UNIVERSITY OF MASSACHUSETTS AMHERST

HONOR THESIS

Translating Literature with LLMs: Maintaining Consistency on Named Entities And Integrating External Knowledge

Author:

Jiarui LIU

Supervisor:

Mohit Nagaraja IYYER

in the

Manning College of Information & Computer Sciences

Commonwealth Honors College

November 29, 2024

Committee in Charge

Committee Chair: Mohit Nagaraja IYYER

Committee Member: James ALLAN

Committee Member: Benjamin MARLIN

Project Advisor: Marzena KARPINSKA

“*Translation is the circulatory system of the world’s literatures.*”

Susan Sontag

UNIVERSITY OF MASSACHUSETTS AMHERST

Abstract

Manning College of Information & Computer Sciences

BS-CS Computer Science(BS)

Translating Literature with LLMs: Maintaining Consistency on Named Entities And Integrating External Knowledge

by Jiarui LIU

In literary translation, the accurate and consistent translation of named entities, such as character names and significant proper nouns, is essential. Recent advancements in Large Language Models (LLMs) have shown promise in translation tasks; however, they often struggle to maintain consistency over long texts due to limitations in context windows and computational budgets.

This study introduces an efficient approach that combines the advanced capabilities of LLMs with human translation practices using In-Context Learning (ICL) and Retrieval-Augmented Generation (RAG). We integrating Termbase, pre-editing methods, and external knowledge sources such as Wikipedia into LLMs and our findings demonstrate that this hybrid method significantly improves both the consistency of named entities and the overall quality of translation, offering a practical solution for literature machine translation.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my committee chair, Mohit Nagaraja IYYER and the project advisor Marzena KARPINSKA for their invaluable guidance, continuous support, and encouragement throughout my honor thesis project. Their expertise and insightful feedback have greatly contributed to the success of this thesis.

I am also deeply thankful to my committee members, James ALLAN and Benjamin MARLIN, for their constructive criticism and suggestions, which have been instrumental in refining my work. Their commitment and willingness to share their knowledge have been truly inspiring.

I extend my sincere thanks to the Manning College of Information & Computer Sciences and the Commonwealth Honors College at the University of Massachusetts Amherst. The resources, facilities, and academic environment provided by the institution have been essential to my research.

Finally, I would like to express my heartfelt appreciation to my parents for their unwavering support and encouragement. Their love, patience, and sacrifices have been the foundation upon which I have built my academic journey. Without their belief in me, this achievement would not have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

As an avid reader with a voracious appetite for stories, my nightly ritual has always been immersing myself in the captivating pages of a novel for about an hour before sleep. This habit transcends mere pastime; it represents a journey through diverse cultures and perspectives. The literary world, vast and ever-expanding, teems with an array of creative works, each offering a unique linguistic landscape. In the case of languages in which my proficiency is limited, I find myself reliant on translated versions of these works. Regrettably, as one might expect, a majority of these compelling narratives remain beyond my reach due to the scarcity of available translations, especially for new books, web novels, and fan fiction.

This challenge has led me to think of machine translation as an indispensable tool used in my reading. The emergence of Large Language Models (LLMs) and their capability in machine translation (MT) are astonishing, unlocking doors to literary realms once inaccessible. The translations with LLMs, particularly GPT-based models though not free from errors, have always impressed me with their high quality and near-native fluency. There are, however, still noticeable inaccuracies and inconsistencies, which serve as a stark reminder of the persistent gap between the finesse of human translation and the capabilities of machine translation.

My academic pursuits in Machine Learning (ML) and Natural Language Processing (NLP), coupled with an in-depth exploration of LLMs, have ignited a profound interest and a compelling thought: What if I could play a role in narrowing this divide? Imagine the possibility of enhancing LLMs to produce translations that are not only coherent and accurate but also capable of capturing the subtleties and beauty inherent in original texts.

1.2 Necessity of Machine Translation

In our increasingly interconnected world, the ability to traverse language barriers is paramount. The global exchange of ideas and knowledge heavily relies on our capacity to communicate effectively across diverse linguistic landscapes. Translation, therefore, is not just a tool but a vital conduit for this cross-cultural exchange. Traditionally, this role has been held by human translators, whose skills and expertise have been instrumental in bridging linguistic and cultural divides. However, the sheer volume and variety of contemporary global communications necessitate a translation approach that surpasses the capabilities of traditional methods. This burgeoning need calls for a translation solution that is not just effective, but also salable, adaptable, and capable of handling a multitude of contexts and languages.

Machine Translation, led by the advances in LLM technology, stands at the cusp of this evolution. It offers a cost-effective, rapid, and multilingual pathway to high-quality translation, forging connections between diverse cultures and languages. This technology does not seek to replace the human element in translation but rather to augment and expand our ability to share and understand the rich tapestry of human narratives.

1.3 Topic Focus

While numerous existing studies concentrate on investigating LLM' sentence-level translation ability solely (He et al., 2023; Ghazvininejad et al., 2023; Mu et al., 2023), in the thesis, we are more interested in exploring the capability of LLMs to translate long-form document-level text(text longer than model's text window size), particularly the literature, connecting with modern human workflow and retrieval-augmented generation (RAG) method.

Long-form text translation brings new challenges and barriers including limited length of text windows, inadequate utilization of context, inconsistent translation of Terminology, different translation from the real world, etc. Furthermore, as mentioned in (Karpinska and Iyyer, 2023), translating works of literature presents distinct difficulties because of the complex characteristics of creative pieces and the necessity to accurately convey the author's tone and contextual subtleties with the balance of faithfulness, expressiveness, and elegance.

Our objective is to comprehensively understand the emerging challenges and problems in long-text translation, conduct thorough analysis and evaluations, and strive to overcome these challenges by **combining Large Language Models with contemporary human translation/localization practices** and connecting with external knowledge base.

During human translation, modern CAT tools like Trados and OmegaT are equipped with translation memories and termbases to help translate in a consistent way. Humans also use the Internet and terminology sheets to speed up translation as well as retain high quality. The integration involves:

1. **leverage efficient setup in human translation practices like Termbase and Translation memory.**

2. **Connect with external resources such as Internet dictionaries and search engines.**

Additionally, we intend to investigate the types of demonstrations that are most effective and identify precise methods through which they can improve translation quality.

Chapter 2

Work Of Previous Researchers

2.1 Language Models And Large language Models (LLMs)

A language model (LM) is probabilistic machine learning model of a natural language trained to conduct a probability distribution over words. LM do nothing but predict the most likely word in an utterance position based on the context provided. A significant breakthrough in the development of language models was the introduced of the transformer architecture (Vaswani et al., 2023) which utilizes a self-attention mechanism.

The attention mechanism, originally introduced by (Bahdanau et al., 2016) for machine translation in recurrent neural network (RNN) to address long-range dependencies issues, allows RNN models to dynamically focus on specific parts of input text when generating output. Self-attention extends this idea by empowering the model to weigh the importance of different words in the input sequence relative to each other. This mechanism enables the model to capture long-range dependencies, process calculations in parallel, and understand context efficiently.

In the self-attention mechanism, each word (token) in the sequence is transformed into three vectors: query (Q), key (K), and value (V). The query vector of a word is compared against the key vectors of other words to compute attention weights, which are then used to calculate a weighted sum of the value vectors as result using the Attention function:

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2.1}$$

where $\frac{1}{\sqrt{d_k}}$ is a scaling factor, and the softmax function normalizes the weights to fall between 0 and 1, ensuring that they sum to 1.

In the transformer architecture, instead of computing a single set of self-attention weight, multi-head attention is employed to computes multiple sets of Q, K, V in parallel, enabling the model to focus on different linguistic properties of the input sequence. The results are then concatenated and linearly transformed. The Transformer consists of multiple layers of such multi-head attention and feed-forward neural networks. This design allows the model to capture complex patterns in the data efficiently.

Large language models (LLMs), such as GPT models, are language models trained on large-scale datasets with numerous internal parameters, most of which leverage the Transformer architecture. LLMs are notable for their ability to achieve general-purpose language understanding and generation. A single LLM can perform a wide array of tasks such as question answering, sentiment analysis, and translation in multiple languages, without any task-specific training.

Figure 2.2 illustrates a user interacting with an LLM using ChatGPT-4o. First, the user asks, "What is the capital of Australia?" (The correct answer is Canberra). Next, the user requests a sentiment analysis on the review: "Fresh ingredients. Large portions. Attentive staff. Highly recommend!" (The correct sentiment is Positive). Finally, the user asks the model to translate the sentence "Lolita, light of my life, fire of my loins. My sin, my soul." from English to Chinese. The model successfully completes all these tasks accurately.

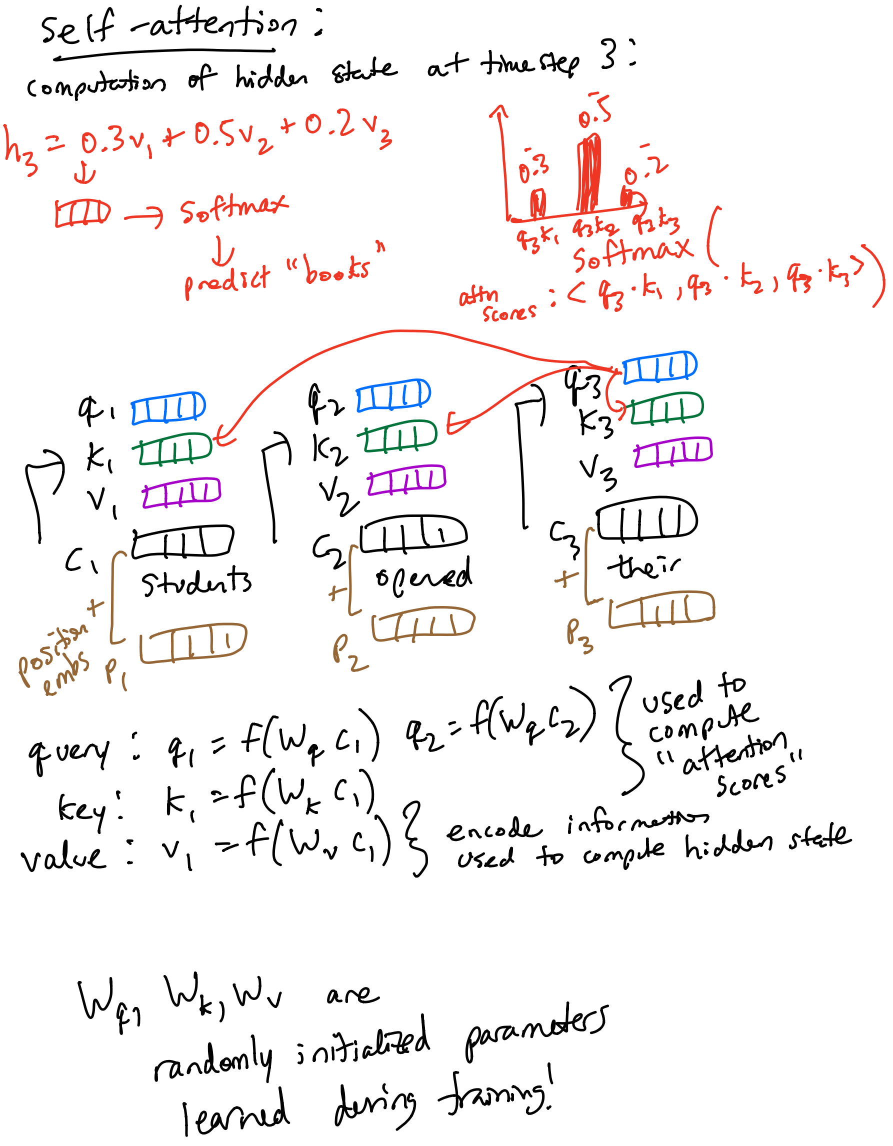
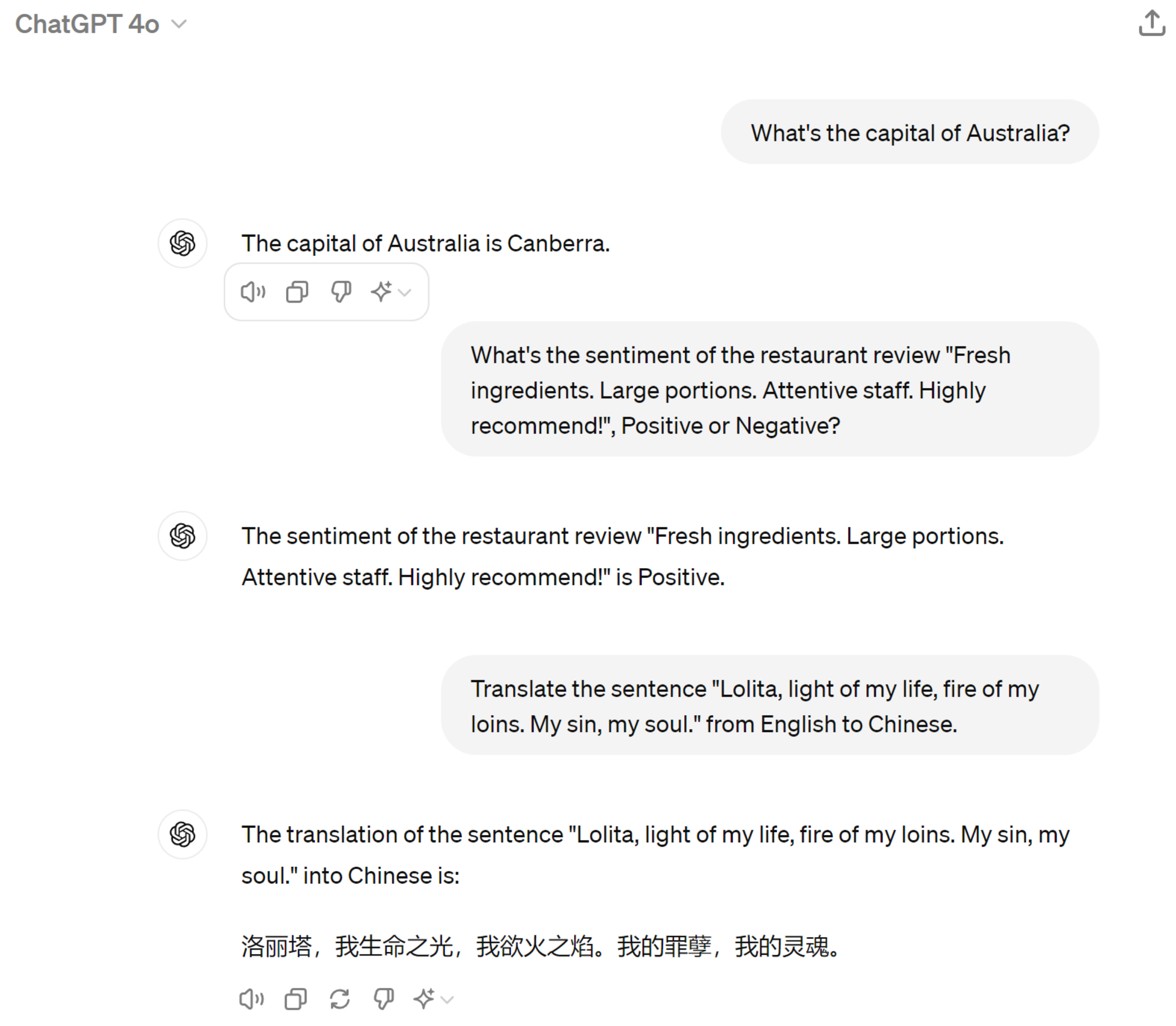

Self-attention:
computation of hidden state at time step 3:
h3 = 0.3v1 + 0.5v2 + 0.2v3
softmax
predict "books"
0.3
0.5
0.2
q3k1 q3k2 q3k3
softmax(
attn scores: < q3·k1, q3·k2, q3·k3 >)
q1
k1
v1
c1
students
position embs
p1
q2
k2
v2
c2
opened
p2
q3
k3
v3
c3
their
p3
query: q1 = f(Wq c1) q2 = f(Wq c2)
used to compute "attention scores"
key: k1 = f(Wk c1)
value: v1 = f(Wv c1)
encode information used to compute hidden state
Wq, Wk, Wv are randomly initialized parameters learned during training!

FIGURE 2.1: The self-attention mechanism from cs689's note by Professor Mohit Iyyer.



What's the capital of Australia?

The capital of Australia is Canberra.

What's the sentiment of the restaurant review "Fresh ingredients. Large portions. Attentive staff. Highly recommend!", Positive or Negative?

The sentiment of the restaurant review "Fresh ingredients. Large portions. Attentive staff. Highly recommend!" is Positive.

Translate the sentence "Lolita, light of my life, fire of my loins. My sin, my soul." from English to Chinese.

The translation of the sentence "Lolita, light of my life, fire of my loins. My sin, my soul." into Chinese is:

洛丽塔，我生命之光，我欲火之焰。我的罪孽，我的灵魂。

FIGURE 2.2: LLM can perform different tasks without any task-specific training.

2.2 Machine Translation As The Downstream Task Of LLMs

Focusing on the machine translation task, Large Language Models (LLMs) trained on massive unlabeled datasets like GPT and Llama demonstrate impressive abilities via simple prompting, eliminating the need for extensive collections of parallel data required by traditional translation models. Just ask LLMs to "Translate the below text from [Source Language] to [Target Language]", and the task is accomplished (Figure 2.2). Additionally, LLMs can enhance their performance through In-Context Learning (ICL), which involves conditioning on provided labeled examples without the need for parameter optimization.

The use of GPT models yields remarkable results in generating context-aware and coherent Translation, especially for high-resource languages and sometimes surpassing commercial models (Vilar et al., 2023; Hendy et al., 2023; Chowdhery et al., 2022).

2.3 Document-level Machine Translation Using LLM

Moreover, LLMs are believed to effectively leverage document-level context and produce fewer errors and inconsistencies while performing document-level translation. (Zhang et al., 2023) and (Karpinska and Iyyer, 2023) show LLMs are able to produce highly fluent and competitive translations and produce even better translations when translating complete paragraphs instead of single sentence. This indicates LLMs can leverage paragraph-level context information during translating.

Using human evaluation, (Karpinska and Iyyer, 2023) found that in literary translation, paragraph-level translation made by GPT-3.5 (GPT-3.5 PARA) is favored by human annotator compared with sentence-by-sentence translation made by GPT-3.5 (GPT-3.5 SENT) and translation produced by google translation. This study also demonstrates that when given paragraph-level context, LLMs can generate more coherent and enjoyable output with fewer mistranslations and grammatical issues. For example, GPT-3.5 PARA excels at assigns appropriate pronouns when translating from language such as Japanese, which often refers to the listener/speaker in the third person rather than second-person pronoun (i.e., you).

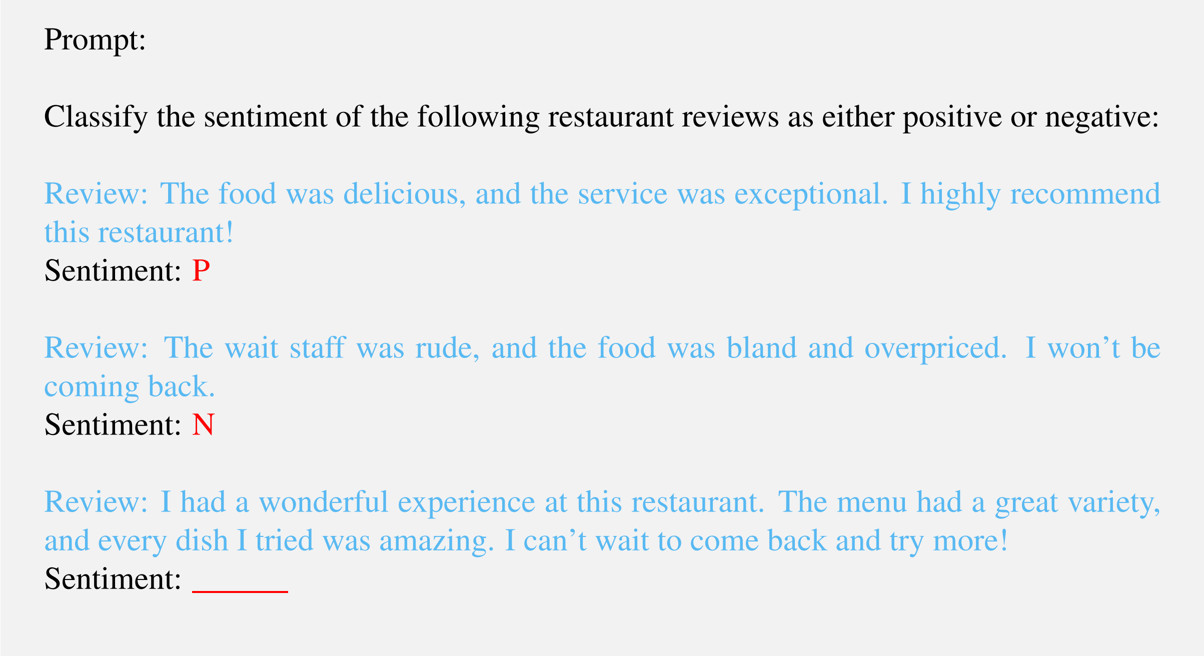
2.4 In-Context Learning (ICL)

In-context learning (ICL), first introduced in the GPT-3 paper by (Brown et al., 2020), is an approach that allows language models (LMs) to learn and perform task without the need for explicit fine-tuning or task-specific training. This is achieved by providing the model with a small list of input-output pairs as demonstrations before the actually task input during inference time.

The overarching idea behind in-context learning is learning by analogy or pattern finding. (Dong et al., 2023) To correctly answer the task question, the model needs to learn the input distribution, output distribution, input-output mapping and the formatting from the given demonstrations. (Xie et al., 2022) Meanwhile, (Min et al., 2022) found in-context learning still words well even when outputs in the demonstration are replaced randomly. This suggests that the model was likely been exposed to similar tasks in its vast training set and can leverage other demonstration components, such as input distribution and format to infer the correct answer.

One of the key advantages of in-context learning is that ICL dramatically simplifies the process of incorporating human knowledge into language models. (Liu et al., 2021; Wu et al., 2023) As the demonstration is written in natural language, user can guide the model's behavior and adapt it for specific tasks without resource-intensive fine-tuning. The provided flexibility enables quick experimentation and adaptation for down stream tasks, which is particularly useful in the case of large language models with hundred of billions of parameters, where self fine-tuning is nearly impossible due to computational constraint.

The prompt in Figure 2.3 start with an instruction to classify the sentiment of restaurant



Prompt:

Classify the sentiment of the following restaurant reviews as either positive or negative:

Review: The food was delicious, and the service was exceptional. I highly recommend this restaurant!
Sentiment: P

Review: The wait staff was rude, and the food was bland and overpriced. I won't be coming back.
Sentiment: N

Review: I had a wonderful experience at this restaurant. The menu had a great variety, and every dish I tried was amazing. I can't wait to come back and try more!
Sentiment: ______

FIGURE 2.3: A example of in-context learning prompt.

reviews as either P (positive) or N (negative). Following the instruction, there are two example reviews along with their sentiment labels serving as demonstration for the model to learn the task. The last restaurant review is the input-only task. When the prompt is fed to language model, the model needs to predict its sentiment based on provided instruction and demonstration.

2.5 Sentence-level Demonstrations And Translation Memory insertion

(Mu et al., 2023) propose Translation Memory Prompting for LLMs which prompt similar sentences from Translation Memory as demonstrations.

(Vilar et al., 2023) investigate a variety strategies for choosing translation examples at sentence level for few-shot prompting. They compared the k-nearest neighbors (kNN) and random selection strategies on two corpus: one is the full dataset with higher possibility of getting low quality example, the other is the choice high quality part of the former, concluding that the quality of examples is the most critical factor.

(Agrawal et al., 2022) conducted an analysis on the influence of the selection and quantity of few-shot in-context examples of overlap sentence on output translation quality, finding task-level prompts can complement example-specific prompts and kNN-MT baseline. Furthermore, To yield enhanced demonstrations, they proposed a BM25-like approach to better retrieval of similar sentence examples from the corpus.

2.6 Dictionary-based Prompting For Machine Translation

(Ghazvininejad et al., 2023) introduces Dictionary-based Prompting for Machine Translation (DIPMT) that leverages bilingual dictionaries to provide phrase-level knowledge in prompts during translation with LLMs. DIPMT appends possible translations for specific input words in the prompt which can improve translation quality particularly in low-resource and out-of-domain scenarios where LLMs often struggle due to limited relevant training data.

2.7 Termbase And Terminology Management

Termbase (TB) and terminology management are essential components of Computer-Assisted Translation (CAT) tools used in human translation practice. A termbase is a database that stores and manages terminology specific to the project, which stores terms, definitions, context, and other relevant information.

When the human translator encounters a term in the source text that could be found in TB, the tool could automatically provide with the corresponding target translation, reducing time and effort while maintaining translation consistency.

Termbase allow human translators ensure their translation adhere to domain-specific requirements, maintain consistency throughout the project, and improve the overall quality.

2.8 Retrieval-Augmented Generation (RAG)

LLMs have demonstrated impressive abilities in Natural Language Processing (NLP) and Natural Language Understanding (NLU) tasks, drawing on vast amounts of data to acquire in-depth knowledge. However, despite their capabilities, these models have certain limitations. They cannot dynamically expand their knowledge base, lack transparency in explaining their predictions, and are prone to producing "hallucinations"(Guerreiro et al., 2023), where the generated content might be unrelated or factually incorrect. (Lewis et al., 2021; Marcus, 2020)

The Retrieval-Augmented Generation (RAG) approach aims to address these limitations by combining the strengths of LLMs with external data sources. RAG integrates a retrieval system with a sequence-to-sequence (seq2seq) model, which allows the model to access and incorporate relevant external information.

For translation, Some researchers also propose to retrieval relevant samples from external datasets which can further provide additional information that are not contained in the training corpus. (Zheng et al., 2021; Cai et al., 2021)

2.8.1 Formal Explanation Of RAG

Text generation tasks can be conceptualized as a mapping from input text to output sequences:

$$y = f(x)$$

A typical sequence-to-sequence language model, for example, is represented as:

$$\begin{aligned} p_\theta(y|x) &= p_\theta(y_1, y_2, \ldots, y_N|x) \\ &= \prod_{n=1}^{N} p_\theta(y_n|x, y_{1:n-1}) \end{aligned}$$

Retrieval Augmented Generation (RAG) introduces a novel approach that enhances models with external memory (knowledge) through information retrieval (IR), so that they can acquire more information during prediction. The idea of such framework is retrieved relevant documents Z is beneficial for the model. This approach can be mathematically expressed as:

$$y = f(x, z)$$

Where $z \in Z$ represents a set of documents retrieved based on the query dependencies of x from a retrieval model π. We can model the sequence-to-sequence framework as a mixture model:

$$\begin{aligned} p_{\theta,\pi}(y|x) &= \sum_{z \in Z} p_\pi(z|x) p_\theta(y|x,z) \\ &\approx \sum_{z \in \text{top-k}(p_\pi(|x))} p_\pi(z|x) p_\theta(y|x,z) \\ &= \sum_{z \in \text{top-k}(p_\pi(|x))} p_\pi(z|x) \prod_{n=1}^{N} p_\theta(y_n|x,z,y_{1:n-1}) \end{aligned}$$

Consider the "retrieval model" π as a combination of U distinct queries, each query $q_u \in Q$ carrying a weight w_u dependent on the input x. Each query returns the top M documents $d_{u1}, d_{u2}, \ldots, d_{uM}$ with corresponding weights (ranks); thus, the model yields a list of documents D: $d_{11}, \ldots, d_{1M}, \ldots, d_{UM}$ along with their normalized weights $w_{11}, \ldots, w_{1M}, \ldots, w_{UM}$.

This process can be interpreted as a random variable with probabilities (normalized weights), leading to the following model formulation:

$$\begin{aligned} p_{\theta,\pi}(y|x) &= \sum_{d \in D} w_d p_\theta(y|x,z) \\ &= p_{\theta,\pi}(y|x) = \sum_{d \in D} p_\pi(d|x) p_\theta(y|x,z) \end{aligned}$$

For simplicity, consider each query returning only the top result, functioning as a mapping from input x to a single document v. The weight of each query q_u is computed as $w_u = Q(q_u) \cdot K(x)$, resembling the attention mechanism in transformers (Vaswani et al., 2023) with Q, K, V.

Different M values can be set for each query, provided that normalization is maintained. Specifically, if $M = 1$, the query function essentially maps input x to result v. Any model that can be formulated as such a function, including question-answering models or another LLM, could be used here.

A helpful figure from LlamaIndex (Figure 2.4) may aid in understanding this process better.

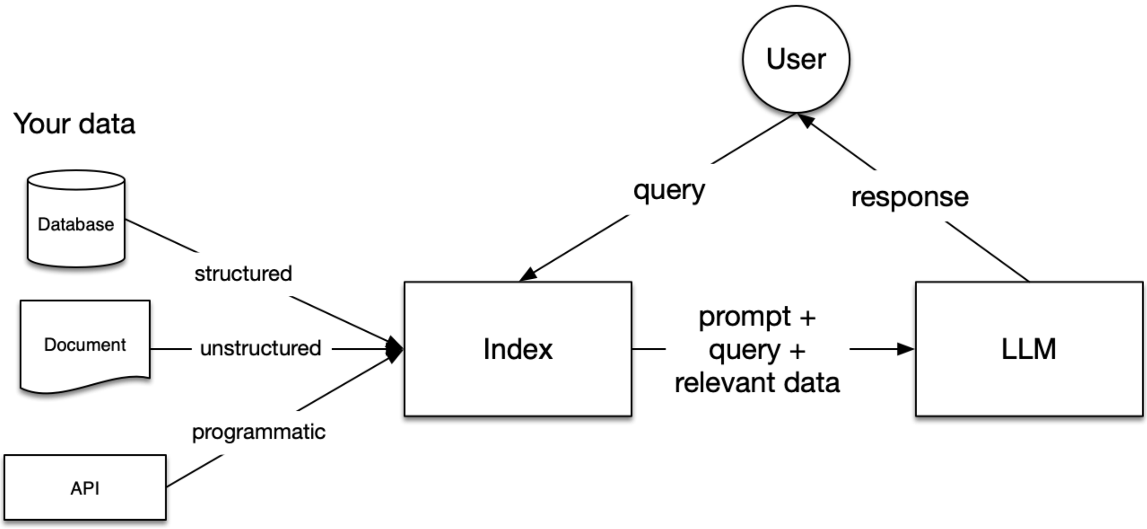


FIGURE 2.4: RAG explanation from LlamaIndex

Chapter 3

Methodology

3.1 Issue On Named Entity Translation

3.1.1 Translating Named Entities Necessitates Different Approaches

Named entities, such as proper nouns, main character names, and locations, present a significant challenge in document-level translation.

Translating named entities often necessitates different approaches and methods for translation models(Babych and Hartley, 2003): Some require retrieve official transactions from knowledge base, while others may presuppose subtle intension meanings. While LLMs can leverage their vast training data and knowledge base to address some of these challenges, there are still many scenarios where they fall short.

A notable example is the challenge of character selection when translating from alphabetic languages into logographic languages, such as Chinese or Japanese. In these cases, the same name can have multiple reasonable corresponding characters in the target language, each carrying different meanings, connotations or even encoding gender information. For example, the name "Alex" could be transliterated into Chinese using characters that emphasize masculinity, femininity, or neutrality, depending on the context or the translator's intent.

Another example is name ordering in different languages. Chinese and Japanese names follow a different convention compared to English. In English, names typically appear in the [*given name*] [*family name*] format, whereas in Chinese and Japanese, the [*family name*] precedes the [*given name*]. This difference requires specific attention in translation.

TABLE 3.1: Conventions of Name Order in Different Environments

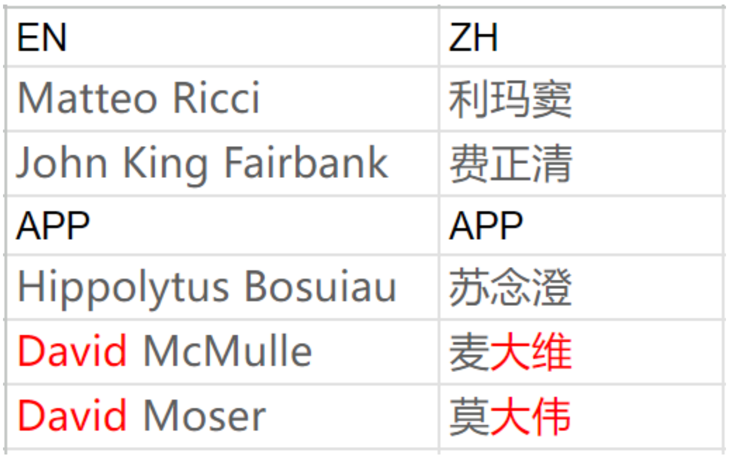
	English Environment	**Chinese Environment**
Chinese Name	[Given Name] [Family Name] Dong Li	[Family Name] [Given Name] Li Dong
Western Name	[Given Name] [Family Name] Marilyn Monroe	[Given Name] [Family Name] Marilyn Monroe

When written within an English context, both Chinese/Japanese and English names follow the *[Given name] [Family name]* format. However, during the translation process from English to Chinese/Japanese, the order of Chinese/Japanese names is reverted to the original *[Family name] [Given name]* format. English names, on the other hand, retain their original *[Given Name] [Family Name]* order in the translated text. Such conventions can lead to confusion for models in practices, especially when names from different cultures are presented together in a text. For instance, "Dong Li and Marilyn Monroe (in English)" would translate to "Li Dong and Marilyn Monroe(in Chinese)". (Table 3.1)

Machine-generated translations can conflict with official translations, particularly in the context of person names, historical events, and organizational names. The translation of person names usually follows either *phonological translation* (translating the name based on its pronunciation) or *transliteration* (converting the name from one writing system to another).

However, some individuals may choose to establish distinct names in other languages or create their own translated names. For example (Figure 3.1), the sinologist John King Fairbank's Chinese name is Fei4 Zheng4qing1 and the Jesuit missionary Hippolytus Bosuiau adopted the Chinese name Su1 Nian4cheng2, neither of them translated by phonological translation nor transliteration. The "David" in David McMulle and David Moser also translated differently in Chinese. While LLMs usually can handle famous person name itself, the model fail to synchronize with the official translation in many cases. During our tests, GPT-4 and

Claude 3 Opus successfully identified and translated John King Fairbank but failed on Hippolytus Bosuiau. (Figure A.1)



EN	ZH
Matteo Ricci	利玛窦
John King Fairbank	费正清
APP	APP
Hippolytus Bosuiau	苏念澄
David McMulle	麦大维
David Moser	莫大伟

FIGURE 3.1: Special cases of Named-Entities Translation

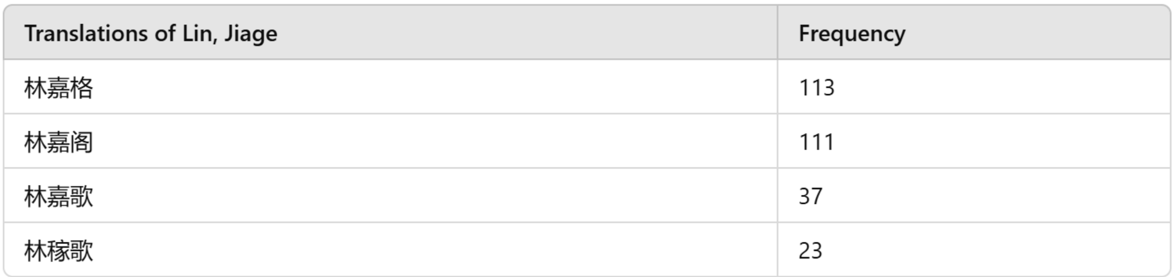
Certain entities, such as proper nouns or abbreviations, should remain untranslated as per convention but may inadvertently translated by machine translation models. This can lead to confusion or inconsistencies in the translated text. For instance, abbreviations like "APP" (application) or "FBI" (Federal Bureau of Investigation) should typically remain in their original form, as they are widely recognized and understood.

Another challenge emerges with fictional entities, which often require different translations depending on the domain or style. These terms may lack standardized translations, leading to inconsistencies when interpreted across various settings. The same name might have multiple translations depending on the genre (e.g., fantasy, science fiction), environment and intended audience.

Furthermore, anti-languages, cant and argots—specialized forms of language used by particular groups—pose additional difficulties for machine translation models. These terms, as they carry meanings or intentions not immediately evident from their surface structure, often rely on external knowledge or cultural understanding for specific groups to translate accurately.

3.1.2 Inconsistency Of Named-Entities Translation

Due to limitations on a model's output context window[1] and computational budgets, long-form texts like literature can hardly be processed by models within a single inference. The most commonly used practical method is to chunk the text into smaller segments and translate each segment individually. However, when a translation model runs with a positive temperature setting (introduces variability in its responses), it could produce different translations for the same entity across different segments–even though each translation might individually make sense. Depending on the context of a particular name instance, LLM models map the same name to different characters in different chunk runs of a same book. (Table 3.2)



Translations of Lin, Jiage	Frequency
林嘉格	113
林嘉阁	111
林嘉歌	37
林稼歌	23

TABLE 3.2: Four different reasonable translations Qwen 2.5 mapped the name "Lin, Jiage" to

Inconsistencies in named-entity translation can lead to confusion, break immersion, and disrupt the flow of the narrative. While mistranslating or omitting named entities may not significantly impact the BLEU or COMET score in automatic evaluations, it can greatly diminish human readability and the overall quality of the translation, particularly in the case of literature. This underscores the importance of developing models that can accurately handle named entities to ensure coherent and reliable translations.

[1] For example, although GPT-4 has a 128k input-token capacity, it only has a 4096 output context window, meaning it can output around 4k tokens at most for a single prompt.

3.2 Our proposed methods

3.2.1 Hypothesis: Wikipedia As Dictionary for Named Entity Translation

When translating, humans often rely on internet searches and relevant data sources to find accepted translations for specific terms. Similarly, machines might benefit from access to a vast, multi-language knowledge base. Integrating external resources like *Wikipedia* [2] into machine-generated translations could be an approach to address issues.

Wikipedia, despite not being regarded as a formal academic source, offers a comprehensive collection of terms in multiple languages with widely accepted translations. This makes it a valuable resource for translation tasks, especially for terms related to historical events, geographic locations, and organization names.

Wikipedia's official API allows for automated searches of terms and retrieval of information, including the pages of same terms in different languages which could be seamlessly incorporated into the translation pipeline with minimal additional effort. We could identify real-exist named entities from Wikipedia and retrieve the official translations form it.

3.2.2 Hypothesis: TermBase (TB) Augmented Translation

As named entity translation remains a challenging aspect for recent models, it may be worth considering a separation of named entity translation from the main text translation inference.

Building on the concept of term bases[3] and insights from (Ghazvininejad et al., 2023), we hypothesize that translating named entities first and then inserting the translated entities into the prompt could enhance LLMs' translation quality and help LLMs handle inconsistencies in named entity translation by ensuring uniformity before the main text translation begins.

[2] https://www.wikipedia.org/

[3] Also known as a translation glossary, a term base is a database that contains words, expressions, or terms in multiple languages.

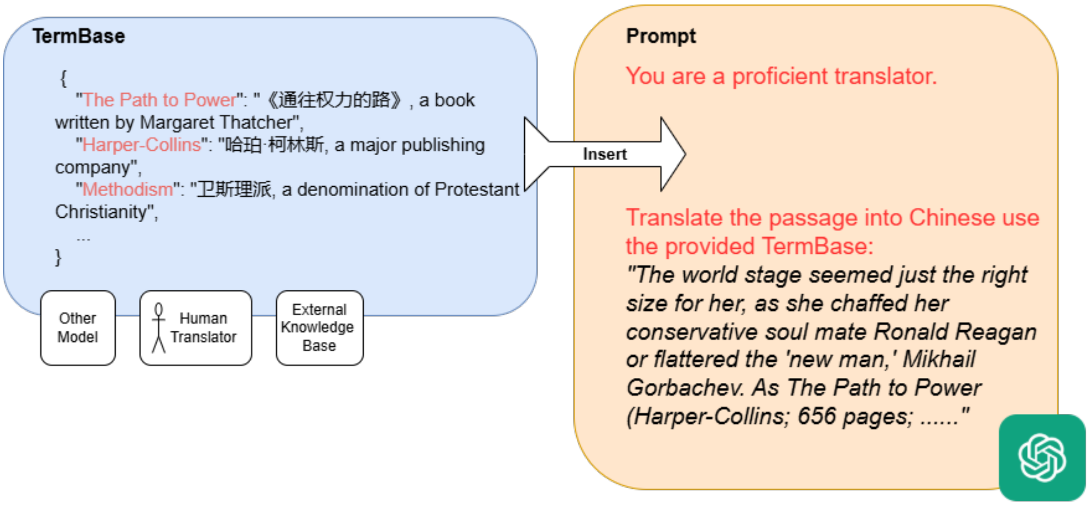

TermBase
{
"The Path to Power": " 《通往权力的路》 , a book written by Margaret Thatcher",
"Harper-Collins": "哈珀·柯林斯, a major publishing company",
"Methodism": "卫斯理派, a denomination of Protestant Christianity",
...
}
Other Model
Human Translator
External Knowledge Base
Insert
Prompt
You are a proficient translator.
Translate the passage into Chinese use the provided TermBase:
"The world stage seemed just the right size for her, as she chaffed her conservative soul mate Ronald Reagan or flattered the 'new man,' Mikhail Gorbachev. As The Path to Power (Harper-Collins; 656 pages; ......"

FIGURE 3.2: Our TermBase augmented method: Translate named entities and main text separately; Insert translated named entities as TermBase in our translation prompt

We are particularly interested in 4 type of named entities:

1. **Person name**: Includes names of individuals, both real and fictional. This category can also include titles and honorifics when they are attached to a name (e.g., 'Sherlock Holmes', 'President Obama').
2. **Location**: Geographical locations such as cities, countries, or natural landmarks (e.g., 'Paris', 'Mount Everest').
3. **Organization** Names of companies, institutions, governmental bodies, or groups (e.g., 'Google', 'The League of Explorers').
4. **Creative Work**: Titles of creative or intellectual works, such as books, films, paintings, software applications, and songs (e.g., 'War and Peace', 'Titanic', 'The Starry Night', 'Photoshop').
5. **Events**: Refers to historical, cultural, or notable events, such as wars, ceremonies, natural disasters, and significant conferences. (e.g., 'World War II', 'The Oscars').

We are **NOT** interested in named entities of type **'LANGUAGE', 'DATE', 'TIME', 'PERCENT', 'MONEY', 'QUANTITY', 'ORDINAL', 'CARDINAL'** as they are either well-known nouns that can be handled by LLM itself (e.g. "English") or not influence translation quality at all (e.g. "May 22 2024" vs "2024/05/22").

Leveraging LLMs' in-context learning capability Liu et al. (2021), we directly add a termbase (named entity translation table) to the original prompt. If the model successfully utilizes the inserted high-quality named entity translations, we expect this method to resolve most inconsistencies in named entity translation and may improve the overall translation quality and readability.

For example, in Figure 3.3, we added the blue Term Base section to the original translation prompt (purple part). The blue section provides the named entity translations in advance, ensuring the model uses consistent translations throughout the text.

You are translating a book part by part

Translate named entities with the below translations:

{ "named-entity1" : "translation of named-entity1",
"named-entity2" : "translation of named-entity2",
. . . }

Please translate the following part from English into Chinese. Use natural and fluent Chinese.

$PART_TO_TRANSLATE$

FIGURE 3.3: Template for experiment 1, Purple part (Baseline Template) represent the original template used for translation. Blue part (Termbase Insertion) is the added termBase section containing named entities and their translations.

We first validated in Section 4.1.1 that LLMs can follow instructions and utilize the term base effectively using a small dataset. After confirming the model's capability, we conducted two experiments in section 4.2 to evaluate this method for improving translation quality and consistency.

3.3 TermBase Augmented Translation Pipeline

Based on the results of our experiments in 4.2, we developed a streamlined pipeline for incorporating a termbase into prompt-based translation models. The pipeline (Figure 3.4) involves the following steps:

1. **Named Entity Recognition (NER)**
2. **Translation of Named Entities (and brief explain them)**
3. **Merge & Store named Entity Translations and Human Proofreading**
4. **Termbase generation and Final translation**

3.3.1 Named Entity Recognition (NER) and Translation

The first step involves extracting relevant named entities from the entire source text. This step is crucial to ensure consistent translation of specific terms, names, and locations throughout the text.

In our experiments, we utilized the latest *GPT-4-1106-preview* as the NER model for English. While this model performed well, other NER models, such as Stanza Qi et al. (2020), can also be employed, depending on the requirements. Details on our evaluation of NER models are provided in Section 4.1.1. It is important to note that different languages may necessitate the use of language-specific NER models.

Once the named entities are identified, each is translated based on its contextual meaning. To address information loss, such as gender, honorifics, or cultural nuances during translation, users can optionally include brief explanations for each named entity.

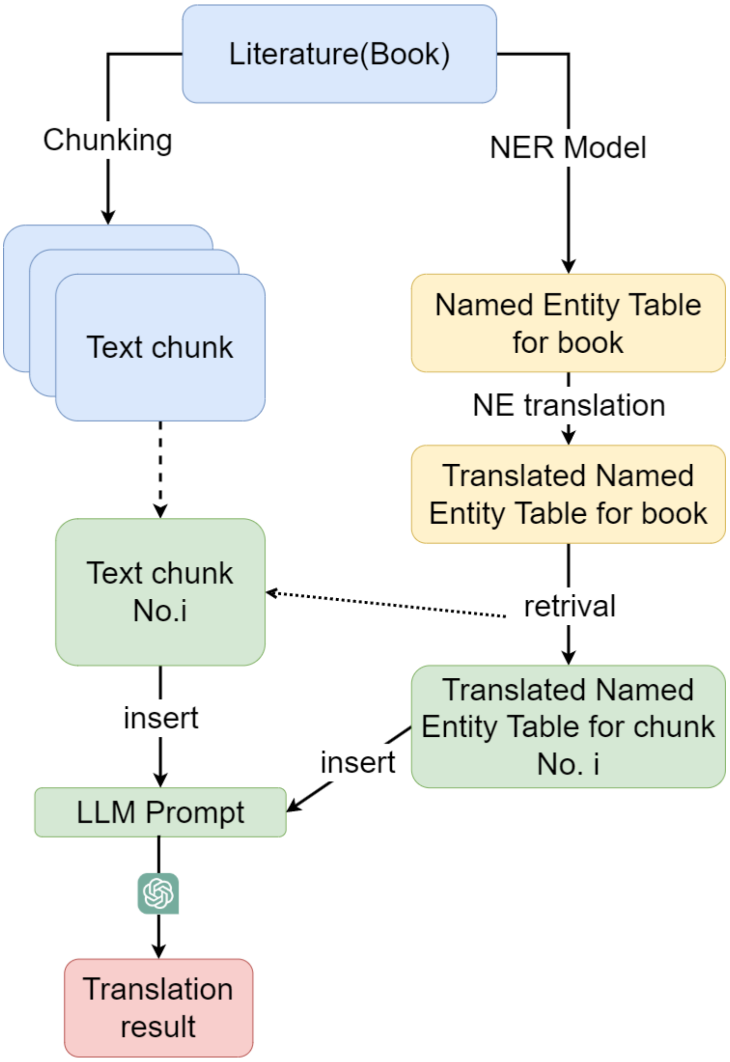


FIGURE 3.4: TermBase augmented translation pipeline

For translation, we continued using GPT models. As analyzed in Section 4.1.2, retrieval-augmented (RAG) with Wikipedia resources is not very helpful for this translation task. For simplicity, we opted not to integrate Wikipedia as Dictionary hypothesis into our pipeline.

We tested two translation setups:

1. Sentence as context Translation: For each named entity, we retrieved several sentences from the source text where the entity appeared. During the translation of named entities, these sentences were included in the prompt to provide context, reducing ambiguity. The

prompt we use here is Figure A.4.

2. Chunk as context Translation: The source text was divided into chunks of approximately 1000–2000 words. Each chunk, along with a named entity list, was fed into GPT. The model was tasked with translating the named entities within the list using the chunk text as context. While this approach facilitated contextual translation, it sometimes caused inconsistencies, as the same named entity could be translated differently across chunks. These inconsistencies were resolved in subsequent steps.

While both setups produced promising results at first glance, we did not perform a thorough evaluation due to a lack of defined evaluation criteria.

3.3.2 Merge & Store Named Entity Translations and Human Proofreading

After translation, the named entities (and its translation & description) are merged into a named entities table per book and stored in a database for future reference.

To ensure accuracy and contextual appropriateness, People might hire human translators to review and revise the named entity translations. Reviewing only the named entity table is a relatively light workload for human translators and can significantly improves translation quality by ensuring critical terms are correctly translated.

If fully automated translation is not a strict requirement, asking human translator for named entity translation is also a highly effective and efficient strategy, especially given current LLMs still struggle to address many issues on named entity translation

3.3.3 Termbase Generation and Final Translation

For the final translation phase, each chunk of the source text was processed as follows:

1. We filter named entities appearing in this chunk from the database

2. Retrieve those named entities' translations and explanations from the database.

3. Generate a termbase table and insert it into the translation prompt, as illustrated in Figure 3.3.

To assist the model in understanding this task, we use 1-shot paragraph-level demonstrations, as shown in Figure A.3. This approach attempt to utilized the pre-translated named entities effectively, maintaining consistency and readability across the final output.

Chapter 4

Experiments And Result Report

4.1 Validation of Hypothesis

4.1.1 LLM's Ability To Follow The TermBase Augmented Prompt

To assess the viability of the term base idea, we first designed an experiment to integrate a termbase into prompting-based machine translation (MT) models to **Test model's ability to follow the instruction**

We first modified the baseline prompt template to include an additional *Termbase* section

The updated prompt template, as illustrated in Figure 1, comprises two main parts:

1. **The Termbase Section**: This part lists pairs of named entities, with each line a phrase pair: *source language* - *target language*. We hope it serves as a reference for the model to ensure consistent translation of specific terms.

2. **The Source Text section**: The segment of original text needs to be translated.

We use the *GPT-4-1106-preview* [1] model to translate the first 20 chunks of the newly published book *Yellow Face*, with each chunk containing approximately 1,000 words (which is around 4,000 token for Chinese translated result using GPT-4-1106-preview's tokenizer). This

[1] GPT-4-turbo in 2023, before the update of tokenizer

book was chosen for its recency, ensuring it was not included in the model's training dataset. Additionally, the book's content, featuring a mix of Chinese names, Western names, and real organization names such as "Xiao Li", "Google" and "Yale University" makes it an ideal source text for testing our hypothesis.

We run the experimental procedure in 2 steps:

1. **Named Entity Recognition and Translation**: In the first run, we tasked *GPT-4-1106-preview* with identifying and translating only the named entities into Chinese within a single prompt (as shown in Figure A.2).

2. **Termbase Insertion and Text Segment Translation**: Then we created a termbase from these translations and inserted it into the prompt for the second run, using the prompt template in Figure 3.3

In step 1, we manually labeled named entities of above 4 categories 3.2.2 in the 20 chunks of text. We then employed two tools, Stanza (Qi et al., 2020) and GPT-4-1106-preview, to perform English Named Entity Recognition (NER) on the labeled data. For Stanza, we created an exclude list to filter out named entities belonging to the following categories: **'LANGUAGE', 'DATE', 'TIME', 'PERCENT', 'MONEY', 'QUANTITY', 'ORDINAL', 'CARDINAL'**.Upon comparing the performance of the two tools, we discovered that GPT-4-1106-preview outperformed Stanza in all metrics except Recall. Stanza demonstrated a tendency to output more named entities in the aforementioned categories, while GPT-4 produced fewer false results (Table 4.1).

TABLE 4.1: Performance Comparison of Different Models

Model	**Precision**	**Recall**	**F1 Score**
GPT-4	0.7719	0.7242	0.7369
Stanza with exclude list	0.6396	0.7815	0.6900
Stanza without exclude list	0.3774	0.7925	0.5048

As we placed a special emphasis on consistency in translation, we also runs NER task while filtering out phrases that appeared only once throughout the entire book. This filter is based on the rationale that phrases appearing merely once cannot demonstrate inconsistency across translations of different chunks and are likely less critical to the overall narrative of the book.

Similarly, in this task, the *GPT-4-1106-preview* model demonstrated high precision (0.9149) in the task with a slightly lower recall at 0.8217, suggesting the output missed some entities. On the other hand, the *Stanza* model and the union of two models showed a higher recall as *Stanza* seems to identify a broader range of entities. The *Stanza* model, however, came with a trade-off, as evidenced by its lower precision (0.7706), implying a greater likelihood of including false positives. In an attempt to leverage the strengths of both models, we combined their results by simply uniting their outputs. This approach helped us reach a high high recall of 0.9299. But the precision downed to 0.7725.

Upon deeper analysis of the results, we found:

1. The false negatives set from *GPT-4-1106-preview* model is less informative and are unlikely to impact the final translation result.

```
false_positives_set: {'japan', 'uber', 'xiao', 'chinese american ', 'q&a', 'germany', 'ah lung', 'us', 'china', 'cate', ' russia', 'korean american'}
```

2. The *Stanza* model captures more person name but also tend to include pronouns like "she's" in the "PERSON" category. The model might be achievable with more refined filtering to exclude such useless phrases, thus increasing the precision without compromising recall.

We hence selected *GPT-4-1106-preview* as our preferred NER model for the subsequent stages of this experiment.

After Step 2, we meticulously labeled and then counted the mistakes and inconsistencies in translations produced by both the baseline prompt (Figure 3.3 purple part) and the termbase-enhanced prompt (Figure 3.3). Our primary focus was on assessing the model's ability to adhere to the TermBase included in the prompt. We category errors in this experiment into 2 category:

1. **NE Mistakes**: These refer to instances where the model did not correctly follow the provided term base translation in Step 1, regardless of whether the translation in the TermBase itself was correct. For example, if the term "Jack" was incorrectly provided with translation "Cat" in the prompt, the model would not be penalized for translating "Jack" to "Cat."

2. **Inconsistencies**: These refer to instances where the same phrase was translated differently across different chunks. For example, both "Jun1 Ni1" and "Zhu1 Ni1" are acceptable Chinese translations of the name Jannie, but to maintain consistence, only one translation is acceptable throughout. (In some scenarios, translating the full name to a nickname or pronoun might be acceptable, but we still counted these as inconsistencies in this context.)

Mistake type	**Baseline**	**Termbase**
NE Mistakes	83	36
Inconsistencies	79	9

TABLE 4.2: Comparison of Translation Results. "NE Mistakes" refers to the model not correctly following the provided term base translation in Step 1, regardless of whether the translation in the TermBase itself was correct. "Inconsistencies" refer to instances where the same phrase was translated differently across different chunks.

Summarized in Table 4.2, these results demonstrate a significant reduction in both NE mistakes and inconsistencies when using **the termbase-enhanced prompt** compared to the baseline, **highlighting the ability of LLMs to faithfully follow the TermBase**. Upon further analysis, we also found that many NE mistakes made by the termbase-enhanced prompt involved nicknames or pronouns, which might be acceptable in a literary context. Additionally, it was observed that the *GPT-4-1106-preview* model could effectively leverage its extensive knowledge base to accurately translate common named entities like "Google", "Barack Obama", and "Treaty of Versailles" and likely to maintain phrase consistency within a single run of the prompt.

Later in section 4.2, we performed evaluation on WMT23 Shared Task: Discourse-Level Literary Translation, showing high consistency in terminology translation (cTT) for the TermBase-inserted prompt. This further verifies GPT-4's ability to follow our instructions accurately.

4.1.2 Integrate Wikipedia Helps Named Entity Recolonization and Translation, But Helps Little

To leverage Wikipedia as an additional resource for named entity translation, we propose updating our Named Entity Recognition and translation prompt A.5 to incorporate supplementary information from Wikipedia using Retrieval-Augmented Generation (RAG). Specifically, for each retrieved named entity, we include the following two components in the named entity list provided in the prompt (as illustrated in Figure A.5):

1. **First paragraph from the Wikipedia page**: This provides a concise description and explanation of the named entity, potentially resolving ambiguities and enhancing the model's understanding.

2. **Wikipedia page-name in the target language**: This provides the official translation.

We expect that this additional context will help LLM model better understand named entities and improve the accuracy of their translations.

Testing the integration of Wikipedia in the NER process for the book *Yellow Face* revealed interesting outcomes. The Wikipedia retrieval function (via their API) successfully identified 88 real named entities from the book, including terms which were incorrectly translated by our *GPT-4-1104-preview* model. For instance, the GPT-4 model failed to identify the author "Maxine Hong Kingston" and did not translate "Random House" or "El Centro (California)" with their official translations.

However, several potential issues were noted that could affect the practical application of this approach:

1. **Ambiguity of Named Entities**: Correctly identified named entities might not always correspond to the same object or concept on the Wikipedia page. This lack of contextual consideration can lead to inaccuracies. For instance, "yellow fever" could be refer to one infectious disease, but in certain contexts, it might refer to "Asian fetish" instead. Such nuances are missed when relying solely on a Wikipedia API call.

2. **Redundancy for Most Entities**: LLMs probabily already has wikipedia in their training data and Previous runs have shown that advanced LLMs are can often translate well-know entities accurately without extra assistance.

3. **Limitation of Wikipedia Database**: Wikipedia coverage limited range of specialized or less-known terms such as fictional names, niche jargon.

We found that the primary challenge in named entity translation actually arises from entities that fall outside the scope of Wikipedia, such as fictional names, niche jargon, or context-specific terms that lack standardized or widely accepted translations. These out-of-Wikipedia entities often require additional contextual knowledge or domain-specific experience, which current LLMs struggle to address effectively.

Hence we decided NOT integrate this add-on into our pipeline.

4.2 Evaluation on TermBase Augmented Translation

To further analyze our TermBase-augmented method, we evaluated the Translation Quality and Terminology Consistency on the WMT23 Shared Task: Discourse-Level Literary Translation dataset. The dataset, containing original Chinese web novels translated into English by human translators, features a golden standard human translation and contains many real and fictional named entities.

1. **Translation Quality** We utilized the evaluation metrics BLEU (Post, 2018) and COMET (Rei et al., 2020). BLEU measures the correspondence between models' output and reference translation, providing a score based on n-gram precision. COMET utilizes pre-trained multilingual models to evaluate translation quality based on human judgment, offering a more nuanced assessment than BLEU.

2. **Terminology Consistency** To measure the consistency of terminology translation, we followed the translation consistency metric (Wang et al., 2023; Lyu et al., 2021):

$$cTT = \frac{\sum_{w \in TB} \frac{\sum_{i=1}^{n} \sum_{j=i+1}^{n} \mathbb{1}(t_{wi}=t_{wj})}{C_n^2}}{|TB|} \tag{4.1}$$

   For each named entity w in the TermBase (TB), $t(w) = \{t_{w1}, t_{w2}, \cdots, t_{wn}\}$ denotes the list of translations for each w in the source text, and C_n^2 denotes the number of combinations of the translation list $t(w)$. The function $\mathbb{1}(t_{wi} = t_{wj})$ returns 1 if and only if $t_{wi} = t_{wj}$, otherwise, it returns 0. This metric measures the frequency of any two translations of w being the same throughout the book.

   To calculate cTT , we first use used awesome-align (Dou and Neubig, 2021) for word alignment and match each translated word of named entities in MT model's output. Then we calculated cTT using formula 4.1. The Python implementation of this calculation function is provided in B.1.

To perform the evaluation, we first used Bertalign (Liu and Zhu, 2023) to align the model output with the golden standard human translation in WMT format. Bertalign utilizes sentence transformers to map sentences into vector space and then use dynamic programming (DP) method to find the most similar pairs. We then verified the package results by setting a threshold on cosine similarity and do a post-editing by human annotators.

For the BLEU score, we used the following model

```
from sacrebleu.metrics import BLEU
bleu = BLEU(tokenize='zh')
```

For the COMET score, we used the *Unbabel/XCOMET-XL* model with a batch size of 32, running on an A100 GPU on Colab.

4.2.1 Experiment 1: WMT23 test set 1 and Book114 chapter 01 - 10

We first performed a small evaluation on English to Chinese translation using the WMT23 Shared Task: Discourse-Level Literary Translation test set 1 and Book114 chapters 01–10. The dataset contains both English and Chinese text and is already line-aligned. We first divided text into chunks of approximately 1000 words. These chunks were translated using the latest *gpt-4-0125-preview*[2] with the following settings: temperature 0.3, presence penalty 0, and frequency penalty 0.

Each Text was translated using two prompts:

1. The TermBase-augmented prompt (Figure A.3), which included pre-translated named entities, with one-shot paragraph-level demonstrations.

2. The baseline normal translation prompt (Figure A.6), which used no TermBase augmentation, with one-shot paragraph-level demonstrations.

[2] Experiment conducted in January 2024

Both setups employed one-shot paragraph-level demonstrations. To tests whether inserting an additional TermBase in the prompt affects translation quality, the TermBase experiment was conducted with two setups:

1. **Low quality translated TermBase**: The translation of entities was directly generated by a simple machine translation system like Google Translation, which likely does not match the golden standard. This setup tests whether inserting an additional TermBase in the prompt affects translation quality.

2. **Golden Standard TermBase**: This setup uses the golden standard translation of named entities. By comparing this with the baseline, we can assess the influence of the TermBase-augmented prompt on translation quality and consistency.

   To obtain the golden standard TermBase:

   (a) We first performed Named Entity Recognition (NER) on the source text (English) using GPT-4 (prompt A.2) to extract named entities from each chunk.

   (b) For each named entity, we then retrieved several sentences containing the entity from the source text (English) and its corresponding reference text (Chinese). These reference sentences included the translated entities.

   (c) Using GPT-4 and a specific prompt A.7, we identified the translation of each named entity within the reference sentences. We use this as the golden standard named entity translation.

   Note although the named entity translation we obtained is golden standard, we might not recognize the correct set of named entities in step one, which caused several failure case in later experiment and we analyze them in 4.2.2.

The results are shown in Tables 4.3 and 4.4.

Experiment	BLEU ↑	COMET ↑	cTT ↑
Baseline	26.892	0.756	0.811
With TermBase (Low Quality)	26.070	0.751	-
With TermBase (Gold Standard)	29.328	0.785	0.921

TABLE 4.3: Evaluation results on WMT23 test set 1

Experiment	BLEU ↑	COMET ↑	cTT ↑
Baseline	17.696	0.8460	0.7179
With TermBase (Gold Standard)	18.539	0.8577	0.9078

TABLE 4.4: Evaluation results on WMT23 Book114 chapters 01–10

Findings When using a low quality translated TermBase, the BLEU and COMET scores slightly decreased by 4% and 1%, respectively. This slight decrease indicates **the TermBase in the prompt (which make the prompt more complex) will have very little influence on the translation quality**. The decrease here is likely due to the "wrong" guidelines provided by the inaccurately translated named entities.

In the *WMT23 test set 1*, using a TermBase with gold standard translations increased the BLEU score from 26.892 to 29.328 and the COMET score from 0.756 to 0.785. Similarly, for *Book114 chapters 01 - 10*, using a gold standard TermBase resulted in improvements, with the BLEU score increasing from 17.696 to 18.539 and the COMET score from 0.8460 to 0.8577. This highlights **the importance of the quality of TermBase translations**, as a high-quality TermBase indeed increase the overall translation quality.

Moreover, the cTT score, which measures terminology consistency, showed significant improvement with termbase prompt. For the *WMT23 test set 1*, the cTT score improved from 0.811 to 0.921, increase for around 13%. For *Book114 chapters 01 - 10*, the cTT score increased from 0.7179 to 0.9078, resulting a 26% improvement in named entities translation consistency. This further demonstrating incorporating a TermBase significantly improves terminology consistency.

4.2.2 Experiment 2: 27 books from wmt23 dataset

Next, we conducted a large-scale evaluation using 27 randomly selected books [3] from the WMT23 Shared Task: Discourse-Level Literary Translation dataset. Each selected book contained 50 chapters, resulting in a comprehensive test set for evaluation.

We tested four models:

1. **GPT-4o-2024-08-06 OpenAI (2023)**: One of the most advanced LLM.
2. **GPT-4o-mini-2024-07-18**: A smaller, cost-efficient model in GPT-4o series.
3. **Qwen-Plus-2024-09-19 Team (2024) Team (2024)**: One of the most advanced LLM; developed by Chinese company Alibaba, with strong training on Chinese data.
4. **Qwen-Turbo-2024-09-19**: A smaller, cost-efficient model in Qwen series.

We included Qwen models due to their focus on Chinese data, which may result in better performance in English-to-Chinese translation tasks. Additionally, we included cost-efficient smaller models (*GPT-4o Mini*, *Qwen-Turbo*) to evaluate whether a smaller model could effectively follow our TermBase instructions and perform consistent translations.

We obtained the TermBase using the same method described in 2. All models were evaluated using their batch APIs with consistent settings: temperature 0.3, presence penalty 0, and frequency penalty 0. For *GPT-4o*, we used JSON-style prompts (templates A.8, A.9). However, we observed that the other three models did not consistently produce valid JSON outputs when handling output of thousands of tokens. As a result, we converted these prompts into equivalent string-style formats (templates A.10, A.11) and use them in this experiment.

[3] 27 book id: '50-nswdllbw', '129-gwltq', '75-sghndhm', '167-wdbz', '151-gfsy', '97-jmtj', '31-yxxgzyscks', '35-kbds', '139-cfdw', '153-ldyb98k', '177-syfjytlnl', '46-yghknct', '64-sh', '135-czyxsj', '68-mcxw', '59-ywxcbjn', '73-ylwsyed', '87-wdyty4xs', '140-yjqzyds', '76-dyxl', '98-wyctUp', '16-cjsjy', '123-whrqysj', '154-nsxhn', '147-wnyxs', '120-yzjd', '114-wyzjzfs'

Model	BLEU ↑	COMET ↑	cTT ↑
GPT-4o Mini (Baseline)	20.90	0.8089	0.6127
GPT-4o (Baseline)	23.24	0.8236	0.6453
GPT-4o Mini (Termbase)	23.97	0.8384	**0.8660**
GPT-4o (Termbase)	**25.78**	**0.8450**	**0.8750**
Qwen-Turbo (Baseline)	21.87	0.8126	0.5762
Qwen-Plus (Baseline)	23.92	0.8236	0.5842
Qwen-Turbo (Termbase)	24.20	0.8360	**0.8246**
Qwen-Plus (Termbase)	**27.18**	**0.8499**	**0.8685**

TABLE 4.5: Comparison of *BLEU*, *COMET*, and *cTT* scores across different models with and without termbase augmentation.

Results

The results are presented in Table 4.5. We found that **incorporating the TermBase improved *BLEU*, *COMET*, and *cTT* scores across all four models.**

Interestingly, the cost-efficient smaller models (e.g., *gpt4omini* and *Qwen-Turbo*) performed remarkably well on this task. These models effectively interpreted our TermBase instructions and **outperformed** larger models without TermBase augment (baseline) in terms of both translation **quality** and **consistency**.

One notable finding was the substantial improvement in *cTT* scores when using the TermBase.

1. **GPT-4o Series:**

   (a) **GPT-4o Mini**: baseline achieved a *cTT* score of **0.6127**, which rose to **0.8660** with TermBase augmentation—a relative increase of **41.4%**.

   (b) **GPT-4o**: baseline scored **0.6453** for *cTT*, which increased to **0.8750** with TermBase—a relative improvement of **35.6%**.

2. **Qwen Series:**

   (a) **Qwen-Turbo**: baseline scored **0.5762**, which increased to **0.8246** with TermBase—a relative increase of **43.1%**.

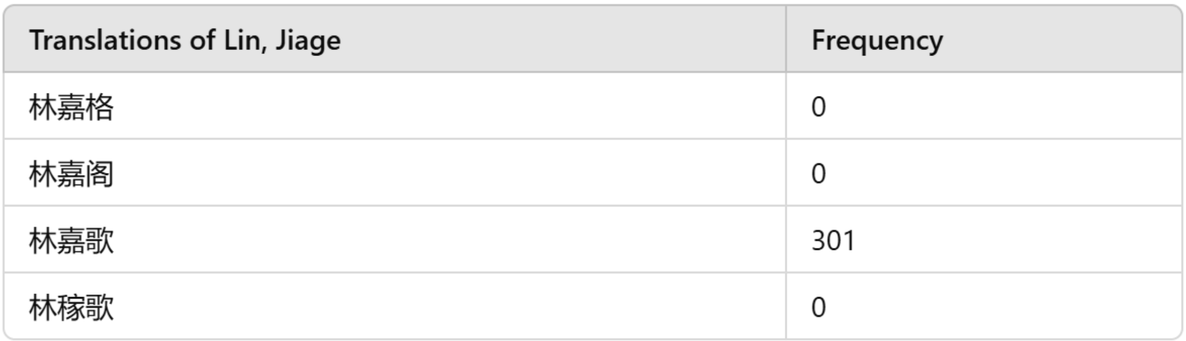
(b) **Qwen-Plus**: baseline improved from **0.5842** to **0.8685**, a **48.6%** relative improvement.

Meanwhile, our *cTT* scores measurement might underestimate the actual consistency due to limitations in the evaluation method:

- In our *cTT* calculation process 2, we first preformed the word alignment using Dou and Neubig (2021) and then calculated the *ctt* score based on those aligned words. In this process, some correctly translated named entities were not aligned accurately, which negatively affected the *cTT* score in Table 4.5.

- Cases where person names were translated into pronouns (e.g., "he" or "she") were counted as inconsistencies in our evaluation, even though these translations were contextually appropriate.

- As noted in 2, the TermBase used was not a de facto "golden standard": It sometimes included phrase that is not named entities– which do not need to translated consistently; sometimes it included conflicts: two different named entities-translation pairs referred to the same object are both in the termbase.

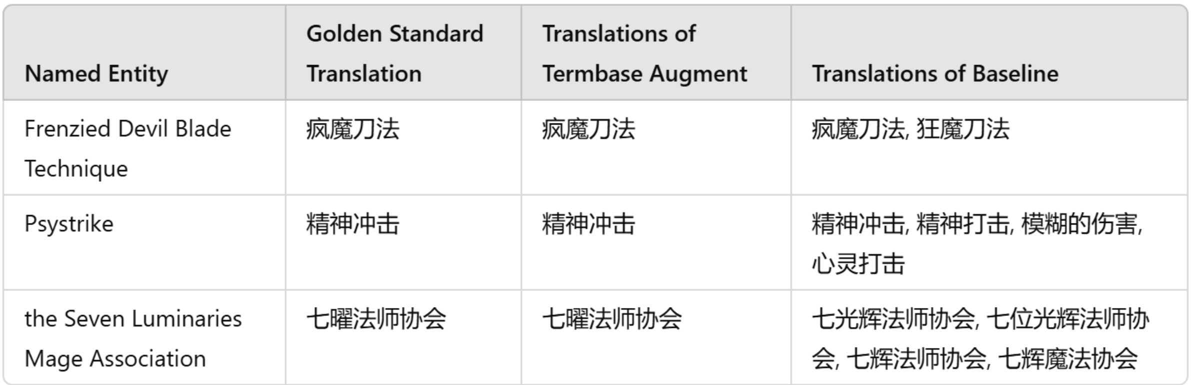


Translations of Lin, Jiage	Frequency
林嘉格	0
林嘉阁	0
林嘉歌	301
林稼歌	0

TABLE 4.6: Example: the name "Lin, Jiage" was translated consistently with the TermBase-augmented approach, whereas the baseline produced four different, albeit reasonable, translations (see Table 3.2).

Upon closer analysis of the results by human, we observed that **most**[4] named entities were consistently translated when using the TermBase. In contrast, the baseline models produced multiple reasonable but inconsistent translations for the same entity. (Table 4.6)

Beyond common named entities, the TermBase also maintain the consistency of translations for specialized terms of fiction: such as weapon names, skill names, and fictional concepts (Table 4.7). Which make this method very useful during fiction and web-novel translation.



Named Entity	Golden Standard Translation	Translations of Termbase Augment	Translations of Baseline
Frenzied Devil Blade Technique	疯魔刀法	疯魔刀法	疯魔刀法, 狂魔刀法
Psystrike	精神冲击	精神冲击	精神冲击, 精神打击, 模糊的伤害, 心灵打击
the Seven Luminaries Mage Association	七曜法师协会	七曜法师协会	七光辉法师协会, 七位光辉法师协会, 七辉法师协会, 七辉魔法协会

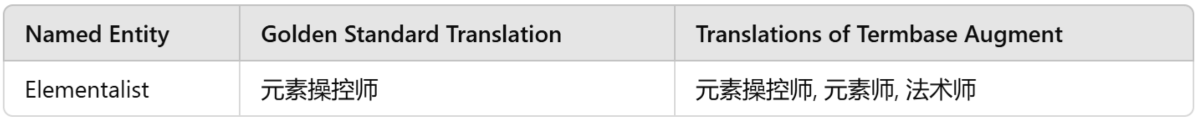
TABLE 4.7: Entities such as weapon names, skill names, and fictional concepts were also translated consistently with the TermBase-augmented approach.

Failure cases However, our TermBase-augmented approach do had some notable failure cases. Some common nouns that were context-specific to a particular book or fiction genre, such as "Elementalist" (a character class) or "Card" (a weapon group), were inconsistently translated even when specified in the TermBase (Table 4.8).

Phrases that are not traditional named entities, such as mantras or catchphrases (e.g., “I Don’t Know Everything, I Just Know What I Know.”, “It’s Time To Duel!”), were also less likely to be translated consistently or uniquely, even when included in the TermBase.

[4] except very few ($\leq$ 5%) failure case

These failure cases suggest that while the TermBase is effective for named entities and structured terms, its utility can be limited for more abstract or context-dependent phrases. We need to be aware of the converge and usage when creating Termbase.



Named Entity	Golden Standard Translation	Translations of Termbase Augment
Elementalist	元素操控师	元素操控师, 元素师, 法术师

TABLE 4.8: The term "Elementalist" was not translated consistently despite being specified in the TermBase.

Chapter 5

Conclusion, Limitation, and Future Research

5.1 Summary of Findings

This thesis investigates and analyzes the challenges of named entity translation (3.1.1) in long-text translation (e.g., novels translation) using LLMs. To address these challenges, we proposes a simple yet effective TermBase-augmented in-context learning method without requiring pre-training or fine-tuning. Our method was evaluated on the WMT23 Shared Task: Discourse-Level Literary Translation dataset, achieving promising results in both translation quality and terminology consistency. Additionally, we proposed a translation pipeline that incorporates a TermBase to enhance the translation of long texts.

1. **Challenges in Named Entity Translation:** Named entities, including proper nouns, character names, and locations, present significant challenges in translation due to cultural, linguistic, and contextual nuances. Meanwhile, document/book-level translation demands high consistency for terminology, which can be difficult for models to maintain autonomously.

2. **Effectiveness of TermBase Integration:** Our TermBase augment method (integrating a high-quality TermBase into prompts) significantly enhances translation quality and terminology consistency. The quality of the TermBase is critical: while using a randomly translated TermBase resulted in slight declines in BLEU and COMET scores, a gold-standard TermBase led to significant improvements. This underscores the importance of accurate named entity translations and the benefits of detailed, contextually relevant resources. Experiments show that high-quality TermBases enabled models like *GPT-4o* and *Qwen* to achieve increased *BLEU*, *COMET*, and *cTT* scores on the WMT23 Shared Task: Discourse-Level Literary Translation dataset, demonstrating improvements in translation quality, terminology consistency, and readability.

3. **Terminology Consistency (cTT)**: The TermBase-augmented method notably reduced inconsistencies in named entity translations. The cTT scores, which measure terminology consistency, showed significant (around 40% to 50%) improvement with the integration of a TermBase.

4. **Performance of TermBase-Augmented Smaller Models**: Smaller, cost-efficient models (e.g., *GPT-4oMini* and *Qwen-Turbo*) performed remarkably well on this task. These models not only interpreted TermBase instructions effectively but also **outperformed** full-scale models without TermBase augmentation (baseline) in terms of translation **quality** and **consistency**. Additionally, the cost of using these smaller models was substantially lower-—translating a single book cost as little as 0.10 USD.[1]

5.2 Limitations and Future Work

While our findings are promising, this study has several limitations:

[1] Cost varies depending on the book; this calculation is based on "Yellow Face."

1. **Model Limitations**: Our experiments only covered commercial closed-source models *GPT-4o* and *Qwen*. We did not evaluate open-source smaller models (e.g., those under 7B parameters) with short input/output token limits, leaving their usability unexplored.

2. **Language and Directionality:** This study focused exclusively on English-to-Chinese translation. We did not explore other translation directions, particularly for low-resource languages. Future work should investigate the applicability of TermBase integration across a broader range of languages.

3. **Pipeline Evaluation:** While we proposed a fully automatic TermBase-augmented translation pipeline, a thorough evaluation of its performance remains pending. A promising alternative approach could be: having humans translate named entities first, followed by using cost-efficient LLMs for full text translation.

We also shared our method with the SakuraLLM team[2], which has since implemented TermBase augmentation support in their v1.0 series models[3]. However, comprehensive evaluations of these models on using the TermBase augmentation method have not yet been conducted, leaving their performance unverified.

This study opens several avenues for future research:

1. **Expansion of Knowledge Bases:** Expanding the range of external resources in retrieval-augmented prompts could enhance translation quality and consistency. Domain-specific resources like the Allan Poe Wordlist or BanG Dream! Wikia offer potential for improving translations in specialized fields.

2. **Automatic Named Entity Translation**: While human translation of named entities requires minimal effort, developing advanced automatic named entity translation systems could eliminate this step and save time and effects.

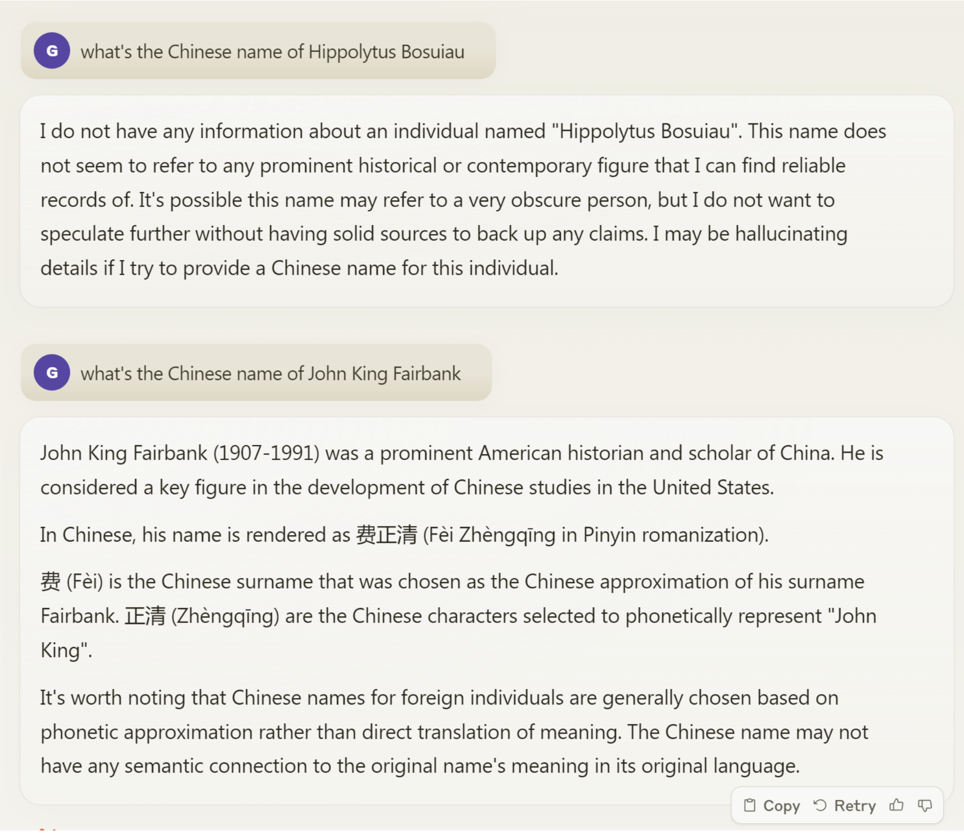
[2] SFT and RLHF models incorporating universal character and relationship attributes for Japanese-to-Chinese translation in the light novel and visual novel domains. https://github.com/SakuraLLM/SakuraLLM.

[3] Sakura-14B-Qwen2.5-v1.0, Qwen2.5-1.5B-v1.0, Sakura-7B-Qwen2.5-v1.0-GGUF.

3. **Incorporating Human Translation Practices**: This study suggests rethinking how established human translation practices—such as translation memory and search engine use—can be integrated into LLMs. These practices might significantly improve translation outcomes when properly implemented.

Appendix A

Prompts



G what's the Chinese name of Hippolytus Bosuiau

I do not have any information about an individual named "Hippolytus Bosuiau". This name does not seem to refer to any prominent historical or contemporary figure that I can find reliable records of. It's possible this name may refer to a very obscure person, but I do not want to speculate further without having solid sources to back up any claims. I may be hallucinating details if I try to provide a Chinese name for this individual.

G what's the Chinese name of John King Fairbank

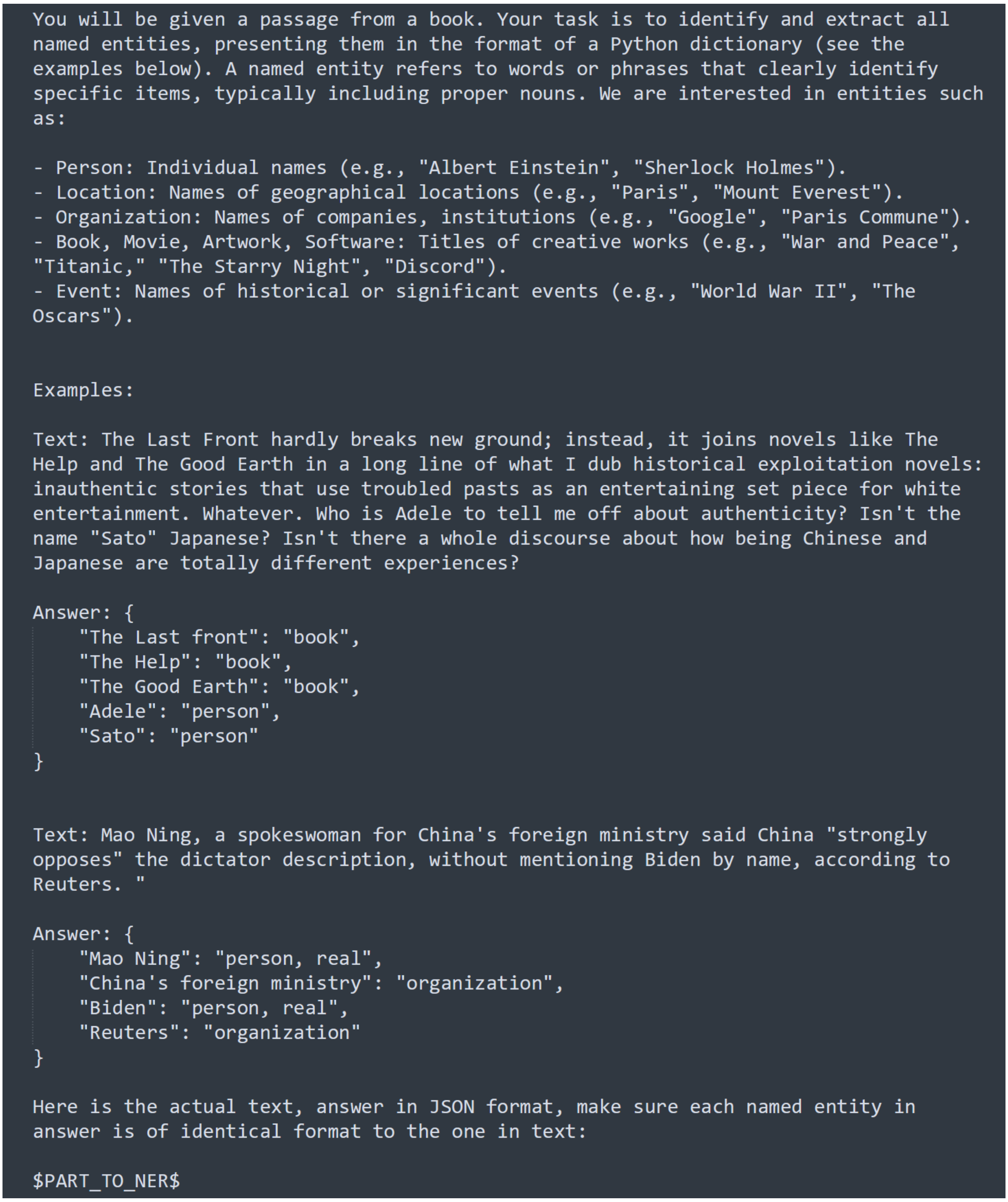
John King Fairbank (1907-1991) was a prominent American historian and scholar of China. He is considered a key figure in the development of Chinese studies in the United States.

In Chinese, his name is rendered as 费正清 (Fèi Zhèngqīng in Pinyin romanization).

费 (Fèi) is the Chinese surname that was chosen as the Chinese approximation of his surname Fairbank. 正清 (Zhèngqīng) are the Chinese characters selected to phonetically represent "John King".

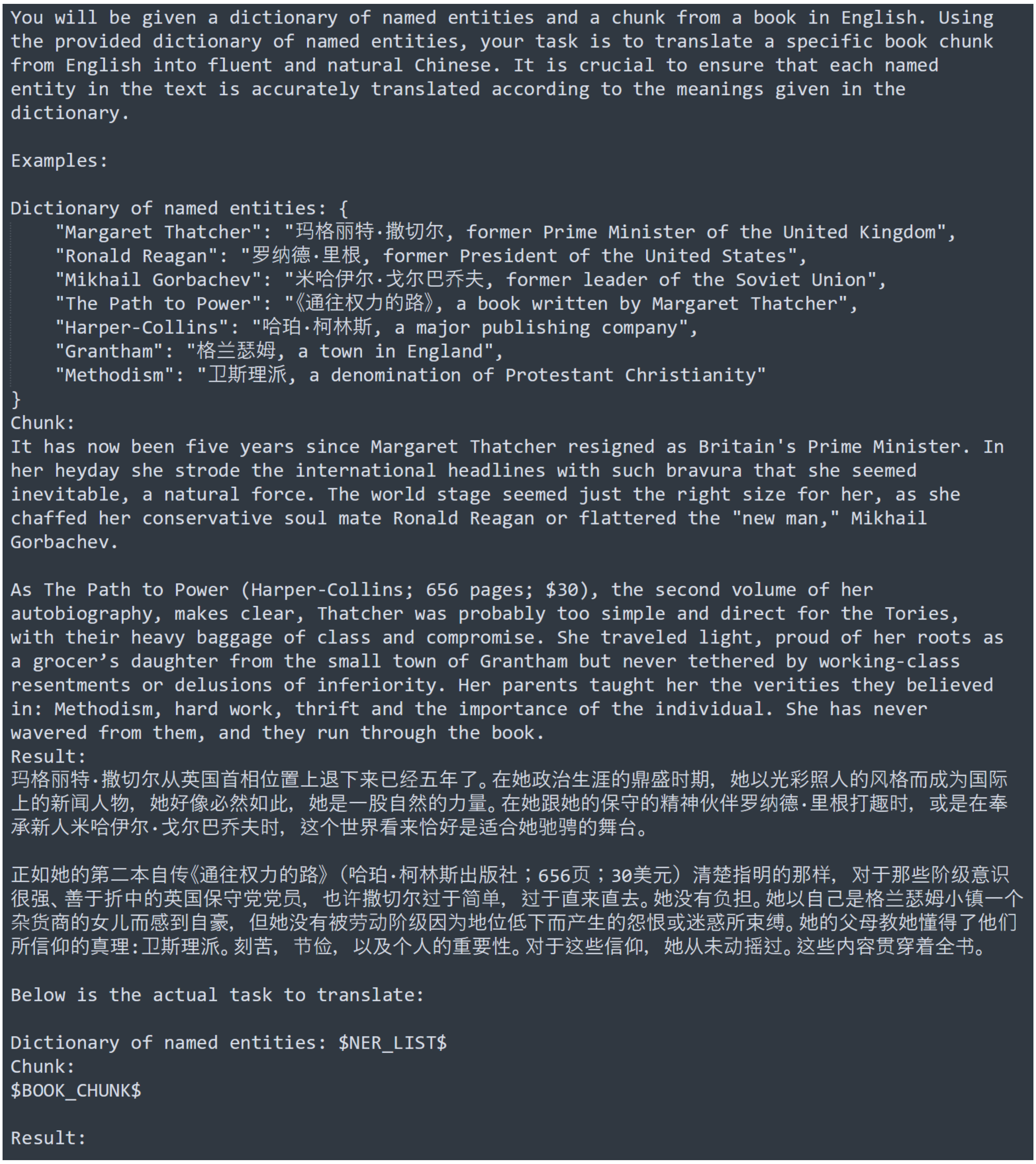
It's worth noting that Chinese names for foreign individuals are generally chosen based on phonetic approximation rather than direct translation of meaning. The Chinese name may not have any semantic connection to the original name's meaning in its original language.

FIGURE A.1: GPT-4 and Claude 3 Opus successfully translated John King Fairbank but failed on Hippolytus Bosuiau



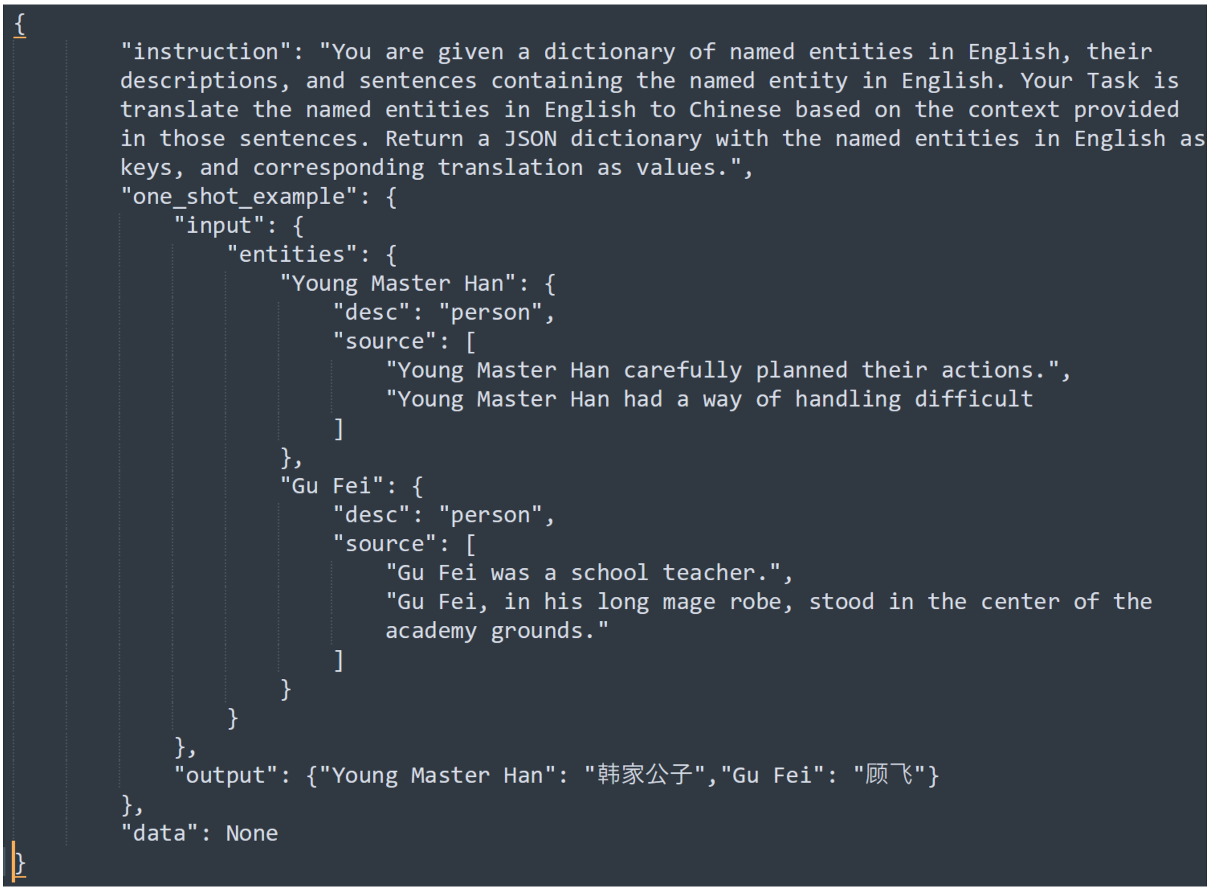
```
You will be given a passage from a book. Your task is to identify and extract all
named entities, presenting them in the format of a Python dictionary (see the
examples below). A named entity refers to words or phrases that clearly identify
specific items, typically including proper nouns. We are interested in entities such
as:

- Person: Individual names (e.g., "Albert Einstein", "Sherlock Holmes").
- Location: Names of geographical locations (e.g., "Paris", "Mount Everest").
- Organization: Names of companies, institutions (e.g., "Google", "Paris Commune").
- Book, Movie, Artwork, Software: Titles of creative works (e.g., "War and Peace",
"Titanic," "The Starry Night", "Discord").
- Event: Names of historical or significant events (e.g., "World War II", "The
Oscars").


Examples:

Text: The Last Front hardly breaks new ground; instead, it joins novels like The
Help and The Good Earth in a long line of what I dub historical exploitation novels:
inauthentic stories that use troubled pasts as an entertaining set piece for white
entertainment. Whatever. Who is Adele to tell me off about authenticity? Isn't the
name "Sato" Japanese? Isn't there a whole discourse about how being Chinese and
Japanese are totally different experiences?

Answer: {
    "The Last front": "book",
    "The Help": "book",
    "The Good Earth": "book",
    "Adele": "person",
    "Sato": "person"
}


Text: Mao Ning, a spokeswoman for China's foreign ministry said China "strongly
opposes" the dictator description, without mentioning Biden by name, according to
Reuters. "

Answer: {
    "Mao Ning": "person, real",
    "China's foreign ministry": "organization",
    "Biden": "person, real",
    "Reuters": "organization"
}

Here is the actual text, answer in JSON format, make sure each named entity in
answer is of identical format to the one in text:

$PART_TO_NER$
```

FIGURE A.2: GPT prompt for Named entity recognition (NER) task



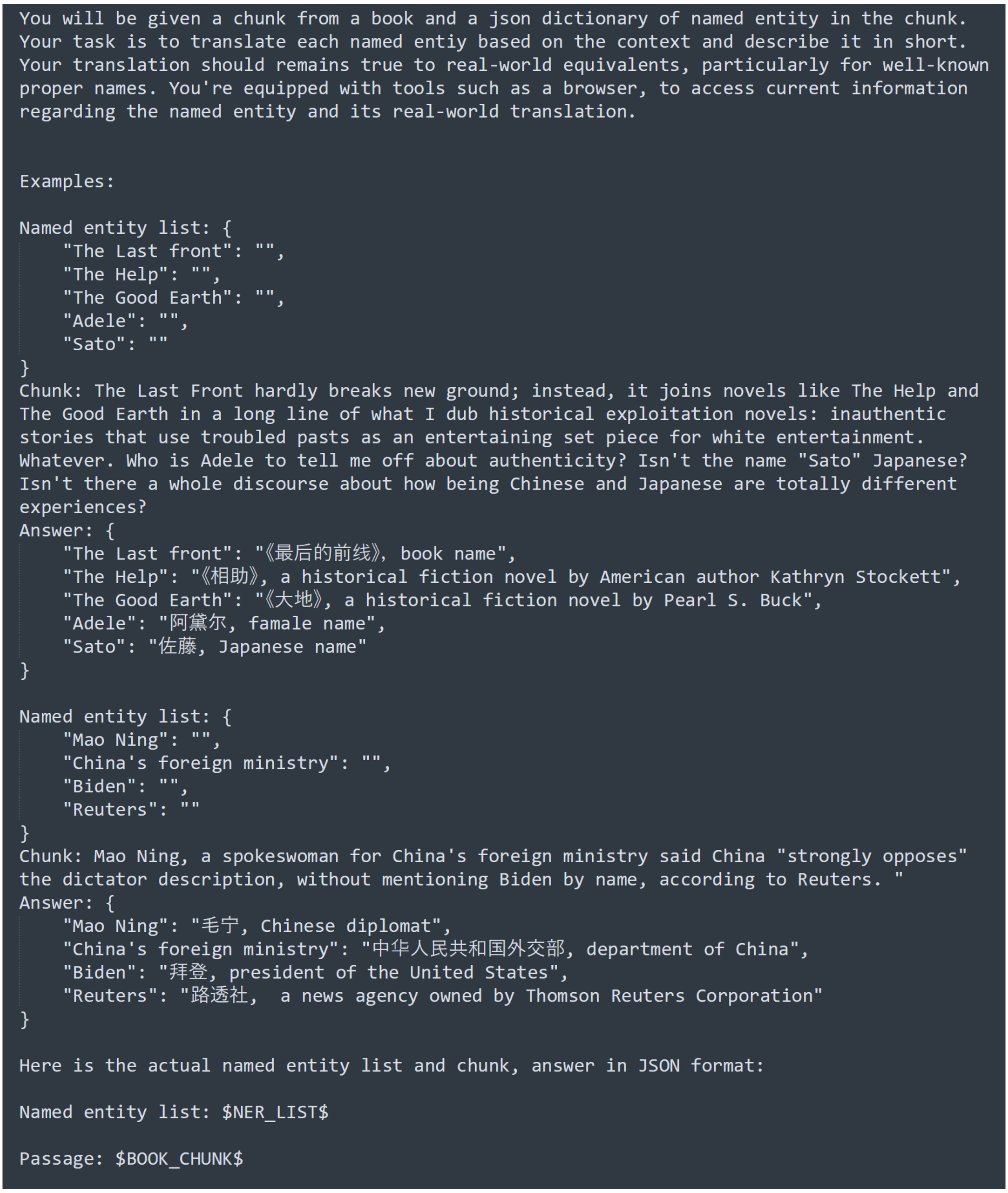
```
You will be given a dictionary of named entities and a chunk from a book in English. Using
the provided dictionary of named entities, your task is to translate a specific book chunk
from English into fluent and natural Chinese. It is crucial to ensure that each named
entity in the text is accurately translated according to the meanings given in the
dictionary.

Examples:

Dictionary of named entities: {
    "Margaret Thatcher": "玛格丽特·撒切尔, former Prime Minister of the United Kingdom",
    "Ronald Reagan": "罗纳德·里根, former President of the United States",
    "Mikhail Gorbachev": "米哈伊尔·戈尔巴乔夫, former leader of the Soviet Union",
    "The Path to Power": "《通往权力的路》, a book written by Margaret Thatcher",
    "Harper-Collins": "哈珀·柯林斯, a major publishing company",
    "Grantham": "格兰瑟姆, a town in England",
    "Methodism": "卫斯理派, a denomination of Protestant Christianity"
}
Chunk:
It has now been five years since Margaret Thatcher resigned as Britain's Prime Minister. In
her heyday she strode the international headlines with such bravura that she seemed
inevitable, a natural force. The world stage seemed just the right size for her, as she
chaffed her conservative soul mate Ronald Reagan or flattered the "new man," Mikhail
Gorbachev.

As The Path to Power (Harper-Collins; 656 pages; $30), the second volume of her
autobiography, makes clear, Thatcher was probably too simple and direct for the Tories,
with their heavy baggage of class and compromise. She traveled light, proud of her roots as
a grocer’s daughter from the small town of Grantham but never tethered by working-class
resentments or delusions of inferiority. Her parents taught her the verities they believed
in: Methodism, hard work, thrift and the importance of the individual. She has never
wavered from them, and they run through the book.
Result:
玛格丽特·撒切尔从英国首相位置上退下来已经五年了。在她政治生涯的鼎盛时期，她以光彩照人的风格而成为国际
上的新闻人物，她好像必然如此，她是一股自然的力量。在她跟她的保守的精神伙伴罗纳德·里根打趣时，或是在奉
承新人米哈伊尔·戈尔巴乔夫时，这个世界看来恰好是适合她驰骋的舞台。

正如她的第二本自传《通往权力的路》（哈珀·柯林斯出版社；656页；30美元）清楚指明的那样，对于那些阶级意识
很强、善于折中的英国保守党党员，也许撒切尔过于简单，过于直来直去。她没有负担。她以自己是格兰瑟姆小镇一个
杂货商的女儿而感到自豪，但她没有被劳动阶级因为地位低下而产生的怨恨或迷惑所束缚。她的父母教她懂得了他们
所信仰的真理:卫斯理派。刻苦，节俭，以及个人的重要性。对于这些信仰，她从未动摇过。这些内容贯穿着全书。

Below is the actual task to translate:

Dictionary of named entities: $NER_LIST$
Chunk:
$BOOK_CHUNK$

Result:
```

FIGURE A.3: TermBase augmented chunk translation prompt



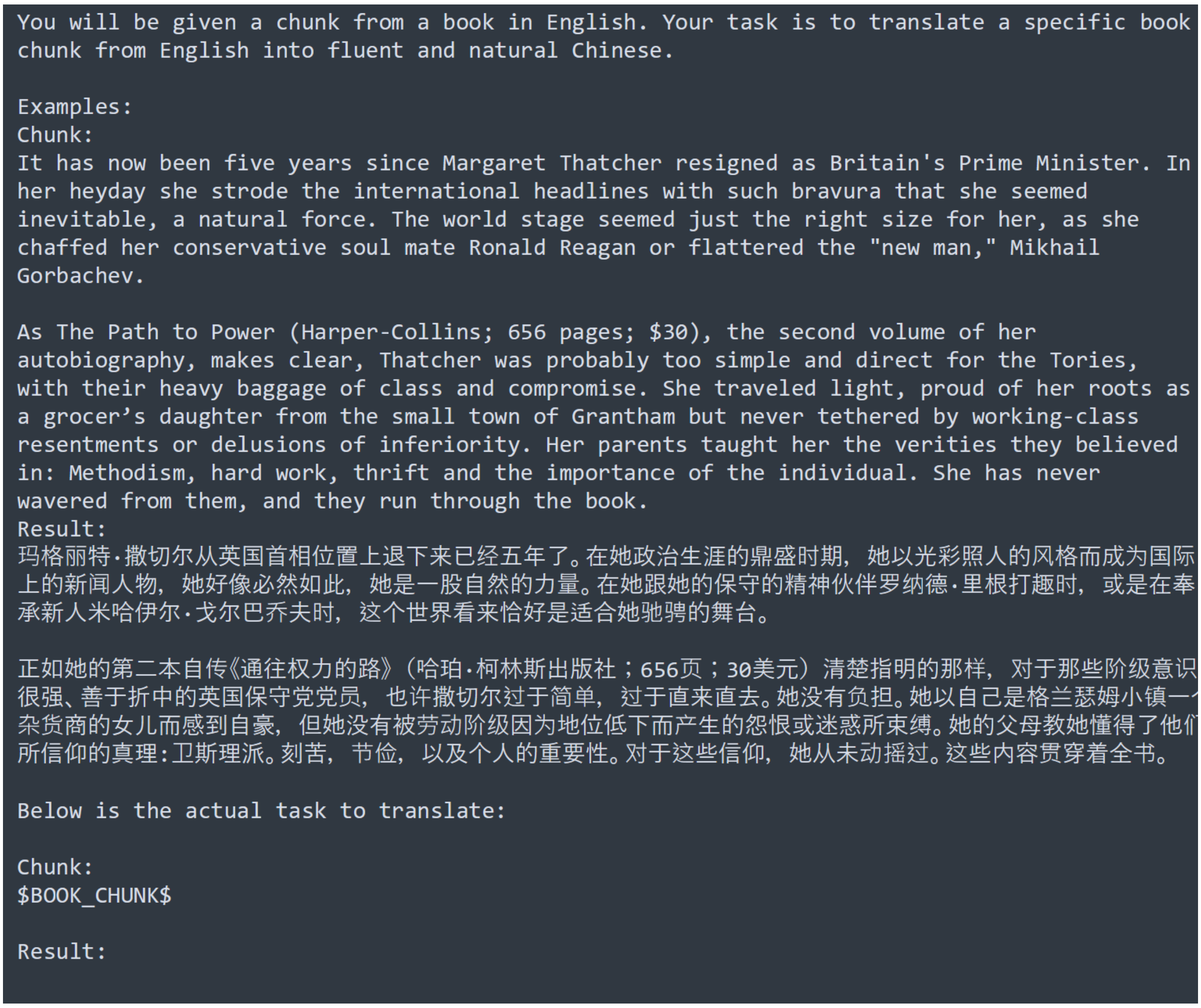
```
{
        "instruction": "You are given a dictionary of named entities in English, their
        descriptions, and sentences containing the named entity in English. Your Task is
        translate the named entities in English to Chinese based on the context provided
        in those sentences. Return a JSON dictionary with the named entities in English as
        keys, and corresponding translation as values.",
        "one_shot_example": {
            "input": {
                "entities": {
                    "Young Master Han": {
                        "desc": "person",
                        "source": [
                            "Young Master Han carefully planned their actions.",
                            "Young Master Han had a way of handling difficult
                        ]
                    },
                    "Gu Fei": {
                        "desc": "person",
                        "source": [
                            "Gu Fei was a school teacher.",
                            "Gu Fei, in his long mage robe, stood in the center of the
                            academy grounds."
                        ]
                    }
                }
            },
            "output": {"Young Master Han": "韩家公子","Gu Fei": "顾飞"}
        },
        "data": None
}
```

FIGURE A.4: Sentence as context Translation prompt



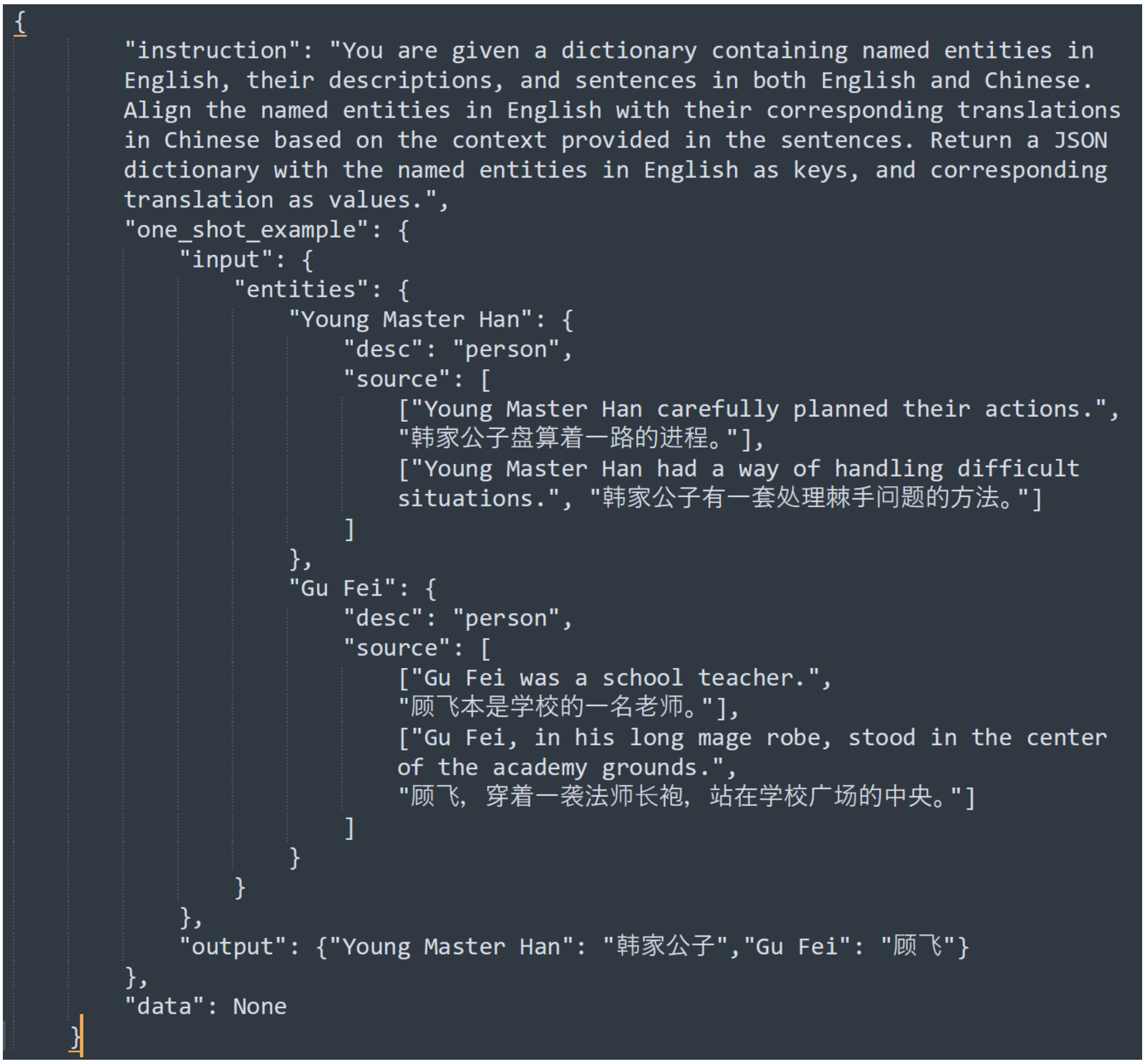
```
You will be given a chunk from a book and a json dictionary of named entity in the chunk.
Your task is to translate each named entiy based on the context and describe it in short.
Your translation should remains true to real-world equivalents, particularly for well-known
proper names. You're equipped with tools such as a browser, to access current information
regarding the named entity and its real-world translation.


Examples:

Named entity list: {
    "The Last front": "",
    "The Help": "",
    "The Good Earth": "",
    "Adele": "",
    "Sato": ""
}
Chunk: The Last Front hardly breaks new ground; instead, it joins novels like The Help and
The Good Earth in a long line of what I dub historical exploitation novels: inauthentic
stories that use troubled pasts as an entertaining set piece for white entertainment.
Whatever. Who is Adele to tell me off about authenticity? Isn't the name "Sato" Japanese?
Isn't there a whole discourse about how being Chinese and Japanese are totally different
experiences?
Answer: {
    "The Last front": "《最后的前线》，book name",
    "The Help": "《相助》, a historical fiction novel by American author Kathryn Stockett",
    "The Good Earth": "《大地》, a historical fiction novel by Pearl S. Buck",
    "Adele": "阿黛尔, famale name",
    "Sato": "佐藤, Japanese name"
}

Named entity list: {
    "Mao Ning": "",
    "China's foreign ministry": "",
    "Biden": "",
    "Reuters": ""
}
Chunk: Mao Ning, a spokeswoman for China's foreign ministry said China "strongly opposes"
the dictator description, without mentioning Biden by name, according to Reuters. "
Answer: {
    "Mao Ning": "毛宁, Chinese diplomat",
    "China's foreign ministry": "中华人民共和国外交部, department of China",
    "Biden": "拜登, president of the United States",
    "Reuters": "路透社,  a news agency owned by Thomson Reuters Corporation"
}

Here is the actual named entity list and chunk, answer in JSON format:

Named entity list: $NER_LIST$

Passage: $BOOK_CHUNK$
```

FIGURE A.5: GPT prompt for Named entity recognition (NER) and translation task



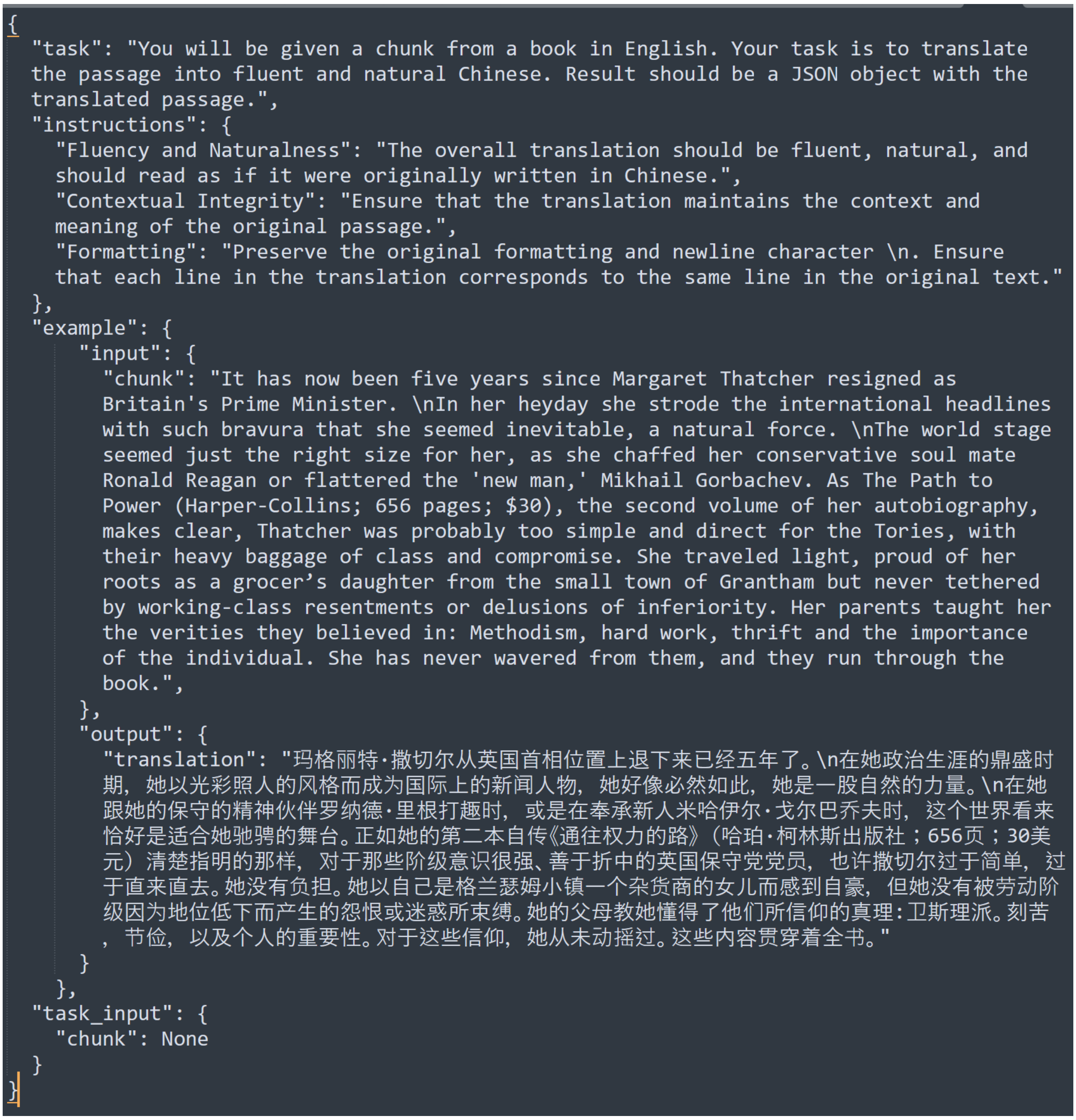
```
You will be given a chunk from a book in English. Your task is to translate a specific book
chunk from English into fluent and natural Chinese.

Examples:
Chunk:
It has now been five years since Margaret Thatcher resigned as Britain's Prime Minister. In
her heyday she strode the international headlines with such bravura that she seemed
inevitable, a natural force. The world stage seemed just the right size for her, as she
chaffed her conservative soul mate Ronald Reagan or flattered the "new man," Mikhail
Gorbachev.

As The Path to Power (Harper-Collins; 656 pages; $30), the second volume of her
autobiography, makes clear, Thatcher was probably too simple and direct for the Tories,
with their heavy baggage of class and compromise. She traveled light, proud of her roots as
a grocer’s daughter from the small town of Grantham but never tethered by working-class
resentments or delusions of inferiority. Her parents taught her the verities they believed
in: Methodism, hard work, thrift and the importance of the individual. She has never
wavered from them, and they run through the book.
Result:
玛格丽特·撒切尔从英国首相位置上退下来已经五年了。在她政治生涯的鼎盛时期，她以光彩照人的风格而成为国际
上的新闻人物，她好像必然如此，她是一股自然的力量。在她跟她的保守的精神伙伴罗纳德·里根打趣时，或是在奉
承新人米哈伊尔·戈尔巴乔夫时，这个世界看来恰好是适合她驰骋的舞台。

正如她的第二本自传《通往权力的路》（哈珀·柯林斯出版社；656页；30美元）清楚指明的那样，对于那些阶级意识
很强、善于折中的英国保守党党员，也许撒切尔过于简单，过于直来直去。她没有负担。她以自己是格兰瑟姆小镇一个
杂货商的女儿而感到自豪，但她没有被劳动阶级因为地位低下而产生的怨恨或迷惑所束缚。她的父母教她懂得了他们
所信仰的真理:卫斯理派。刻苦，节俭，以及个人的重要性。对于这些信仰，她从未动摇过。这些内容贯穿着全书。

Below is the actual task to translate:

Chunk:
$BOOK_CHUNK$

Result:
```

FIGURE A.6: baseline chunk translation prompt



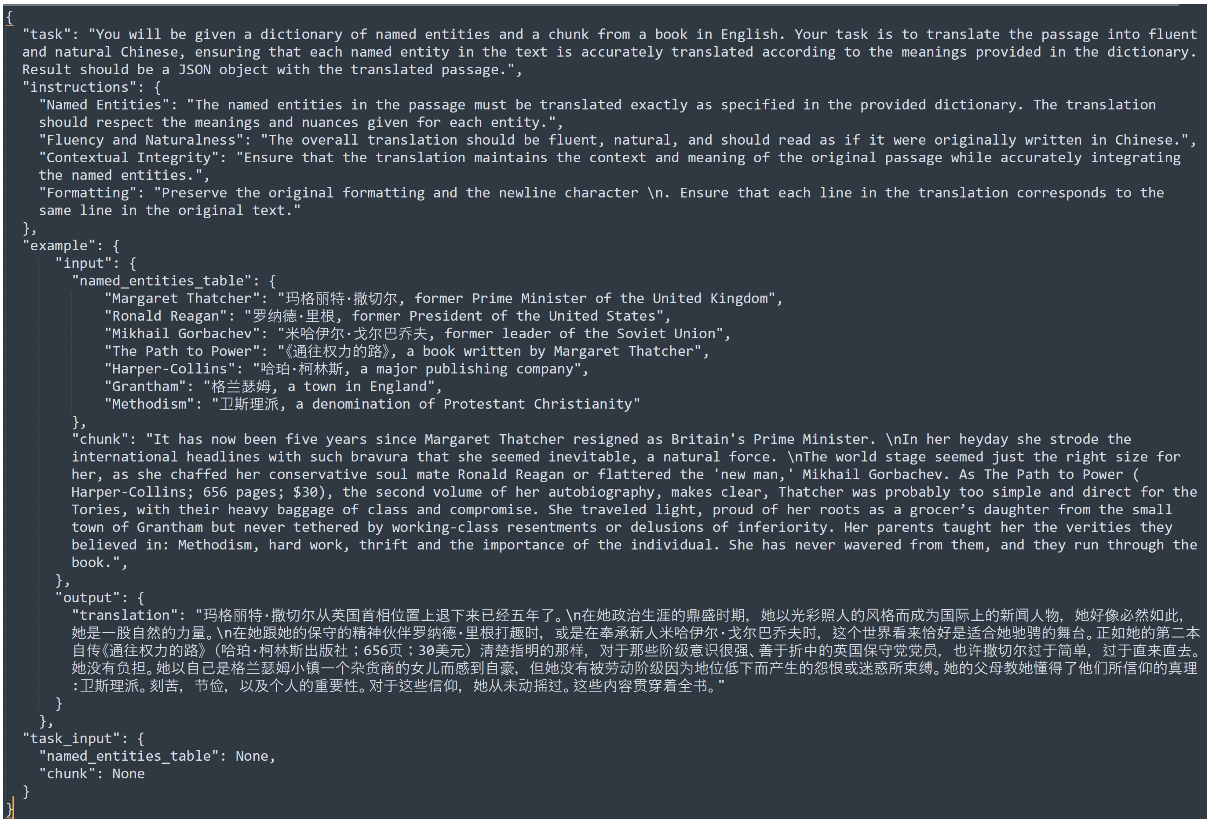
```
{
        "instruction": "You are given a dictionary containing named entities in
        English, their descriptions, and sentences in both English and Chinese.
        Align the named entities in English with their corresponding translations
        in Chinese based on the context provided in the sentences. Return a JSON
        dictionary with the named entities in English as keys, and corresponding
        translation as values.",
        "one_shot_example": {
            "input": {
                "entities": {
                    "Young Master Han": {
                        "desc": "person",
                        "source": [
                            ["Young Master Han carefully planned their actions.",
                            "韩家公子盘算着一路的进程。"],
                            ["Young Master Han had a way of handling difficult
                            situations.", "韩家公子有一套处理棘手问题的方法。"]
                        ]
                    },
                    "Gu Fei": {
                        "desc": "person",
                        "source": [
                            ["Gu Fei was a school teacher.",
                            "顾飞本是学校的一名老师。"],
                            ["Gu Fei, in his long mage robe, stood in the center
                            of the academy grounds.",
                            "顾飞，穿着一袭法师长袍，站在学校广场的中央。"]
                        ]
                    }
                }
            },
            "output": {"Young Master Han": "韩家公子","Gu Fei": "顾飞"}
        },
        "data": None
    }
```

FIGURE A.7: GPT prompt for finding golden standard named entity translation



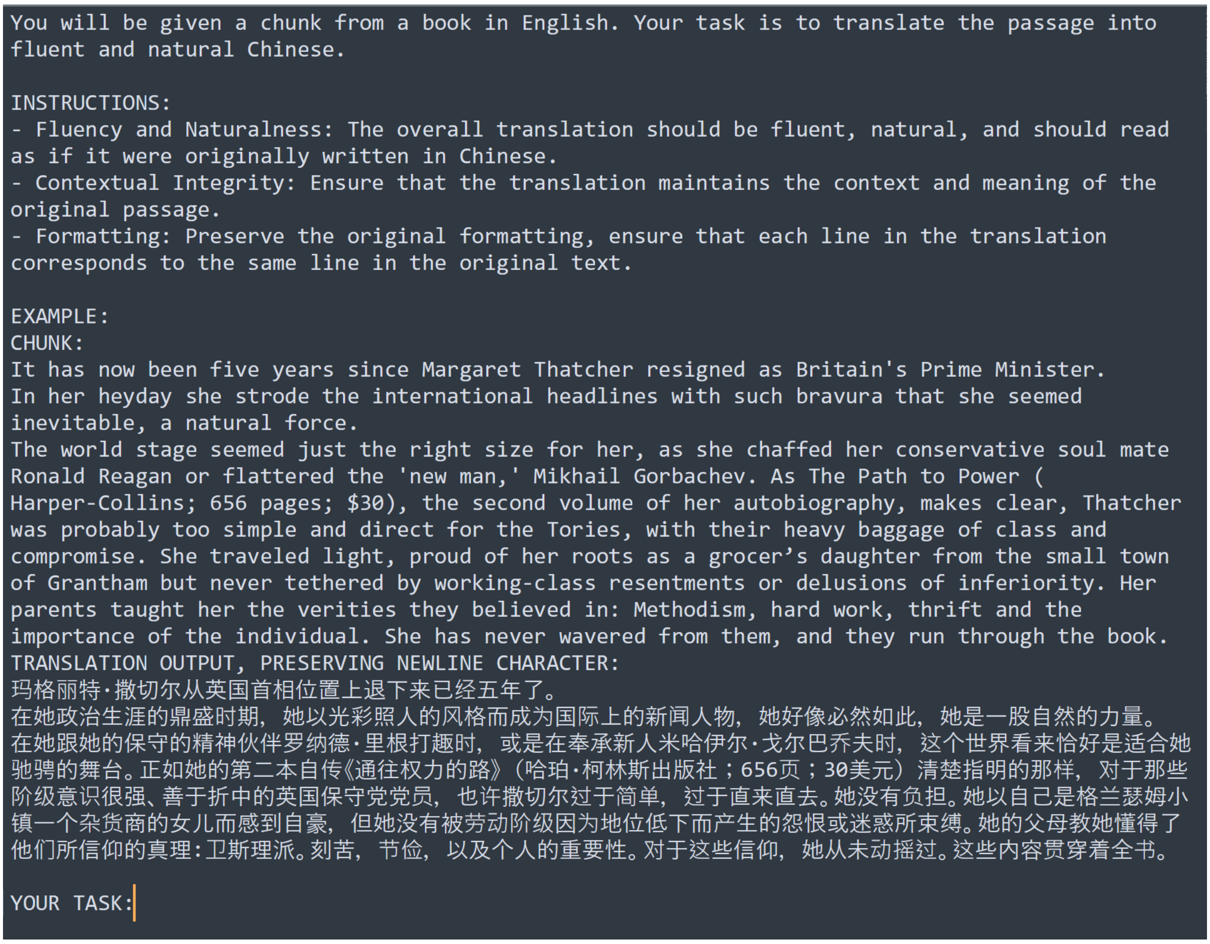
```
{
  "task": "You will be given a chunk from a book in English. Your task is to translate
  the passage into fluent and natural Chinese. Result should be a JSON object with the
  translated passage.",
  "instructions": {
    "Fluency and Naturalness": "The overall translation should be fluent, natural, and
    should read as if it were originally written in Chinese.",
    "Contextual Integrity": "Ensure that the translation maintains the context and
    meaning of the original passage.",
    "Formatting": "Preserve the original formatting and newline character \n. Ensure
    that each line in the translation corresponds to the same line in the original text."
  },
  "example": {
      "input": {
        "chunk": "It has now been five years since Margaret Thatcher resigned as
        Britain's Prime Minister. \nIn her heyday she strode the international headlines
        with such bravura that she seemed inevitable, a natural force. \nThe world stage
        seemed just the right size for her, as she chaffed her conservative soul mate
        Ronald Reagan or flattered the 'new man,' Mikhail Gorbachev. As The Path to
        Power (Harper-Collins; 656 pages; $30), the second volume of her autobiography,
        makes clear, Thatcher was probably too simple and direct for the Tories, with
        their heavy baggage of class and compromise. She traveled light, proud of her
        roots as a grocer’s daughter from the small town of Grantham but never tethered
        by working-class resentments or delusions of inferiority. Her parents taught her
        the verities they believed in: Methodism, hard work, thrift and the importance
        of the individual. She has never wavered from them, and they run through the
        book.",
      },
      "output": {
        "translation": "玛格丽特·撒切尔从英国首相位置上退下来已经五年了。\n在她政治生涯的鼎盛时
        期，她以光彩照人的风格而成为国际上的新闻人物，她好像必然如此，她是一股自然的力量。\n在她
        跟她的保守的精神伙伴罗纳德·里根打趣时，或是在奉承新人米哈伊尔·戈尔巴乔夫时，这个世界看来
        恰好是适合她驰骋的舞台。正如她的第二本自传《通往权力的路》（哈珀·柯林斯出版社；656页；30美
        元）清楚指明的那样，对于那些阶级意识很强、善于折中的英国保守党党员，也许撒切尔过于简单，过
        于直来直去。她没有负担。她以自己是格兰瑟姆小镇一个杂货商的女儿而感到自豪，但她没有被劳动阶
        级因为地位低下而产生的怨恨或迷惑所束缚。她的父母教她懂得了他们所信仰的真理：卫斯理派。刻苦
        ，节俭，以及个人的重要性。对于这些信仰，她从未动摇过。这些内容贯穿着全书。"
      }
    },
  "task_input": {
    "chunk": None
  }
}
```

FIGURE A.8: JSON style prompt for translation (baseline)



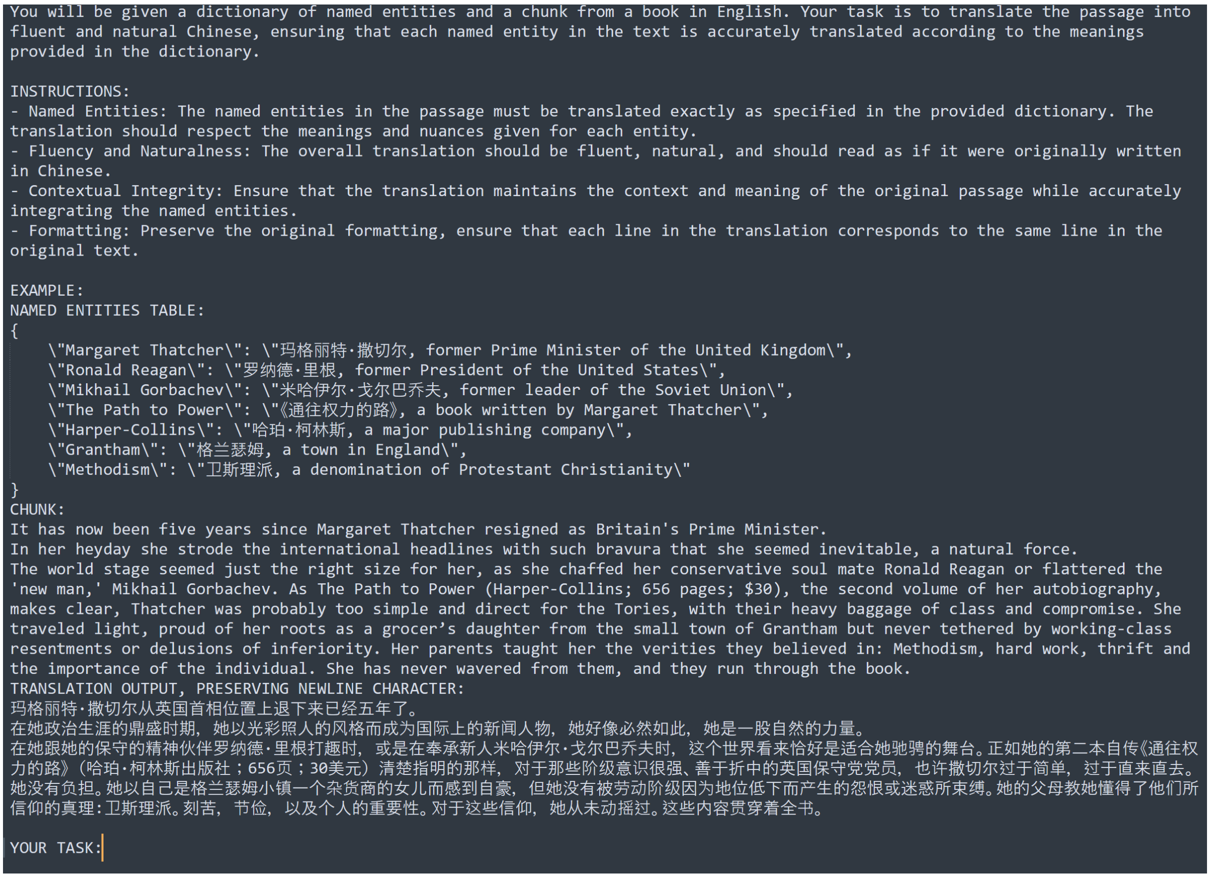
```
{
  "task": "You will be given a dictionary of named entities and a chunk from a book in English. Your task is to translate the passage into fluent
  and natural Chinese, ensuring that each named entity in the text is accurately translated according to the meanings provided in the dictionary.
  Result should be a JSON object with the translated passage.",
  "instructions": {
    "Named Entities": "The named entities in the passage must be translated exactly as specified in the provided dictionary. The translation
    should respect the meanings and nuances given for each entity.",
    "Fluency and Naturalness": "The overall translation should be fluent, natural, and should read as if it were originally written in Chinese.",
    "Contextual Integrity": "Ensure that the translation maintains the context and meaning of the original passage while accurately integrating
    the named entities.",
    "Formatting": "Preserve the original formatting and the newline character \n. Ensure that each line in the translation corresponds to the
    same line in the original text."
  },
  "example": {
      "input": {
        "named_entities_table": {
            "Margaret Thatcher": "玛格丽特·撒切尔, former Prime Minister of the United Kingdom",
            "Ronald Reagan": "罗纳德·里根, former President of the United States",
            "Mikhail Gorbachev": "米哈伊尔·戈尔巴乔夫, former leader of the Soviet Union",
            "The Path to Power": "《通往权力的路》, a book written by Margaret Thatcher",
            "Harper-Collins": "哈珀·柯林斯, a major publishing company",
            "Grantham": "格兰瑟姆, a town in England",
            "Methodism": "卫斯理派, a denomination of Protestant Christianity"
        },
        "chunk": "It has now been five years since Margaret Thatcher resigned as Britain's Prime Minister. \nIn her heyday she strode the
        international headlines with such bravura that she seemed inevitable, a natural force. \nThe world stage seemed just the right size for
        her, as she chaffed her conservative soul mate Ronald Reagan or flattered the 'new man,' Mikhail Gorbachev. As The Path to Power (
        Harper-Collins; 656 pages; $30), the second volume of her autobiography, makes clear, Thatcher was probably too simple and direct for the
        Tories, with their heavy baggage of class and compromise. She traveled light, proud of her roots as a grocer’s daughter from the small
        town of Grantham but never tethered by working-class resentments or delusions of inferiority. Her parents taught her the verities they
        believed in: Methodism, hard work, thrift and the importance of the individual. She has never wavered from them, and they run through the
        book.",
      },
      "output": {
        "translation": "玛格丽特·撒切尔从英国首相位置上退下来已经五年了。\n在她政治生涯的鼎盛时期，她以光彩照人的风格而成为国际上的新闻人物，她好像必然如此，
        她是一股自然的力量。\n在她跟她的保守的精神伙伴罗纳德·里根打趣时，或是在奉承新人米哈伊尔·戈尔巴乔夫时，这个世界看来恰好是适合她驰骋的舞台。正如她的第二本
        自传《通往权力的路》（哈珀·柯林斯出版社；656页；30美元）清楚指明的那样，对于那些阶级意识很强、善于折中的英国保守党党员，也许撒切尔过于简单，过于直来直去。
        她没有负担。她以自己是格兰瑟姆小镇一个杂货商的女儿而感到自豪，但她没有被劳动阶级因为地位低下而产生的怨恨或迷惑所束缚。她的父母教她懂得了他们所信仰的真理
        :卫斯理派。刻苦，节俭，以及个人的重要性。对于这些信仰，她从未动摇过。这些内容贯穿着全书。"
      }
    },
  "task_input": {
    "named_entities_table": None,
    "chunk": None
  }
}
```

FIGURE A.9: JSON style prompt for translation (TermBase augment)



```
You will be given a chunk from a book in English. Your task is to translate the passage into
fluent and natural Chinese.

INSTRUCTIONS:
- Fluency and Naturalness: The overall translation should be fluent, natural, and should read
as if it were originally written in Chinese.
- Contextual Integrity: Ensure that the translation maintains the context and meaning of the
original passage.
- Formatting: Preserve the original formatting, ensure that each line in the translation
corresponds to the same line in the original text.

EXAMPLE:
CHUNK:
It has now been five years since Margaret Thatcher resigned as Britain's Prime Minister.
In her heyday she strode the international headlines with such bravura that she seemed
inevitable, a natural force.
The world stage seemed just the right size for her, as she chaffed her conservative soul mate
Ronald Reagan or flattered the 'new man,' Mikhail Gorbachev. As The Path to Power (
Harper-Collins; 656 pages; $30), the second volume of her autobiography, makes clear, Thatcher
was probably too simple and direct for the Tories, with their heavy baggage of class and
compromise. She traveled light, proud of her roots as a grocer’s daughter from the small town
of Grantham but never tethered by working-class resentments or delusions of inferiority. Her
parents taught her the verities they believed in: Methodism, hard work, thrift and the
importance of the individual. She has never wavered from them, and they run through the book.
TRANSLATION OUTPUT, PRESERVING NEWLINE CHARACTER:
玛格丽特·撒切尔从英国首相位置上退下来已经五年了。
在她政治生涯的鼎盛时期，她以光彩照人的风格而成为国际上的新闻人物，她好像必然如此，她是一股自然的力量。
在她跟她的保守的精神伙伴罗纳德·里根打趣时，或是在奉承新人米哈伊尔·戈尔巴乔夫时，这个世界看来恰好是适合她
驰骋的舞台。正如她的第二本自传《通往权力的路》（哈珀·柯林斯出版社；656页；30美元）清楚指明的那样，对于那些
阶级意识很强、善于折中的英国保守党党员，也许撒切尔过于简单，过于直来直去。她没有负担。她以自己是格兰瑟姆小
镇一个杂货商的女儿而感到自豪，但她没有被劳动阶级因为地位低下而产生的怨恨或迷惑所束缚。她的父母教她懂得了
他们所信仰的真理:卫斯理派。刻苦，节俭，以及个人的重要性。对于这些信仰，她从未动摇过。这些内容贯穿着全书。

YOUR TASK:
```

FIGURE A.10: String style prompt for translation (baseline)



```
You will be given a dictionary of named entities and a chunk from a book in English. Your task is to translate the passage into
fluent and natural Chinese, ensuring that each named entity in the text is accurately translated according to the meanings
provided in the dictionary.

INSTRUCTIONS:
- Named Entities: The named entities in the passage must be translated exactly as specified in the provided dictionary. The
translation should respect the meanings and nuances given for each entity.
- Fluency and Naturalness: The overall translation should be fluent, natural, and should read as if it were originally written
in Chinese.
- Contextual Integrity: Ensure that the translation maintains the context and meaning of the original passage while accurately
integrating the named entities.
- Formatting: Preserve the original formatting, ensure that each line in the translation corresponds to the same line in the
original text.

EXAMPLE:
NAMED ENTITIES TABLE:
{
    \"Margaret Thatcher\": \"玛格丽特·撒切尔, former Prime Minister of the United Kingdom\",
    \"Ronald Reagan\": \"罗纳德·里根, former President of the United States\",
    \"Mikhail Gorbachev\": \"米哈伊尔·戈尔巴乔夫, former leader of the Soviet Union\",
    \"The Path to Power\": \"《通往权力的路》, a book written by Margaret Thatcher\",
    \"Harper-Collins\": \"哈珀·柯林斯, a major publishing company\",
    \"Grantham\": \"格兰瑟姆, a town in England\",
    \"Methodism\": \"卫斯理派, a denomination of Protestant Christianity\"
}
CHUNK:
It has now been five years since Margaret Thatcher resigned as Britain's Prime Minister.
In her heyday she strode the international headlines with such bravura that she seemed inevitable, a natural force.
The world stage seemed just the right size for her, as she chaffed her conservative soul mate Ronald Reagan or flattered the
'new man,' Mikhail Gorbachev. As The Path to Power (Harper-Collins; 656 pages; $30), the second volume of her autobiography,
makes clear, Thatcher was probably too simple and direct for the Tories, with their heavy baggage of class and compromise. She
traveled light, proud of her roots as a grocer’s daughter from the small town of Grantham but never tethered by working-class
resentments or delusions of inferiority. Her parents taught her the verities they believed in: Methodism, hard work, thrift and
the importance of the individual. She has never wavered from them, and they run through the book.
TRANSLATION OUTPUT, PRESERVING NEWLINE CHARACTER:
玛格丽特·撒切尔从英国首相位置上退下来已经五年了。
在她政治生涯的鼎盛时期，她以光彩照人的风格而成为国际上的新闻人物，她好像必然如此，她是一股自然的力量。
在她跟她的保守的精神伙伴罗纳德·里根打趣时，或是在奉承新人米哈伊尔·戈尔巴乔夫时，这个世界看来恰好是适合她驰骋的舞台。正如她的第二本自传《通往权
力的路》（哈珀·柯林斯出版社；656页；30美元）清楚指明的那样，对于那些阶级意识很强、善于折中的英国保守党党员，也许撒切尔过于简单，过于直来直去。
她没有负担。她以自己是格兰瑟姆小镇一个杂货商的女儿而感到自豪，但她没有被劳动阶级因为地位低下而产生的怨恨或迷惑所束缚。她的父母教她懂得了他们所
信仰的真理:卫斯理派。刻苦，节俭，以及个人的重要性。对于这些信仰，她从未动摇过。这些内容贯穿着全书。

YOUR TASK:
```

FIGURE A.11: String style prompt for translation (TermBase augment)

Appendix B

Code

1. Experiment code on Github:

   https://github.com/Catkamakura/honor_thesis

2. Experiment code in Colab:

   https://colab.research.google.com/drive/1AtJwuEL2dyijdiYifP4kjiPg-bJ7Hwny?usp=sharing

LISTING B.1: Python code for cTT calculation

```python
def LTCR(lst):
    if len(lst) < 2:
        return 1
    C = len(lst)*(len(lst)-1)/2
    pairs_count = 0
    n = len(lst)
    for i in range(n):
        for j in range(i+1, n):
            if lst[i] == lst[j]:
                pairs_count += 1
    return pairs_count / C
```

Bibliography

Agrawal, S., Zhou, C., Lewis, M., Zettlemoyer, L., and Ghazvininejad, M. (2022). In-context Examples Selection for Machine Translation. arXiv:2212.02437 [cs].

Babych, B. and Hartley, A. (2003). Improving Machine Translation Quality with Automatic Named Entity Recognition. In *Proceedings of the 7th International EAMT workshop on MT and other language technology tools, Improving MT through other language technology tools, Resource and tools for building MT at EACL 2003*.

Bahdanau, D., Cho, K., and Bengio, Y. (2016). Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473 [cs, stat].

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 [cs].

Cai, D., Wang, Y., Li, H., Lam, W., and Liu, L. (2021). Neural Machine Translation with Monolingual Translation Memory. arXiv:2105.11269 [cs].

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. (2022). PaLM: Scaling Language Modeling with Pathways. arXiv:2204.02311 [cs].

Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., Li, L., and Sui, Z. (2023). A Survey on In-context Learning. arXiv:2301.00234 [cs].

Dou, Z.-Y. and Neubig, G. (2021). Word Alignment by Fine-tuning Embeddings on Parallel Corpora. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Ghazvininejad, M., Gonen, H., and Zettlemoyer, L. (2023). Dictionary-based Phrase-level Prompting of Large Language Models for Machine Translation. arXiv:2302.07856 [cs].

Guerreiro, N. M., Alves, D., Waldendorf, J., Haddow, B., Birch, A., Colombo, P., and Martins, A. F. T. (2023). Hallucinations in Large Multilingual Translation Models. arXiv:2303.16104 [cs].

He, Z., Liang, T., Jiao, W., Zhang, Z., Yang, Y., Wang, R., Tu, Z., Shi, S., and Wang, X. (2023). Exploring Human-Like Translation Strategy with Large Language Models. arXiv:2305.04118 [cs].

Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., and Awadalla, H. H. (2023). How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation. arXiv:2302.09210 [cs].

Karpinska, M. and Iyyer, M. (2023). Large language models effectively leverage document-level context for literary translation, but critical errors persist. arXiv:2304.03245 [cs].

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

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., and Chen, W. (2021). What Makes Good In-Context Examples for GPT-3? arXiv:2101.06804 [cs].

Liu, L. and Zhu, M. (2023). Bertalign: Improved word embedding-based sentence alignment for Chinese–English parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634.

Lyu, X., Li, J., Gong, Z., and Zhang, M. (2021). Encouraging Lexical Translation Consistency for Document-Level Neural Machine Translation. In Moens, M.-F., Huang, X., Specia, L., and Yih, S. W.-t., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3265–3277, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Marcus, G. (2020). The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence. arXiv:2002.06177 [cs].

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. (2022). Rethinking the Role of Demonstrations: What Makes In-Context Learning Work? arXiv:2202.12837 [cs].

Mu, Y., Reheman, A., Cao, Z., Fan, Y., Li, B., Li, Y., Xiao, T., Zhang, C., and Zhu, J. (2023). Augmenting Large Language Model Translators via Translation Memories. arXiv:2305.17367 [cs].

OpenAI (2023). GPT-4 Technical Report. arXiv:2303.08774 [cs].

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., editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. arXiv:2003.07082 [cs].

Rei, R., Stewart, C., Farinha, A. C., and Lavie, A. (2020). COMET: A Neural Framework for MT Evaluation. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Team, Q. (2024). Qwen2.5: A Party of Foundation Models! Section: blog.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention Is All You Need. arXiv:1706.03762 [cs].

Vilar, D., Freitag, M., Cherry, C., Luo, J., Ratnakar, V., and Foster, G. (2023). Prompting PaLM for Translation: Assessing Strategies and Performance. arXiv:2211.09102 [cs].

Wang, L., Lyu, C., Ji, T., Zhang, Z., Yu, D., Shi, S., and Tu, Z. (2023). Document-Level Machine Translation with Large Language Models. In Bouamor, H., Pino, J., and Bali, K., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

Wu, Z., Wang, Y., Ye, J., and Kong, L. (2023). Self-Adaptive In-Context Learning: An Information Compression Perspective for In-Context Example Selection and Ordering. arXiv:2212.10375 [cs].

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. (2022). An Explanation of In-context Learning as Implicit Bayesian Inference. arXiv:2111.02080 [cs].

Zhang, X., Rajabi, N., Duh, K., and Koehn, P. (2023). Machine Translation with Large Language Models: Prompting, Few-shot Learning, and Fine-tuning with QLoRA. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C., editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.

Zheng, X., Zhang, Z., Guo, J., Huang, S., Chen, B., Luo, W., and Chen, J. (2021). Adaptive Nearest Neighbor Machine Translation. arXiv:2105.13022 [cs].