان وقهوة وفرن، وتحفُّ بالجانب الآخر دُكَّان ووكالة، ثمَّ ينتهي سريعًا — كما انتهى مجده الغابر — ببيتَين مُتلاصِقين، يتكوَّن كلاهما من طوابق ثلاثة.

سكنت حياة النهار، وسرى دبيب حياة المساء. همسة هنا وهَمْهَمة هناك: يا ربُّ يا معين. يا رزَّاق يا كريم. حُسْن الختام يا ربُّ. كل شيء بأمره. مساء الخير يا جماعة. تَفَضَّلوا جاء وقت السمر. اصْحَ يا عم كامل وأغلق الدكَّان. غيِّرْ يا سنقر ماء الجوز. أطفئ الفرن يا جعدة. الفص كبَس على قلبي. إذا كنَّا نذوق أهوال الظلام والغارات منذ سنواتٍ خمس، فهذا من شرِّ أنفسنا.

Table 5: Translation of Aya-expanse-8B before and after prompt engineering and application of Davies' 2003 strategies of translating CSIs and Newmark (1988) taxonomy

The Term	Source Text	Target Text (Standard prompting)	Target Text (After one-shot prompting)	Type of CSI (Newmark 1988)	Standard Prompting Davies Strategy	One-shot Prompting Davies Strategy
CSI-1	زقاق المدق	***Narrow alley***	**alleyway (Midaq Alley)***	Social culture	Semantic preservation + Addition	Semantic preservation + formal preservation
CSI-2	الصنادقية	**balcony**	**balcony**	Social Culture	Mistranslation	Mistranslation
CSI-3	دكان	***Shop****	***Shop****	Material Culture	Localization	Localization
CSI-4	وكالة	***Agency***	***Agency***	Material Culture	Mistranslation	Mistranslation
CSI-5	قهوة	***Coffee house****	***Coffee house****	Material Culture	Localization	Localization
CSI-6	فرن	***Bakery****	***Bakery****	Material Culture	Localization	Localization
CSI-7	يارب يا معين	**Oh Lord, help us***	**Oh Lord, grant us aid***	Customs and ideas	Semantic preservation +Transformation	Semantic preservation +Transformation
CSI-8	يا رزاق يا كريم	**Grant us sustenance, O Generous One**	**You are the Provider, the Generous***	Customs and ideas	Semantic preservation +Transformation	Semantic preservation
CSI-9	يا جماعة	**Everyone***	**Everyone***	Social culture	Localization	Localization

CSI-10	كامل	***Kamel****	***Kamil****	Social culture	Formal preservation	Formal preservation
CSI-11	عم	***Uncle****	***Uncle****	Social culture	Localization	Localization
CSI-12	سنقر	***Senker****	***Senker****	Social culture	Formal preservation	Formal preservation
CSI-13	جعدة	***Jada****	***Jaada****	Social culture	Formal preservation	Formal preservation
CSI-14	حُسْن الختام يا ربُّ	**Beauty in conclusion, O Lord**	**Grant us a good ending, Lord***	Customs and ideas	Mistranslation	Addition + Semantic preservation
CSI-15	كل شيء . بأمره	**All is in your hands***	**Everything is in Your hands***	Customs and ideas	Semantic preservation	Semantic preservation
CSI-16	الفص	***Grape***	***Grape***	Customs and ideas	Mistranslation	Mistranslation

"" refers to the accurate translations in the standard prompting and to the improved translations in the post prompting methods*

As shown in Table 5, the one-shot method proves successful in improving the translation of three CSIs. First, CSI-1 "زقاق المدق" is rendered as "narrow alley" using semantic preservation for زقاق and addition through the adjective "narrow" by the standard prompting. Adding another word to give the same meaning as "alley" is considered redundancy in translation as زقاق refers to the narrow path according to *Al Maaany Dictionary*, so there is no need to add "narrow". After applying the one-shot, the LLM translates it into "alleyway" along with "Midaq Alley" which is totally better. The term "alleyway" refers to a narrow passageway according to *Merriam-Webster Dictionary*, and further explaining it as "Midaq Alley" using formal preservation for the word المدق enhances the meaning without losing the ST flavour. Second, CSI-8 "يا رزاق يا كريم" is translated as "Grant us sustenance, O Generous One" using transformation for changing the word class in the adjective "يا رزاق" to a verb "Grant us sustenance" and semantic preservation in "يا كريم" to be rendered as "generous". Through the one-shot, the LLM translates "يا رزاق" as "the provider" which is more accurate than the verb + noun phrase "Grant us sustenance". According to *Collins Dictionary*, the term "sustenance" refers mainly to the food and drink required for living, however people do not ask God for only food and drink, instead they pray to have a peaceful life. Moreover, according to the *Quranic Arabic Corpus*, three translators; Mohsin Khan, Arberry, and Sahih International translated "رزاق" in sūrat l-dhāriyāt verse 58 as "the provider". Therefore, the one-shot translation succeeds in translating this Islamic invocation.

Third, CSI-14 "حسن الختام يارب" is mistranslated as "Beauty in conclusion, O Lord "by the standard prompting because of being overly literal according to the MQM. The use of terms such as "beauty" and "conclusion" is not accurate as "beauty" refers to the quality of being lovely usually given to a person or something according to *Merriam-Webster Dictionary*.

Furthermore, “conclusion” is the final part of something such as an essay or a novel as mentioned by *Cambridge Dictionary*. Through prompt engineering, the LLM then translates it as “Grant us a good ending, Lord” and the use of the verb “grant” in this context is accurate on the lexical level. In accordance with *Cambridge Dictionary*, “grant” means giving or allowing someone something. Moreover, in the Holy Quran, the verb “ارزقنا” is translated into “grant” by Shakir in translating the meaning of verse 114 in Sūrat l-māidah according to the translations provided by the *Quranic Arabic Corpus*. So, using the addition strategy along with semantic preservation to add the verb “grant” succeeds in clarifying the Islamic invocation to the target audiences. However, the model misinterprets the term “الختام” as according to *Al Maany Online Dictionary*, “حسن الختام” is translated as good, pleasant, or gracious end, not ending. “الختام” in *Hans-Wher Dictionary* is “end”, not ending and according to *Merriam-Webster Dictionary*, “end” is the cessation of a course of action, while “ending” is the conclusion of a movie or tv series. Therefore, translating it as “ending” is not accurate because it does not refer to the point at which one’s life ends.

On the other hand, the LLM hallucinates in translating three CSIs. First, CSI-2 “الصنداقية” is rendered into “balcony” by both the standard prompting and one-shot due to MT hallucinations according to the MQM. This results in meaning loss as the target audiences will not get it as a name of a neighbourhood. Second, as Falcon model, CSI-4 “وكالة” is mistranslated as “agency”, while this is not accurate because of being overly literal. Third, CSI-16 “الفص” is misinterpreted as “grape” by both methods; zero-shot and one-shot due to MT hallucinations as per the MQM error typology, while it refers to the hashish people usually smoke.

7. Findings and Discussion

7.1 Prompting Impact

The paper demonstrates a positive impact of the prompt engineering technique to improve the translation of culture-specific items through five different methods. While the zero-shot prompting fails in dealing with some cultural items such as; names, titles, Islamic invocations, and slang language in the testing text, the introduction of advanced methods enhances adequacy and fidelity. Particularly, the CoK and one-shot methods prove successful in enhancing the contextual understanding and the translation process in the two LLMs; Falcon-H1-7B-Instruct and Aya-expanse-8B respectively. The dominance of these two methods supports the argument that LLMs can be directed towards culturally sensitive texts whether by using example-based methods such as the one-shot and the few-shots, or the advance reasoning

methods as CoT and CoK. Consequently, these results underscore that prompt engineering is not just a technological technique, but it is a powerful methodology to optimize the LLMs outputs.

7.2 LLMs Comparison

The numerical results demonstrate the effectiveness of Aya-expanse-8B compared to Falcon-H1-7B-Instruct. Although Falcon LLM is produced by a UAE company, trained on several Arabic datasets, it shows more difficulty in understanding cultural nuances resulting in low scores in almost all the automatic metrics. The dominance of Aya-expanse-8B may be attributed to the number of parameters, 8B compared to the 7B of Falcon. Moreover, such results highlights the conformity to instructions in both LLMs, referring to the strong responsiveness of Aya to instructions than Falcon.

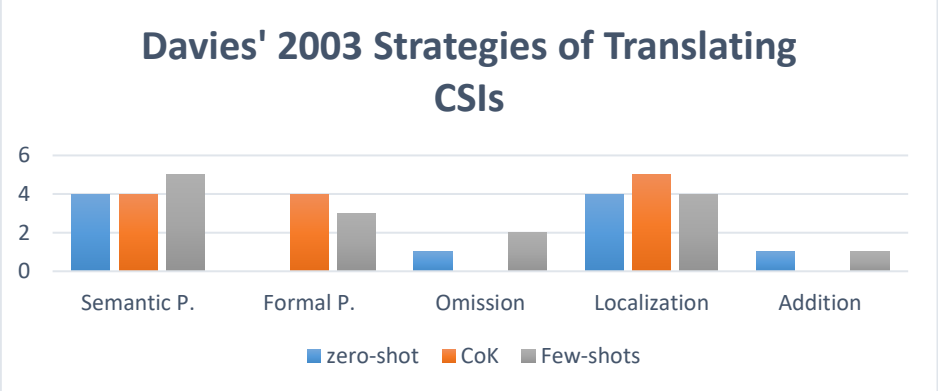
7.2.1 Falcon-H1-7B-Instruct

Table 6: Accurate and inaccurate translations of CSIs by Falcon-H1-7B-Instruct before and after prompt engineering

	No of accurate translations	No of inaccurate translations
Before PE	7	9
After CoK	12	4
After Few-shots	10	6

As shown in Table 6, the translation of Falcon-H1-7B-Instruct is improved by 5 CSIs, through the CoK and by 3 CSIs through few-shots. As a result, the number of inaccurate translations after applying the prompt engineering techniques is fewer than before their application. This suggests the success of prompt engineering techniques in improving the performance of the model.

Figure H: Strategies used by the model in zero-shot, CoK, few-shots prompting



According to Figure H, the model depends on using semantic preservation and localization in most of the prompt engineering methods. However, the CoK employs

localization 5 times instead of 4 like the standard prompting using it more accurately than in zero-shot method which contributes to rendering CSIs to the target audiences. Zero-shot and few-shots are the only methods to use the omission strategy resulting in ignoring the names, while the CoK uses the formal preservation 5 times to translate names of people and places. Furthermore, there is no show of the creation, globalization, and transformation strategies in the three methods. Therefore, the number of mistranslated items by the zero-shot method is 7 and then it is reduced to 3 by both the CoK and few-shots methods.

7.2.2 Aya-expanse-8B

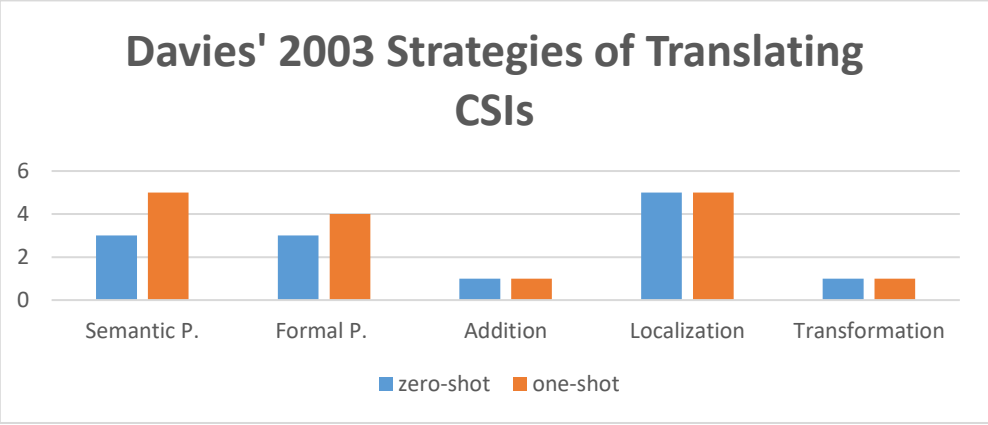
Table 7: Accurate and inaccurate translations of CSIs by Aya-expanse-8B before and after prompt engineering

	No of accurate translations	No of inaccurate translations
Before PE	10	6
After one-shot	13	3

As shown in Table 7, the translation of Aya-expanse-8B is improved by 3 CSIs, through the one-shot method. As a result, the number of inaccurate translations after applying the prompt engineering techniques is reduced to 3 instead of 6 before the application. This proves the success of prompt engineering technique in improving the performance of the model.

According to Figure I, the one-shot method adopts the semantic preservation strategy to deliver the CSIs references more than the zero-shot resulting in successful rendering. Moreover, it uses the formal preservation to transliterate names more than the standard prompting. However, both of them use the same number of addition, localization, and transformation. Figure I also highlights the absence of creation and globalization strategy. Therefore, the number of mistranslations in the zero-shot is reduced to 3 by applying the one-shot method instead of 4.

Figure I: Strategies used by the model in zero-shot, and one-shot prompting



Regarding the two LLMs, the researcher finds that both techniques prove successful in improving the models' hallucinations. In case of Falcon, the model, through CoK and one-shot methods, improve in terms of social culture items including; names such as'كامل', 'سنقر', 'جعدة', titles as 'عم', and places names such as 'زقاق المدق' using semantic and preservation. Moreover, it attains desired results in translating material culture items as exemplified by the item 'قهوة' using localization. Regarding Aya, the model achieves success in enhancing the translation of names of places such as 'زقاق المدق', and the Islamic invocations 'يا رزاق يا كريم', and 'حسن الختام يارب' using semantic preservation.

7.3 Implications and Limitations

The paper contributes to the translation studies, technology and AI fields when designing well-structured prompts to achieve culturally appropriate translations. Such prompts can be adjusted and further used in different translation tasks to test the efficiency in other genres. Moreover, researchers can compare prompt engineering technique with other advanced ones; fine-tuning, and RAG.

Despite the valuable insights into the role of prompt engineering technique in translation tasks, the paper is subject to two limitations. First, the testing data is limited in scope focusing on certain number and type of CSIs that may not cover the complexity of cultural nuances across different languages. Second, the evaluation is employed on two LLMs range from 7B to 8B parameters, therefore the results should not be generalized to other models or versions. Further researches are needed to comprehensively test other datasets of different genres and language to ensure an accurate understanding of using the prompt engineering technique in translation tasks.

8. Conclusion

The paper suggests that prompt engineering plays a crucial role in optimizing the LLMs performance in translating CSIs. Through the comprehensive comparison among zero-shot, one-shot, few-shots, CoT, and CoK across the two LLMs, the researcher finds that the accurately designed prompts do not only enhance the translation on the lexical level, but also in terms of the semantic meaning and cultural fidelity. The improvement comes in different types of CSIs; material culture, social culture, and, customs and traditions items in both LLMs. This is proved through the use of five different evaluation metrics measuring different aspects of translation quality. Each metric plays a role in evaluating certain aspect of the translation quality which is clear in the numerical findings. Moreover, the integration of human evaluation

provides a clear analysis for the accuracy of the automatic evaluation and to what extend researchers can depend on. This contribution highlights how Natural language processing (NLP) can bridge the gap between linguistic equivalence and cultural nuance.

To further enhance the translation of CSIs, prompt engineering different methods can be applied to several LLMs with various parameters to test the effectiveness of each method based on the model knowledge and responsiveness to such advanced technological technique. Well-designed prompts should include culturally relevant expressions and contextual details about the provided text to obtain the desired output. Different text genres; legal, scientific, political, technical, and economic can be tested by varied prompts to decide which of them best conveys the meaning.

Bibliography

Alammar, J., & Grootendorst, M. (2024). *Hands-On Large Language Models*. O'Reilly Media, Inc. Retrieved from https://www.oreilly.com/library/view/hands-on-large-language-models/

Ali, M. S. B. (2024). Culture-specific items and their translation from Arabic into English in Abdulaziz al-Maqaleh's poems. Elect*ronic Journal of University of Aden for Humanity and Social Sciences*, 5(1), 82–93. https://doi.org/10.47372/ejua-hs.2024.1.343

Almaany. (n.d.). *Almaany online dictionary*. Almaany. Retrieved August 15, 2025, from https://www.almaany.com/ar/dict/ar-en/

Alrumayh, A. (2021). Translation by omission and translation by addition in English-Arabic translation with reference to consumer-oriented Texts. *International Journal of Comparative Literature and Translation Studies*, *9*(1), 1. https://doi.org/10.7575/aiac.ijclts.v.9n.1p.1

Amri, A. E. (2024). *LLM Prompt Engineering for Developers*. O'Reilly Online Learning. https://learning.oreilly.com/library/view/llm-prompt-engineering/9781836201731/text/all_chapter_contents.xhtml#whats-next

Blank, I. A. (2023). What are large language models supposed to model? *Trends in Cognitive Sciences*, *27*(11), 987–989. https://doi.org/10.1016/j.tics.2023.08.006

Cambridge Dictionary | English Dictionary, Translations & Thesaurus. (2025). https://dictionary.cambridge.org/

Cespedosa, A. I., Rus, & Mitkov, R. (2023). Machine Translation of literary texts: genres, times and systems. In Lancaster University, *Proceedings of the First NLP4TIA Workshop* (pp. 48–53) [Conference-proceeding]. Proceedings of the First NLP4TIA Workshop. https://aclanthology.org/2023.nlp4tia-1.7.pdf#:~:text=1%20Rationale%20Recent%20advances%20in%20Artificial%20Intelligence%20and

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023, October 23). *Unleashing the potential of prompt engineering in Large Language Models: a comprehensive review*. arXiv.org. https://arxiv.org/abs/2310.14735

Collins English Dictionary. (n.d.). Collins English Dictionary online. Retrieved August 15, 2025, from https://www.collinsdictionary.com/dictionary/english

Davies, E. E. (2003). A Goblin or a Dirty Nose? The Treatment of Culture-Specific References in Translations of the Harry Potter Books. *The Translator*, *9*(1), 65–100. https://doi.org/10.1080/13556509.2003.10799146

He, S. & School of Culture and Communication, Swansea University. (2024). Prompting ChatGPT for translation: A Comparative analysis of translation brief and persona prompts [Journal-article]. https://aclanthology.org/2013.tc-1.6.pdf

Ismail, F. (2011, October 26). Alṣndāqyh. *Al Ahram Gate*. https://gate.ahram.org.eg/News/131052.aspx

Lee, S., Lee, J., Moon, H., Park, C., Seo, J., Eo, S., Koo, S., & Lim, H. (2023). A Survey on Evaluation Metrics for Machine Translation. In *Mathematics* (Vol. 11, p. 1006). https://doi.org/10.3390/math11041006

Li, X., Zhao, R., Chia, Y. K., Ding, B., Joty, S., Poria, S., & Bing, L. (2023, May 22). Chain-of-Knowledge: Grounding Large Language Models via Dynamic Knowledge Adapting over Heterogeneous Sources. arXiv.org. https://arxiv.org/abs/2305.13269

Lin, C.-Y., Information Sciences Institute, & University of Southern California. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In *Information Sciences Institute* [Journal-article]. https://aclanthology.org/W04-1013.pdf

Lommel, A., Uszkoreit, H., Burchardt, A., German Research Center for Artificial Intelligence (DFKI), & Language Technology Lab. (2014). Multidimensional Quality Metrics (MQM): a framework for declaring and describing translation quality metrics [Journal-

article]. Revista Tradumàtica: Tecnologies De La Traducció, Número 12, Traducció i qualitat, 455–463.

https://ddd.uab.cat/pub/tradumatica/tradumatica_a2014n12/tradumatica_a2014n12p455.pdf

Mahfouz, N. (2022). *As-Sukkarīyah [Sugar Street]*. Hindawi Foundation. https://www.hindawi.org/books/18636847/1/

Mahfouz, N. (2022). *Zuqāq al-Midaq [Midaq Alley]*. Hindawi Foundation. https://www.hindawi.org/books/62575295/

Mahfuẓ, N. (1966). *Midaq Alley* (T. Le Gassick, Trans.). The American University in Cairo Press.

Merriam-Webster. (n.d.). Dictionary by Merriam-Webster. In *Merriam-Webster*. https://www.merriam-webster.com/

Newmark, P. (1988). *A Textbook of Translation*. Prentice Hall.

Nguyen, L., & Xu, Y. (2025). Reasoning for Translation: Comparative Analysis of Chain-of-Thought and Tree-of-Thought Prompting for LLM Translation. ACL Anthology, 259–275. https://doi.org/10.18653/v1/2025.acl-srw.17

Nida, E. A., & Taber, C. R. (1969). *The theory and practice of translation* (Vol. 8). Brill Archive.

Oxford Learner's Dictionaries | Find definitions, translations, and grammar explanations at Oxford Learner's Dictionaries. (n.d.). https://www.oxfordlearnersdictionaries.com/

R.M CARTOON TV. (2018, March 11). Cinderella 1 arabic full movie [Video]. YouTube. https://www.youtube.com/watch?v=_dt3PuFbBlc

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2024, February 5). *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications*. arXiv.org. https://arxiv.org/abs/2402.07927

Tan, W., 12, Xu, H., 64, Shen, L., 30, Li, S. S., 136, Kenton Murray, Philipp Koehn, Benjamin Van Durme, Yunmo Chen, & Johns Hopkins University. (2024). Narrowing the Gap between Zero- and Few-shot Machine Translation by Matching Styles. In *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 490–502). https://aclanthology.org/2024.findings-naacl.33.pdf

Tenaijy, M. A., & Al-Batineh, M. (2024). Translating Emirati literature: exploring culture-specific items in Mohammed Al Murr's Dubai Tales. *Humanities and Social Sciences Communications*, 11(1). https://doi.org/10.1057/s41599-023-02555-4

The MQM error typology. (n.d.). https://themqm.org/error-types-2/typology/

The Quranic Arabic Corpus - word by word grammar, syntax and morphology of the Holy Quran. (n.d.-b). https://corpus.quran.com/

Tunstall, L., Von Werra, L., & Wolf, T. (2022). *Natural Language Processing with Transformers, Revised edition*. "O'Reilly Media, Inc." https://landing.deepset.ai/hubfs/Ebooks/oreilly_chapter_excerpt_nlpt.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017b). Attention is All you Need. *arXiv* (Cornell University), 30, 5998–6008. https://arxiv.org/pdf/1706.03762v5

Wang, H., Wu, H., He, Z., Huang, L., & Church, K. W. (2021). Progress in Machine Translation. In *Engineering* (Vol. 18, pp. 143–153). https://doi.org/10.1016/j.eng.2021.03.023

Wang, J., Sun, Q., Li, X., & Gao, M. (2023, June 10). *Boosting Language Models Reasoning with Chain-of-Knowledge Prompting*. arXiv.org. https://arxiv.org/abs/2306.06427

Wehr, H., & Cowan, J. M. (1980). A dictionary of modern written Arabic : Arabic-English. In *Librairie du Liban eBooks* (Issue 1). http://ci.nii.ac.jp/ncid/BA00168454

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022, 28 January). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv.org. https://arxiv.org/abs/2201.11903

Wu, Y., Lan-Bridge, & Hu, G. (2023). Exploring Prompt Engineering with GPT Language Models for Document-Level Machine Translation: Insights and Findings. In Association for Computational Linguistics, *Proceedings of the Eighth Conference on Machine Translation (WMT)* (pp. 166–169). https://aclanthology.org/2023.wmt-1.15.pdf

Zhang, Y. (2020). A Study of Cultural Translation from the Perspective of Cultural Fax. *OALib*, *07*(06), 1–12. https://doi.org/10.4236/oalib.1106450