Laughing Across Languages! A Psychological Theory-driven Humour Translation Approach with Large Language Models

Anonymous ACL submission

Abstract

Humour translation plays a vital role that can serve as a bridge between different cultures, fostering understanding and communication. However, although most existing Large Language Models (LLMs) are capable of general translation tasks, they still struggle with humour translation, especially for linguistic interference and lacking humour in translated text. In this paper, we propose a Humour Decomposition Mechanism (HDM) that utilises Chain-of-Thought (CoT) to imitate the ability of the human thought process, stimulating LLMs to optimise the readability of translated humorous texts. Moreover, we integrate humour theory in HDM to further enhance the humorous elements in the translated text. Our experimental evaluation involves both automatic and human evaluation on open-source humour datasets, demonstrating that our method effectively enhances the quality of humour translation, showing an average improvement of 7.75% in humour, 2.81% in fluency, and 6.13% in coherency. Finally, we release a new humour Chinese dataset which has been translated from English using HDM.

1 Introduction

Humour plays an important role in human interaction. Humour studies can actually gain greater insight into the linguistic, social and psychological factors of humour (Zabalbeascoa, 2005). A comprehensive understanding of humour necessitates a deep grasp of both semantic information and cultural background (Chen et al., 2024b) and effective humour translation serves as a bridge across cultural divides, facilitating communication and fostering cross-cultural understanding (Vandaele, 2016). Pym (2023) mentions that the study of humour translation can enhance the understanding of language transfer and the process of meaning reconstruction, while enriching the translation theories, especially for dynamic equivalence and functionalist translation strategies. Moreover, an effective humour translation strategy can accurately convey its intended humorous effect in the target language (Zabalbeascoa, 2005) and contribute to advancements in general translation research.

Nida (1964) emphasises two fundamental approaches to translation: formal equivalence, which prioritizes literal translation, and dynamic equivalence, which focuses on emotional or contextual translation. However, the majority of existing studies focus on literal translation, with limited research exploring emotional translation, particularly in the context of humour. Chen et al. (2022) use cross-language transfer to enable zero-shot neural machine translation and Wang et al. (2022a) explore a more efficient kNN-MT for translation. With the advent of large language models (LLMs) such as ChatGPT[1] and GPT-4 (Achiam et al., 2023), translation has become a prominent domain where LLMs demonstrate remarkable capacity and competence (Zhang et al., 2023; Karpinska and Iyyer, 2023; Lu et al., 2023; Jiao et al., 2023; Agrawal et al., 2022; Vilar et al., 2022; He et al., 2024). However, they still lack proficiency in humour translation in some cases. In Figure 1a, for example, the punchline of the joke is "Invisibull". Traditional translation often results in the loss of original humour and has noticeable language interference issues.

We claim that humour loss is a challenge in humour translation. Due to linguistic and cultural barriers, humour translation often results in the loss of humour in the translated content (Xia et al., 2023). The reason is that jokes often rely on extensive knowledge and common sense, and the punchline is usually hidden in the semantics of the sentence, such as cultural context, wordplay, and metaphorical expressions. These elements are challenging to identify and translate accurately (Hasan et al.,

[1] https://chat.openai.com/chat

2021), which weakens the humour of the joke to some extent. Additionally, the issue of linguistic interference is a factor in humour translation (Hopkinson, 2007), which is a non-standard version of the target language in the product of translation. Ma and Cheung (2020) indicates that linguistic interference is linked to reduced lexical variety and less cohesive discourse, while the traditional method of translation usually involves merely a linear arrangement of words or phrases (Gambier, 2016), which can result in a lack of fluency and coherence in the translated text. This requires a process that can provide a human thinking process to reconstruct the translated text.

Therefore, to address the challenge of humour translation across different languages, we propose a novel Humour Decomposition Mechanism (HDM) to improve linguistic interference, which introduces a three-step paradigm through the Chain-of-Thoughts (CoT) prompting method (Wei et al., 2022; Zhang et al., 2022; Wang et al., 2022b) by utilising LLMs: (1) mining intrinsic knowledge related to the joke; (2) translating the intrinsic knowledge text; and (3) constructing a new joke based on the translated content. This method mimics a human thinking process for understanding, translating and generating to reconstruct the translated text. Furthermore, to enhance the humour in translated texts, we integrate humour theory into intrinsic knowledge by defining corresponding topics, angles, and punchlines. This approach enables the model to perform humour translations effectively based on the mined knowledge.

We assess our approach both in automatic and human evaluation. For automatic evaluation, we use the *Estimation Metric Based Assessment* (GEMBA) (Kocmi and Federmann, 2023), a type of LLM evaluation, to assess humour, fluency and coherence. For human evaluation, we design a general Five-point Likert Scale evaluation to assess the quality of source language jokes and target translation jokes in humour, coherency and fluency. Experimental results reveal that our method is demonstrably superior to existing solutions, showing an average improvement of 7.75% in humour, 2.81% in fluency, and 6.13% in coherency from English to Chinese. These findings indicate that the approach effectively mitigates humour loss and linguistic interference. Finally, we utilize HDM to generate a new tiny translation dataset from English to Chinese, providing innovative approaches for extending the humour dataset. Overall, the main contributions are summarized as follows:

- We propose an efficient Humour Decomposition Mechanism to guide LLMs to translate jokes, mimicking the human thought process.
- We make the first attempt to incorporate the Psychological theory of constructing humour into the Chain-of-Thought process to improve the humour factors.
- Our approach provides the potential method of extending the dataset and contributes a new Chinese joke translation dataset from English.

2 Methodology

Figure 1b illustrates an overview of the Humour Decomposition Mechanism. Instead of directly asking LLMs for the final translation result, we hope that the LLMs can analyze the latent humour interpretations and intrinsic knowledge before translating the jokes, and then generate the translated jokes based on this. We present two key contributions in this section.

2.1 Humour Decomposition Mechanism

We design three-step paradigm using Chain-of-Thought (CoT) prompting, which mimics the human thought process in solving complex reasoning tasks (Wei et al., 2022; Wang et al., 2022b), to enhance humour translation outcomes.

2.1.1 Humour Decomposition

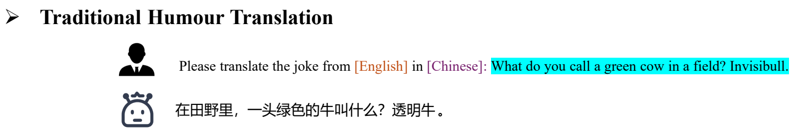
Humour decomposition is one of the important cores for HDM. Specifically, our approach initiates the LLM with a specific task of joke analysis. The request is formulated as follows:

> You are a humour assistant. Please analyze the following joke: [Given joke $\mathcal{L}_i$]

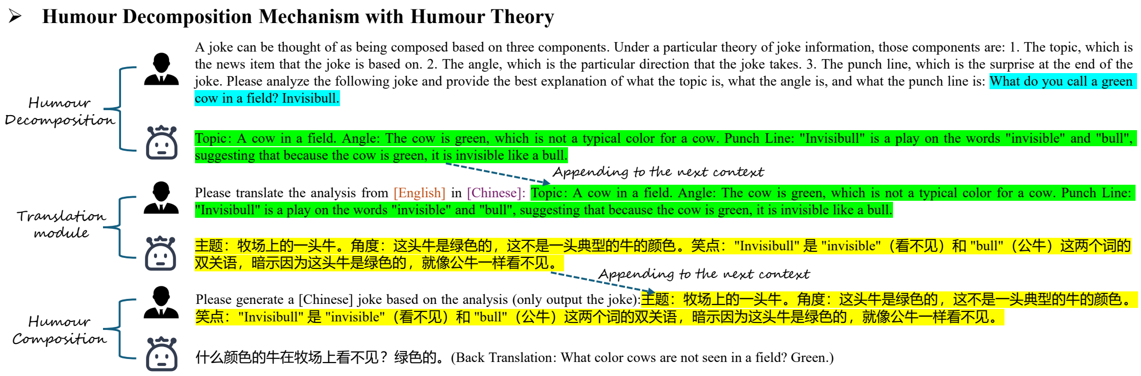
Given a joke $\mathcal{L}_i$, we first claim the role of LLM in humour. Furthermore, we introduce an analysis process to generate the sequence of corresponding knowledge a, which is organized into the final analysis $\mathcal{A}$. The formulation of our *Humour Decomposition* method can be expressed as follows:

$$\mathcal{A}_i = \arg\max p\,(a \mid \mathcal{L}_i) \quad (1)$$

where $\mathcal{L}_i$ and $\mathcal{A}_i$ denote the i_{th} joke and its final analysis.



(a) Traditional translation prompting.



(b) The overview of Humour Decomposition Mechanism.

Figure 1: Comparison of the traditional translation and our HDM, taking the translation from English to Chinese as an example. Lightblue represents the original English joke. Green indicates the analysis in English and yellow corresponds to the Chinese translation of the analysis.

2.1.2 Translation module

After achieving *Humour Decomposition*, we use *Translation Module* to convert the source language analysis into the target language analysis. To illustrate, given the analysis $\mathcal{A}_i$ and the type of source language $\mathcal{S}$, we prompt the LLMs to translate $\mathcal{A}_i$ into target language $\mathcal{T}$, with the prompt defined as:

> Please translate the analysis from [source language $\mathcal{S}$] into [target language $\mathcal{T}$]: [text $\mathcal{A}_i$]

Formally, the translation is determined as:

$$\mathcal{A}_i' = \arg\max p\left(a' \mid \mathcal{A}_i, \mathcal{S}, \mathcal{T}, \right) \tag{2}$$

where $\mathcal{A}_i'$ represents the final translation of the analysis, generated from all potential translation results a'.

2.1.3 Humour Composition

Once the translation is generated, we further propose *Humour Composition* to facilitate the generation of jokes. Given the translation version of the analysis, we design the prompt to make LLMs generate the joke of the target language. This is the structure of the prompt:

> Please generate a [target language $\mathcal{T}$] joke based on the analysis: [text $\mathcal{A}_i'$]

Formally, the humour composition can be defined as:

$$\mathcal{F} = \arg\max p\left(f \mid \mathcal{A}_i', \mathcal{T}\right) \tag{3}$$

where $\mathcal{F}$ is the final generation of the target language joke, generated from all potential generation results f.

2.2 Integrating Humour Theory

In this section, we incorporate humour theory inspired by (Toplyn, 2014) to enhance humour factors. The basic structure of the humorous text consists of the topic $\mathcal{X}$, angle $\mathcal{Y}$ and punchline $\mathcal{Z}$. The topic $\mathcal{X}$ is the news item that the joke is based on and the angle $\mathcal{Y}$ is the particular direction that the joke takes, while the punchline $\mathcal{Z}$ which is the surprise at the end of the joke. Therefore, the *Humour Decomposition* module in HDM can be further improved as follows:

> You are a humour assistant. A joke can be thought of as being composed based on three components. Under a particular theory of joke information, those components are:

1. The topic, which is the news item that the joke is based on.
2. The angle, which is the particular direction that the joke takes.
3. The punchline, which is the surprise at the end of the joke.

Similarly, with *Humour Decomposition*, we first claim the LLM's role in humour. Then, we describe the components under the particular theory and give these components some details. Finally, we provide an instruction to format the model's outputs, which are defined as:

Please analyze the following joke and provide the best explanation of what the topic is, what the angle is, and what the punchline is: [Given joke $\mathcal{L}_i$]

Formally, The improved formulation of the Humour Decomposition can be expressed as follows:

$$\mathcal{A}_i = \arg\max p\left(\mathcal{X}_i, \mathcal{Y}_i, \mathcal{Z}_i \mid \mathcal{L}_i\right) \tag{4}$$

where $\mathcal{A}_i$ denotes the analysis of the i_{th} joke, including the con-cat of topic $\mathcal{X}_i$, angle $\mathcal{Y}_i$ and punchline $\mathcal{Z}_i$.

HDM leverages the advanced generative capabilities of LLMs (Hagos et al., 2024) to reconstruct humour translation, overcoming the limitations of traditional translation methods, which are often constrained by linear word or phrase arrangements and linguistic interference, to improve the fluency and coherency of jokes. Additionally, the integration of humour theory defines the general structure of joke composition within the prompts, enabling the large language model to better comprehend background and punchline information. It theoretically enhances the LLM's ability to generate more humorous jokes, and we will also be demonstrated in our experiments.

3 Dataset Generation

In this section, we translate the English humour dataset and construct the Chinese humour dataset by using the Humour Decomposition Mechanism.

3.1 Humour Corpus Preprocessing

To prepare our dataset, we choose the public dataset of Short Jokes (Moudgil, 2016) as raw data. Before proceeding with the formal tasks, we observe that some jokes in the dataset contain offensive and aggressive content. Therefore, we need to remove these instances first. The binary classification is used to accomplish this goal. We use SemEval 2021 (García-Díaz and Valencia-García, 2021) as the dataset for joke offense detection with a total of 6000 training data and 3000 validating data. Then, we train LoRA (Hu et al., 2021) for LLaMA3 (Dubey et al., 2024) to conduct the task of binary classification.

3.2 Data Translation

In this section, we employ GPT4-Turbo in conjunction with HDM to translate the source language humour dataset. Initially, we perform offensive corpus detection on the source data. Based on the model trained in the previous step, we select 2000 jokes from the Short Jokes Dataset [2] after filtering out harmful content. Subsequently, we conduct the humour translation task. Specifically, following the methodology outlined earlier, each filtered text will be fed into GPT4 in a fixed format and generate the final results.

3.3 Dataset Construction

The structure of the humour dataset is as follows:

JokeDataset = (ID, Content, Topic, Angle, Punchline, DataSource, Link, Original Version). We encapsulate the data in a semi-structured JSON format.

In our dataset, the *Topic*, *Angle*, and *Punchline* constitute the intermediary stage as described in the Methodology section. These elements are decomposed and translated by LLMs from the source language jokes. The *Content*, *DataSource* and *Link* provide the translation joke, the name of the dataset and its source link. We also include the *Original Version* as an original reference text. All details can be found in Appendix.

4 Experiments

4.1 Experiment Setup and Baselines

We select four representative state-of-the-art LLMs from the Chatbot Arena Leaderboard (Zheng et al., 2023) as backbone references for our study: Gemini1.5-Pro (Team et al., 2024), Yi-Large (AI et al., 2024), GPT3.5-Turbo and GPT4-Turbo. Additionally, we use Zero-shot (Hendy et al., 2023), DUAL-REFLECT (Chen et al., 2024a) and MAPS

[2] https://www.kaggle.com/datasets/thedevastator/short-jokes-dataset

LLM	Method	SQM-H	STAR-H	SQM-F	STAR-F	SQM-C	STAR-C
Gemini1.5-Pro	Z-shot (Hendy et al., 2023)	49.82	2.53	96.74	4.81	89.30	4.50
	DUAL (Chen et al., 2024a)	50.86	2.69	92.98	4.46	84.74	4.18
	MAPS (He et al., 2024)	57.98	3.01	96.35	4.74	89.95	4.48
	HDM	**63.80**	**3.19**	**98.54**	**4.93**	**94.27**	**4.74**
Yi-Large	Z-shot (Hendy et al., 2023)	53.40	2.57	95.37	4.76	86.58	4.42
	DUAL (Chen et al., 2024a)	56.34	2.85	94.30	4.63	87.01	4.34
	MAPS (He et al., 2024)	58.08	2.94	95.24	4.67	87.09	4.36
	HDM	**67.99**	**3.22**	**98.99**	**4.95**	**95.56**	**4.85**
GPT3.5-Turbo	Z-shot (Hendy et al., 2023)	50.03	2.52	94.33	4.72	86.83	4.41
	DUAL (Chen et al., 2024a)	54.63	2.77	92.02	4.48	83.42	4.16
	MAPS (He et al., 2024)	57.66	2.87	94.58	4.59	85.90	4.31
	HDM	**61.73**	**3.05**	**96.07**	**4.80**	**88.75**	**4.49**
GPT4-Turbo	Z-shot (Hendy et al., 2023)	53.20	2.58	94.95	4.76	87.70	4.67
	DUAL (Chen et al., 2024a)	58.33	2.95	91.60	4.43	83.30	4.13
	MAPS (He et al., 2024)	59.34	3.02	95.12	4.68	88.62	4.45
	HDM	**70.54**	**3.45**	**99.45**	**4.99**	**97.73**	**4.96**

Table 1: Main results of the automatic metrics GEMBA-SQM and GEMBA-STARS in humour, fluency and coherency for translating from English to Chinese on the Short Joke Dataset. Both higher evaluation metrics indicate better performance.

(He et al., 2024), which are the state-of-the-art translation approaches, as our baselines. Given budget constraints, we randomly select 500 samples on the Short Jokes Dataset for experiments. Finally, we evaluate their performance by using automatic metrics and manual metrics, respectively.

4.2 Metrics

4.2.1 Automatic metrics.

Since our approach specializes in humorous translation tasks, traditional automatic evaluation methods, such as COMET (Rei et al., 2020) and BLEURT (Sellam et al., 2020), have difficulty evaluating elements like humour. Therefore, inspired by Kocmi and Federmann (2023), we evaluate the final results by using GEMBA which is a GPT4-based metric for generation quality. We choose the open area no-reference metrics GEMBA-SQM and GEMBA-STARS for their superior performance in (Kocmi and Federmann, 2023). Specifically, GEMBA-SQM evaluates scalar quality metrics by dividing the assessment results into several stages, where 0 and 100 represent the lowest and highest scores, respectively. GEMBA-STARS is a classification task based on a one-to-five star ranking, which is a style often used when users are asked to review various services or products (Kocmi and Federmann, 2023). In this section, SQM-H, SQM-F and SQM-C represent GEMBA-SQM metrics and STAR-H, STAR-F and STAR-C represent GEMBA-STARS metrics in humour, fluency and coherency.

To adapt to the evaluation of humour translation in linguistic interference and humour factor, we modify the original translation prompts and use the keywords of humour, coherence and fluency based on Chen et al. (2024b). We report the performance by averaging the results over three runs in each type of experiment. Additionally, Kocmi and Federmann (2023) observe that some answers occasionally fall outside these ranges because of the LLM's hallucination. For example, instead of providing predicted scores, the model occasionally outputs explanations as results. Therefore, we omit the invalid responses and retain only the valid results in this research.

4.2.2 Manual metrics.

Issues with hallucinations in LLMs (Bender et al., 2021), combined with the variability in evaluation results depending on the phrasing of prompts, make it difficult to rely on automatic scores for deriving accurate measures of performance. Thus, we also incorporate five human evaluators and randomly select 40 samples in the manual evaluation process to refine the evaluation criteria [3].

The five-point Likert scale is used to assess the quality of humour generation in three dimensions (Zhang et al., 2020a): (1) Humorous (Is the joke funny?); (2) Fluency and Coherency (Does the joke

[3]Human evaluators correspond to all authors in this paper.

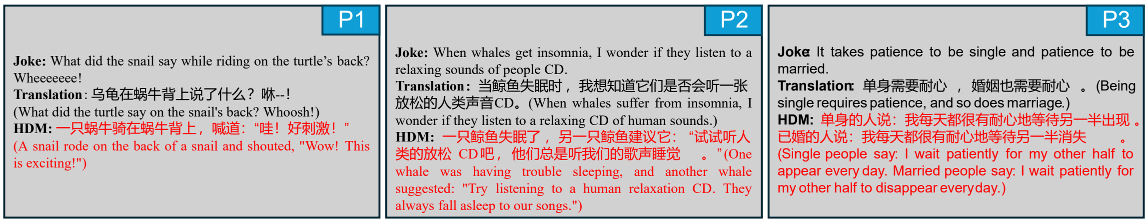


Figure 2: Some correct Chinese cases generated by HDM. We present the original jokes, traditional translations and their back translation and the results of HDM and their back translation.

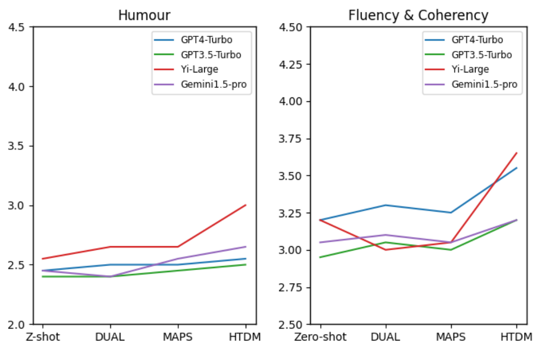


Figure 3: The results of the manual evaluation in humour, fluency and coherency. The x-axis represents the human evaluation categories: Z-shot (Hendy et al., 2023), DUAL (Chen et al., 2024a), MAPS (He et al., 2024) and HDM. The y-axis shows the corresponding evaluation scores.

exhibit overall fluency and coherence?); Each aspect is rated on a scale from 1 to 5, with higher scores indicating better performance, and the final statistical result is the average value of the human evaluation samples. The human evaluation is used to compare the results in the baseline and our method.

4.3 Main Results

The overall results are shown in Table 1 and Figure 3. As shown in Table 1, HDM outperforms all baselines in terms of humour, fluency, and coherency in automatic metrics. This is particularly evident in the translation from English to Chinese in GPT4-Turbo, where the degree of humour improves by an average of 11.2%. These results show that HDM can go beyond the other state-of-the-art translation methods, both enhancing the humour of translated text in humour translation, and also alleviate the problem of linguistic interference.

Table 3 shows the results of human evaluation on the baselines and HDM with the differences in their performance. We observe that our method has some improvements over all the baselines in each metrics. It is worth noting that in the specific evaluation of humor, the Yi-Large model shows superior performance than other LLMs. We also apply Weighted Cohen's Kappa to compute the inter-evaluator agreement. Averaging across all 40 samples and metrics, we achieve a Cohen's Kappa of 0.32, indicating a fair level of agreement as defined by (Landis and Koch, 1977). These results demonstrate the effectiveness of HDM in humour translation.

5 Analysis

5.1 Generality Analysis of HDM

To further investigate the generality of our work, we verify the generality of HDM from two perspectives [4]. MAPS (He et al., 2024) is selected as the baseline for the Generality Analysis based on the comprehensive metrics evaluated in the experiment:

5.1.1 HDM works well on other datasets.

We conduct experiments on other datasets, namely the Question-Answer Jokes dataset (Roznovjak, 2016) and SemEval 2021 (García-Díaz and Valencia-García, 2021). Table 2 shows that HDM can obtain better performance across all LLMs and metrics in different datasets, achieving the improvements of at least 1.84% in humour, 1.7% in fluency and 2.15% in coherency.

5.1.2 HDM works well on other languages.

To better assess the model's generalization capabilities, we conduct the experiments in different languages, including Spanish and German. As shown in Table 3, the experimental results demonstrate that HDM consistently performs significantly well across these languages, for instance, with improvements of 2.75% in humour, 3.25% in fluency, and

[4] Given budget constraints, we have randomly selected 100 samples in each dataset and language.

LLM	SQM-H		SQM-F		SQM-C	
	base	ours	base	ours	base	ours
Question-Answer Joke						
Gemini1.5-Pro	60.00	**64.02**	97.67	**99.53**	85.29	**90.63**
Yi-Large	61.10	**67.30**	96.00	**99.00**	82.30	**93.00**
GPT3.5-Turbo	62.30	**64.14**	95.45	**97.37**	82.12	**87.68**
GPT4-Turbo	64.70	**68.70**	96.40	**99.10**	87.37	**95.05**
SemEval-2021						
Gemini1.5-Pro	61.20	**64.60**	97.90	**99.00**	90.95	**93.10**
Yi-Large	57.90	**67.50**	96.50	**99.20**	92.85	**95.35**
GPT3.5-Turbo	56.30	**66.06**	96.80	**98.50**	88.05	**93.35**
GPT4-Turbo	59.90	**70.10**	96.70	**99.10**	91.70	**97.20**

Table 2: Generality analysis of automatic metric in translating from English to Chinese in different Datasets.

LLM	SQM-H		SQM-F		SQM-C	
	base	ours	base	ours	base	ours
EN=>SP						
Gemini1.5-Pro	59.50	**64.70**	94.00	**97.90**	87.20	**91.60**
Yi-Large	58.25	**68.35**	**96.50**	96.30	89.55	**91.15**
GPT3.5-Turbo	57.90	**68.20**	95.70	**97.00**	88.90	**89.48**
GPT4-Turbo	61.40	**69.80**	95.53	**98.88**	89.50	**95.50**
EN=>GE						
Gemini1.5-Pro	62.80	**65.20**	95.10	**95.90**	89.00	**89.80**
Yi-Large	61.80	**64.55**	94.25	**97.50**	87.40	**90.40**
GPT3.5-Turbo	61.80	**65.30**	92.90	**97.30**	85.35	**87.50**
GPT4-Turbo	61.30	**68.50**	95.90	**98.00**	88.70	**89.85**

Table 3: Generality analysis of automatic metric in different languages. *SP* represents Spanish and *GE* represents German.

3% in coherency in Yi-Large when translating from English to German. Those further demonstrate the effectiveness and broad applicability of HDM.

5.2 Ablation Study

This analysis aims to investigate the effects of the results on Humour Theory and the Humour Decomposition Mechanism. We randomly select 100 samples to conduct the ablation study, as shown in Table 6, where:

- "-HT" denotes removing the part of humour theory. Our approach will only use the analyzes for the intermediary stage.
- "-HDM" denotes removing the Humour Decomposition Mechanism. We directly input the prompt of decomposing humour to conduct the translation.
- "base" denotes both removing the Humour Decomposition Mechanism and humour theory.

From Table 6 we observe that HDM demonstrates significant performance gains across all LLMs and evaluation metrics and plays a critical component of our approach, especially in humour. We attribute these improvements to CoT prompts, which help LLMs refine translated text by enhancing their parsing and reconstruction abilities.

Humour Theory (HT) further delivers some improvements after HDM. For example, Gemini1.5-Pro achieve gains of +3.3%, +1.00%, and +3.10% in humour, fluency, and coherency, respectively. However, we find that the improvements are less pronounced after removing HDM compared to the baseline. In some cases, such as with GPT4, there are even declines. This indicates that HT works more effectively when combined with HDM, leading to better overall performance.

5.3 How does prompt selection affect HDM?

We also validate the robustness of the zero-shot Humour Decomposition Mechanism against the different humour translation prompting.

Figure 4 illustrates the performance of four different prompts in HDM by using GPT4-Turbo. The experimental findings reveal that despite fluctuations in GEMBA-SQM evaluation of reasoning across different prompts, all humour translation prompts consistently enhance performance compared to the traditional CoT approach. This further verifies the effectiveness of HDM.

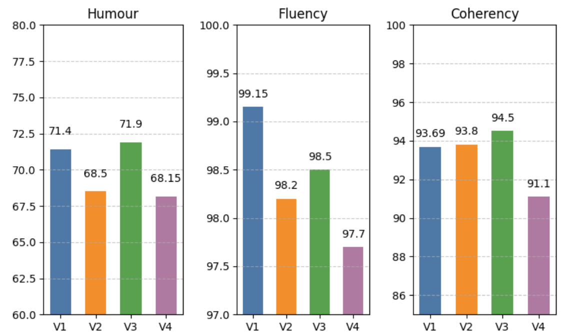


Figure 4: Performance comparisons of four various prompts of HDM in humour, fluency and coherency, marked by V1, V2, V3 and V4. The y-axis is the score on the GEMBA-SQM. We evaluate the performance on the Short Jokes Dataset using the GPT4-Turbo setting.

5.4 Case Study and Error Analysis

In this section, we present some correct examples generated by using HDM as shown in figure 2 and make some analysis for some bad cases. For instance, the generated translation of P_1 describes the background sentence as "the snail say while riding on the turtle's back", while the snail shouting "Wheeeeeee" reflects the snail's feeling that the turtle is fast, which highlights the humorous effect.

Setting	EN=>ZH			EN=>SP		
	SQM-H	SQM-F	SQM-C	SQM-H	SQM-F	SQM-C
GPT4-Turbo						
-	**70.50**	**98.80**	**96.70**	**68.00**	**99.10**	**95.70**
-HDM	54.60	95.65	88.67	57.30	97.00	89.10
-HT	69.15	96.67	91.77	67.10	98.67	94.05
base	51.60	93.30	87.20	55.20	96.67	88.80
Gemini1.5-Pro						
-	**66.50**	**97.30**	**93.90**	**66.20**	**98.70**	**94.61**
-HDM	57.40	94.00	93.50	60.80	98.30	89.40
-HT	63.20	96.30	90.80	65.80	98.40	92.00
base	56.30	93.70	87.70	53.60	97.23	90.83

Table 4: Ablation results on Humour Decomposition Mechanism with various LLMs settings on Short Joke Dataset.

In the traditional translation, the onomatopoeia of "Wheeeeeee" is translated into "Whoosh (back translation)", while in HDM, the snail more intuitively reflects the language humour effect by saying "Wow This is exciting! (back translation)". The jokes generated by using HDM are more informative and coherent than directly translated text, thus allowing people to better understand the humorous connotations of the texts.

In addition, there are still some samples that HDM is hard to address. One situation involves the judgment of the source language based on the pronunciation and shape of characters within the context of puns. For example, the joke is "How do sheep in Mexico say Merry Christmas? Fleece Navidad!". The punchline of this joke relies on the auditory similarity between "Fleece" and "Feliz." By substituting "Feliz" with "Fleece" it creates a humorous image of sheep celebrating Christmas in their own way. In this case, HDM struggles to generate jokes that combine puns with cultural and linguistic elements.

6 Related works

6.1 Humour Theory

Raskin (1979) proposes the incongruity theory, which believes that the key to humour is the incongruity between readers' expectations and the ending of one story (Amir et al., 2016). Toplyn (2014) further proposes the monologue joke generation theory, which defines the structure of a joke as the topic, angle and punchline. There are currently some studies that incorporate humour theory into natural language processing for humour generation (Zhang et al., 2020b; Zhong et al., 2024; Wang et al., 2024; Chen et al., 2023; Chain-of Thought) and humour recognition (Zhao et al., 2019; Alnajjar et al., 2022; Kenneth et al., 2024). According to this theory, we explore how to translate the jokes across different languages.

6.2 Translation for LLMs

Extensive research has been conducted to evaluate the translation capabilities of LLMs. Some people study issues specific to LLMs, including the selection of prompt templates (Jiao et al., 2023; Zhang et al., 2023) and In-Context Learning(Vilar et al., 2022; Zhang et al., 2023). Other researchers investigate translation across diverse scenarios, such as low-resource translation (Jiao et al., 2023; Zhu et al., 2023), document-level (Hendy et al., 2023; Karpinska and Iyyer, 2023; Wang et al., 2023) and Multilingual machine translation (Zhu et al., 2023; Jiao et al., 2023).

6.3 Chain-of-Thought (CoT)

CoT prompting involves either providing instruction or a few chain-of-thought examples (Ji et al., 2024). Recently, a series of studies (Ye and Durrett, 2023; Zhou et al., 2022a; Kojima et al., 2022; Zhang et al., 2022; Fei et al., 2023) have proposed their respective prompting strategies, breaking down the entire task into smaller components and then systematically addressing, strategizing, and carrying out each of these components. With the improvement of model capabilities, some works (Zhou et al., 2022b; Gao et al., 2023; Zelikman et al., 2022) treat the instruction as the "program" for searching, optimization, generating programs and bootstrapping the ability to perform successively more complex reasoning.

7 Conclusion and Future Work

In this paper, we introduce a novel approach named Humour Decomposition Mechanism (HDM) for humour translation. Specifically, HDM consists of *humour decomposition* and *translation module* and *humour composition*, which creates a three-step paradigm of mining intrinsic knowledge of jokes, translating the intrinsic knowledge and then composing the jokes based on the translation. Moreover, we integrate humour theory into HDM to boost performance further. Experimental results in automatic and human evaluation both reveal our method can attain promising performance in humour translation. In the future, we will explore the methods for incorporating automatic and human review in HDM to further improve the quality of humour translation.

8 Limitations

Although our methods have demonstrated significant advantages in experimental evaluations, in human evaluation, the evaluators of our researcher correspond to all authors in this paper. This may result in potential evaluation bias. The evaluator-researcher overlap may affect the objectivity of the results. Therefore, it needs further validation of the fairness of human evaluation in the future.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation. *arXiv preprint arXiv:2212.02437*.

01. AI, :, Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. 2024. Yi: Open foundation models by 01.ai. *Preprint*, arXiv:2403.04652.

Khalid Alnajjar, Mika Hämäläinen, Jörg Tiedemann, Jorma Laaksonen, and Mikko Kurimo. 2022. When to laugh and how hard? a multimodal approach to detecting humor and its intensity. *arXiv preprint arXiv:2211.01889*.

Silvio Amir, Byron C Wallace, Hao Lyu, and Paula Carvalho Mário J Silva. 2016. Modelling context with user embeddings for sarcasm detection in social media. *arXiv preprint arXiv:1607.00976*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Using Incongruity Resolution Chain-of Thought. Content-specific humorous image captioning using incongruity resolution chain-of-thought.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024a. DUAL-REFLECT: Enhancing large language models for reflective translation through dual learning feedback mechanisms. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 693–704, Bangkok, Thailand. Association for Computational Linguistics.

Guanhua Chen, Shuming Ma, Yun Chen, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. 2022. Towards making the most of cross-lingual transfer for zero-shot neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 142–157.

Yuetian Chen, Bowen Shi, and Mei Si. 2023. Prompt to gpt-3: Step-by-step thinking instructions for humor generation. *arXiv preprint arXiv:2306.13195*.

Yuyan Chen, Yichen Yuan, Panjun Liu, Dayiheng Liu, Qinghao Guan, Mengfei Guo, Haiming Peng, Bang Liu, Zhixu Li, and Yanghua Xiao. 2024b. Talk funny! a large-scale humor response dataset with chain-of-humor interpretation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17826–17834.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*.

Yves Gambier. 2016. Translations| rapid and radical changes in translation and translation studies. *International Journal of Communication*, 10:20.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.

José Antonio García-Díaz and Rafael Valencia-García. 2021. UMUTeam at SemEval-2021 task 7: Detecting and rating humor and offense with linguistic features and word embeddings. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1096–1101, Online. Association for Computational Linguistics.

Desta Haileselassie Hagos, Rick Battle, and Danda B Rawat. 2024. Recent advances in generative ai and large language models: Current status, challenges, and perspectives. *IEEE Transactions on Artificial Intelligence*.

Md Kamrul Hasan, Sangwu Lee, Wasifur Rahman, Amir Zadeh, Rada Mihalcea, Louis-Philippe Morency, and Ehsan Hoque. 2021. Humor knowledge enriched transformer for understanding multimodal humor. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 12972–12980.

Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Chris Hopkinson. 2007. Factors in linguistic interference: A case study in translation. *SKASE Journal of translation and interpretation*, 2(1):13–23.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Bin Ji, Huijun Liu, Mingzhe Du, and See-Kiong Ng. 2024. Chain-of-thought improves text generation with citations in large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18345–18353.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is chatgpt a good translator? a preliminary study. *arXiv preprint arXiv:2301.08745*, 1(10).

Marzena Karpinska and Mohit Iyyer. 2023. Large language models effectively leverage document-level context for literary translation, but critical errors persist. *arXiv preprint arXiv:2304.03245*.

Mary Ogbuka Kenneth, Foaad Khosmood, and Abbas Edalat. 2024. A two-model approach for humour style recognition. *arXiv preprint arXiv:2410.12842*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2023. Chain-of-dictionary prompting elicits translation in large language models. *arXiv preprint arXiv:2305.06575*.

Xingcheng Ma and Andrew KF Cheung. 2020. Language interference in english-chinese simultaneous interpreting with and without text. *Babel*, 66(3):434–456.

Abhinav Moudgil. 2016. Short jokes. Kaggle, Data set. Available at: https://www.kaggle.com/datasets/abhinavmoudgil95/short-jokes.

Eugene Albert Nida. 1964. *Toward a science of translating: with special reference to principles and procedures involved in Bible translating*. Brill Archive.

Anthony Pym. 2023. *Exploring translation theories*. Routledge.

Victor Raskin. 1979. Semantic mechanisms of humor. In *Annual Meeting of the Berkeley Linguistics Society*, pages 325–335.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *arXiv preprint arXiv:2009.09025*.

Jiri Roznovjak. 2016. Question-answer jokes. Kaggle, Data set. Available at: https://www.kaggle.com/datasets/jiriroz/qa-jokes.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

J. Toplyn. 2014. *Comedy Writing for Late-night Tv: How to Write Monologue Jokes, Desk Pieces, Sketches, Parodies, Audience Pieces, Remotes, and Other Short-form Comedy*. Twenty Lane Media, LLC.

Jeroen Vandaele. 2016. *Translating humour*. Routledge.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting palm for translation: Assessing strategies and performance. *arXiv preprint arXiv:2211.09102*.

Dexin Wang, Kai Fan, Boxing Chen, and Deyi Xiong. 2022a. Efficient cluster-based k-nearest-neighbor machine translation. *arXiv preprint arXiv:2204.06175*.

Han Wang, Yilin Zhao, Dian Li, Xiaohan Wang, Gang Liu, Xuguang Lan, and Hui Wang. 2024. Innovative thinking, infinite humor: Humor research of large language models through structured thought leaps. *arXiv preprint arXiv:2410.10370*.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. *arXiv preprint arXiv:2304.02210*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022b. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Chenri Xia, Mansour Amini, and Kam-Fong Lee. 2023. Humor translation: A case study on the loss of humorous loads in spongebob squarepants. *Cadernos de Tradução*, 43:e89705.

Xi Ye and Greg Durrett. 2023. Explanation selection using unlabeled data for chain-of-thought prompting. *arXiv preprint arXiv:2302.04813*.

Patrick Zabalbeascoa. 2005. Humor and translation—an interdiscipline.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Hang Zhang, Dayiheng Liu, Jiancheng Lv, and Cheng Luo. 2020a. Let's be humorous: Knowledge enhanced humor generation. In *Annual Meeting of the Association for Computational Linguistics*.

Hang Zhang, Dayiheng Liu, Jiancheng Lv, and Cheng Luo. 2020b. Let's be humorous: Knowledge enhanced humor generation. *arXiv preprint arXiv:2004.13317*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Zhenjie Zhao, Andrew Cattle, Evangelos Papalexakis, and Xiaojuan Ma. 2019. Embedding lexical features via tensor decomposition for small sample humor recognition. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2024. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13246–13257.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022a. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022b. Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

A Appendix

A.1 Dataset Analysis

The structure of the humour dataset is as follows:

JokeDataset = (ID, Content, Topic, Angle, Punchline, DataSource, Link, Original Version). Figure 5 illustrates an example from the translation Chinese dataset.

In our dataset:

- ID: The ID of the target language joke.
- Content: The content of the target language joke.
- Topic, Angle and Punchline: The humour theory elements of the target language joke.
- DataSource: The name of the source language joke dataset.
- Link: The link of the source language joke dataset.
- Original Version: The source language version of the joke.

Additionally, we analyze the most frequently appearing vocabulary in each dataset to determine whether the translated text deviates from the meaning of the source language text. As figure 6 shows, some high-frequency words both appear in the source and target language dataset, including terms such as “time”, “day” and “good”. Also, some new words appear in the target language dataset such as “friend” and “new”. We attribute this to the fact that LLMs tend to expand the translated text while preserving the essence of the source language, consequently resulting in the emergence of new words that overshadow the original high-frequency words.

A.2 The Prompt Details of GEMBA.

We use the *Estimation Metric Based Assessment* (GEMBA), a type of LLM evaluation, to formalize the definitions of evaluation prompts. Based on these definitions, we report several of our prompt strategies for evaluation metrics, as shown in Table 5. For GEMBA-SQM, a continuous scale from 0 to 100 is used to define four stages. For instance, GEMBA-SQM-F categorizes these stages as “No Fluency“, “Some Fluency“, “Most Fluency“ and “Perfect Fluency“. GEMBA-STARS is a classification task based on a one-to-five star ranking. For example, GEMBA-STAR-F is a five-star evaluation metric of fluency, with one star representing “No Fluency,“ two stars indicating “Less fluency“, three stars signifying “Some fluency“, four stars denoting “Most fluency“, and five stars indicating “Perfect fluency“.

A.3 Ablation Study

Table 6 shows the prompt details of removing HDM, removing HT, and baseline in the ablation study.

- -HDM denotes removing the Humour Decomposition Mechanism.
- -HT denotes removing the part of humour theory.
- “Base“ denotes both removing the Humour Decomposition Mechanism and humour theory

A.4 Prompt Selection in HDM

To further verify the robustness and effectiveness of HDM, we perform an analysis of the final outcomes across a range of HDM with varying expressions. Specifically, we utilize GPT4 to rewrite the prompts of *Humour Decomposition* module in HDM. Our instructions is like as follows:

> Please rewrite the following prompt into a new version: “You are a humour explanation assistant. A joke can be thought of as being composed based on three components. Under a particular theory of joke information, those components are:
> 1. The topic, which is the news item that the joke is based on.
> 2. The angle, which is the particular direction that the joke takes.
> 3. The punchline, which is the surprise at the end of the joke.
> Please analyze the following joke and provide the best explanation of what the topic is, what the angle is, and what the punchline is:“

As shown in table 7, we report four different prompt selections in HDM, which correspond to the V1, V2, V3 and V4 in the paper, respectively.

<Humour>
<ID>H0001**</ID>**
<Content>你知道为什么《侏罗纪世界》比《侏罗纪公园》差吗？因为据说每一卷《侏罗纪世界》的电影胶卷都比《侏罗纪公园》轻21克——看来是少了那部分灵魂的重量啊！**</Content>**
<Topic>比较电影《侏罗纪公园》和《侏罗纪世界》的质量。**</Topic>**
<Angle>这个笑话从引用灵魂重21克的观念入手，暗示《侏罗纪世界》与《侏罗纪公园》相比缺失了一些重要或有灵魂的东西。**</Angle>**
<punchline>《侏罗纪世界》的电影胶卷比《侏罗纪公园》轻21克"，这句话幽默地暗示《侏罗纪世界》缺少了《侏罗纪公园》所具有的深度或本质（"灵魂"），并用21克的重量来量化。**</punchline>**
<DataSource>Short Joke Dataset**</DataSource>**
<Link>https://www.kaggle.com/datasets/abhinavmoudgil95/shortjokes**</Link>**
<Original Version>The soul weighs 21 grams. We know this because the Jurassic World film reels are 21 grams lighter than Jurassic Park.**</Original Version>**
</Humour>

Figure 5: An example of the Translation Joke Dataset

(a) Source Language Dataset

(b) Target Language Dataset

Figure 6: Word cloud in the Short Joke dataset (a) and target translation dataset (b). For an intuitive comparison, the word cloud (b) has been translated from Chinese.

<Below is the prompt text of GEMBA-SQM-H>
Instruction: Score the following {target_lang} sentence on a continuous scale from 0 to 100 that starts on "No humour", goes through "Some humour", then "Humour", up to "Most humour".
Sentence: "sentence"
Score (0-100):

<Below is the prompt text of GEMBA-SQM-F>
Instruction: Score the following {target_lang} sentence on a continuous scale from 0 to 100 that starts on "No fluency", goes through "Some fluency", then "Most fluency", up to "Perfect fluency".
Sentence: "sentence"
Score (0-100):

<Below is the prompt text of GEMBA-SQM-C>
Instruction: Score the following {target_lang} sentence on a continuous scale from 0 to 100 that starts on "No coherency", goes through "Some coherency", then "Most coherency", up to "Perfect coherency".
Sentence: "sentence"
Score (0-100):

<Below is the prompt text of GEMBA-STAR-H>
Instruction: Score the following {target_lang} sentence with one to five stars. Where one star means "No humour", two stars mean "Less humour", three stars mean "Some humour", four stars mean "Most humour", and five stars mean "Perfect humour".
Sentence: "sentence"
Stars:

<Below is the prompt text of GEMBA-STAR-F>
Instruction: Score the following {target_lang} sentence with one to five stars. Where one star means "No fluency", two stars mean "Less fluency", three stars mean "Some fluency", four stars mean "Most fluency", and five stars mean "Perfect fluency".
Sentence: "sentence"
Stars:

<Below is the prompt text of GEMBA-STAR-C>
Instruction: Score the following {target_lang} sentence with one to five stars. Where one star means "No coherency", two stars mean "Less coherency", three stars mean "Some coherency", four stars mean "Most coherency", and five stars mean "Perfect coherency".
Sentence: "sentence"
Stars:

Table 5: The prompt details of GEMBA in our approach

<Below is the prompt text of removing HDM>

Instruction: You are a humour explanation assistant. A joke can be thought of as being composed based on three components. Under a particular theory of joke information, those components are:

1. The topic, which is the news item that the joke is based on.
2. The angle which is the particular direction that the joke takes.
3. The punch line which is the surprise at the end of the joke.

Please translate the following joke in Spanish based on this theory: [source language joke]

<Below is the prompt text of removing HT>

Instruction: Please analyze the following joke: [source language joke]

Instruction: Please translate the analysis from English to Spanish: [Analysis]

Instruction: Please generate a Spanish joke based on the analysis: [Translated analysis]

<Below is the prompt text of the baseline>

Instruction: Please translate the joke from English to Spanish: [source language joke]

Table 6: The prompt details in Ablation Study

<Prompt Selection V1>

Instruction: As a humour explanation assistant, jokes can be analyzed based on three key components according to a specific theory of humour:
1. The topic, which represents the news item or subject the joke revolves around.
2. The angle, which indicates the specific perspective or approach the joke takes.
3. The punch line, which delivers the unexpected twist or surprise at the end of the joke.

Please analyze the following joke and provide your best estimate of its topic, angle, and punch line: [source language joke]

Instruction: Please translate the analysis from English to Spanish: [Analysis]

Instruction: Please generate a Spanish joke based on the analysis: [Translated analysis]

<Prompt Selection V2>

Instruction: As a humour analysis assistant, jokes can be broken down into three essential elements according to a particular theory of humour:
1. The Topic: This refers to the main subject or context around which the joke is centered.
2. The Angle: This represents the unique perspective or approach that the joke takes toward the topic.
3. The Punch Line: This is the unexpected twist or conclusion that provides humour, often through a surprising or witty remark.

Please explain the following joke by identifying its topic, angle, and punch line: [source language joke]

Instruction: Please translate the text from English to Spanish: [Analysis]

Instruction: Please generate a Spanish joke based on the analysis: [Translated analysis]

<Prompt Selection V3>

Instruction: According to a specific theory of humour, jokes can be analyzed into the topic, which is the news item that the joke is based on, the angle, which is the particular direction that the joke takes, and the punchline, which is the surprise at the end of the joke.

Please analyze the following joke and provide your best estimate of its topic, angle, and punchline: [source language joke]

Instruction: Please translate the text from English to Spanish: [Analysis]

Instruction: Please generate a Spanish joke based on the analysis: [Translated analysis]

<Prompt Selection V4>

Instruction: Jokes can be decomposed into the topic, angle and punchline According to a specific theory of humour. Specifically, the topic is the news item that the joke is based on, the angle is the particular direction that the joke takes, and the punchline is the surprise at the end of the joke.

Please decompose the following joke and provide the decomposition of its topic, angle, and punchline: [source language joke]

Instruction: Please translate the text from English to Spanish: [Analysis]

Instruction: Please generate a Spanish joke based on the analysis: [Translated analysis]

Table 7: The prompt selection in HDM