SlangDIT: Benchmarking LLMs in Interpretative Slang Translation

Anonymous ACL submission

Abstract

The challenge of slang translation lies in capturing context-dependent semantic extensions, as slang terms often convey meanings beyond their literal interpretation. While slang detection, explanation, and translation have been studied as isolated tasks in the era of large language models (LLMs), their intrinsic interdependence remains underexplored. The main reason is lacking of a benchmark where the two tasks can be a prerequisite for the third one, which can facilitate idiomatic translation. In this paper, we introduce the interpretative slang translation task (named SlangDIT) consisting of three sub-tasks: slang detection, cross-lingual slang explanation, and slang translation within the current context, aiming to generate more accurate translation with the help of slang detection and slang explanation. To this end, we construct a SlangDIT dataset, containing over 25k English-Chinese sentence pairs. Each source sentence mentions at least one slang term and is labeled with corresponding cross-lingual slang explanation. Based on the benchmark, we propose a deep thinking model, named SlangOWL. It firstly identifies whether the sentence contains a slang, and then judges whether the slang is polysemous and analyze its possible meaning. Further, the SlangOWL provides the best explanation of the slang term targeting on the current context. Finally, according to the whole thought, the SlangOWL offers a suitable translation. Our experiments on LLMs (*e.g.*, Qwen2.5 and LLama-3.1), show that our deep thinking approach indeed enhances the performance of LLMs where the proposed SLangOWL significantly surpasses the vanilla models and supervised fine-tuned models without thinking. [1]

1 Introduction

The slang includes words, phrases, idioms, and expressions that are not typically found in formal language or standard dictionaries, conveying subtle shades of meaning, tone, and attitude, which is often used in social interactions, particularly among specific groups or communities (Mashhady and Pourgalavi, 2013). The natural characteristic leads to some challenges of slang translation since it is hard to model context-dependent semantic extensions. Although plenty of studies on them have been carried out based on slang detection (Ishiwatari et al., 2019; Pei et al., 2019), or cross-lingual slang explanation (Gluck et al., 2025), or slang translation (Sun et al., 2022) in the era of large language models (LLMs) (Jhirad et al., 2023; Pei et al., 2019; Sun et al., 2022, 2024), to our knowledge, little research work has been devoted to slang translation with the help of slang detection and cross-lingual slang explanation. One important reason is the lack of such slang translation datasets.

Source Sentence: As I recall, you've always carried a torch for Oldsmobiles.
Slang Term: carried a torch
Cross-lingual Slang Explanation: “carried a torch” 是一个俚语，意思是对某人或某物有持续的爱慕或热情，通常用于描述对某人的暗恋或长期的喜爱。在这个句子中，它表示你一直对Oldsmobiles这款车有特别的情感。(‘carried a torch’ is a slang term that means having a sustained admiration or passion for someone or something, usually used to describe a secret love or long-term affection for someone. In this sentence, it means that you have always had a special emotional attachment to the Oldsmobile car.)
Chinese Translation: 我记得，你一直特别喜欢奥兹摩比。

Source Sentence: The athlete proudly carried a torch during the opening ceremony of the Olympics.
Slang Term: No slang
Explanation: “carried a torch” 是手持火炬 - 字面意思，指手里拿着一个火炬。 (‘carried a torch’ means holding a torch - literally means holding a torch in one's hand..)
Chinese Translation: 这位运动员在奥运会开幕式上自豪地手持火炬。

Source Sentence: She carried a torch for environmental conservation, dedicating her life to protecting nature.
Slang Term: No slang
Explanation: “carried a torch”坚持某种信念 - 引申含义，指坚持某种信仰、理想或目标，通常是为了纪念或延续某种精神。 (‘carried a torch’ means adhering to a certain belief - its extended meaning refers to adhering to a certain faith, ideal, or goal, usually to commemorate or continue a certain spirit...)
Chinese Translation: 她坚持环境保护的信念，毕生致力于保护自然。

Figure 1: Some examples of the SlangDIT benchmark.

Meanwhile, the previous work generally neglects the polysemy present in slang terms. In different contexts, the slang term conveys differ-

[1] Source codes and data will be freely available at github upon acceptance.

ent ideas. For example, in the Figure 1, the term '*carried a torch*' is a slang term in the top box while it is not in the middle ('superficial semantic meaning') and bottom box ('extended meaning'). In fact, for humans, to identify and understand slang, one firstly needs to be familiar with the cultural, social, and historical context in which it is used. This involves recognizing the nuances of language and also requires an awareness of the ever-changing nature of language, as slang terms and expressions may quickly evolve or become outdated (Lėgaudaitė, 2010; Keidar et al., 2022). For instance, if we take the slang term (*i.e.* '*carried a torch*') as general words in the top box of Figure 1, we could not capture the real sense the speaker said (*i.e.*, 'sustained admiration or passion for someone or something'). Secondly, to translate it, one requires not only conveying the literal meaning of the words but also capturing the tone, connotation, and implied meaning that is often embedded in slang expressions (Mattiello, 2009). Even for professional human translators, they sometimes fail to convey the intended meaning in practice. As shown in the example of Figure 1, directly translate the slang term can not show the subtle meaning the sentence reflects. And with the help of the slang detection and cross-lingual slang explanation, the translation will express the original intention and become satisfactory. All of the above call for a such data resource that can encourage further research in slang understanding and translation.

In this work, we propose a new task named **I**nterpretative **Slang T**ranslation (SlangDIT), with the goal to produce more accurate translations by taking the detected slang and the corresponding cross-lingual slang explanation. To this end, we firstly construct a SlangDIT dataset. Specifically, based on the large-scale movie subtitles[2], about 28M English-Chinese sentence pairs (Liang et al., 2022), (1) we use advanced LLMs (*e.g.*, Qwen2.5-72b (Qwen Team, 2024)) to judge whether the English sentence contains a slang. (2) for the sentence with a slang (~776k), we further utilize Qwen2.5-72b and Llama3.3-70b to extract the slang terms. To ensure the data quality, we only maintain the sentence where both of Qwen2.5-72b and Llama3.3-70b predict the same slang term and GPT-4o agrees that it is a slang. (3) We further utilize Qwen2.5-72b produce the Chinese explanation. (4) To explore the context impact on the slang term, we utilize GPT-4o to annotate whether each slang term is polysemous. Consequently, we obtain over 25k English-Chinese sentences with 13,580 genenal slang terms and 7,818 polysemous slang terms.

Based on the constructed SlangDIT dataset, we propose a deep thinking model (SlangOWL) that simulates cognitive process of humans. Specifically, the model firstly identifies whether the sentence contains a slang, and then judges whether the slang is polysemous and analyze its possible meaning. Further, it provides the best cross-lingual explanation of the slang term targeting on the current context. Finally, according to the thought, the model offers a suitable translation. To achieve this goal, we need the long thought samples to train our models. In view of promising reasoning ability in existing o1-like LLMs, we decide to provide the four key elements (slang term, polysemy, cross-lingual explanation and translation) to DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI et al., 2025), and collect the thought process.

Experiments on three LLM-based systems (Qwen Team, 2024; Dubey et al., 2024), *i.e.*, Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct and Qwen2.5-14B-Instruct, show the effectiveness of deep thinking on translation. It significantly outperforms the vanilla models and simply supervised fine-tuned models in terms of BLEU (Papineni et al., 2002), CometKivi (Rei et al., 2022), Comet (Rei et al., 2020) and evaluators via GPT-4o.

In summary, our main contributions are:

- We propose a new task: interpretative slang translation named SlangDIT, consisting of three subtasks, to advance slang understanding and translation research.
- We are the first that contributes the SlangDIT dataset, which contains 25k <English sentence, Chinese sentence, slang terms, cross-lingual slang explanation> quadruplets. Particularly, it offers 7,818 polysemous slang terms.
- We propose a SlangOWL model that achieves the best performance with the help of deep thinking during translation. We also show that the slang detection and cross-lingual explanation play a key role in translating the sentence with a slang term.

2 SlangDIT Task

In this section, we firstly clarify the symbol definition, and then define the proposed *Interpretative*

[2] https://www.jubenz.com/

	Type	#Sentences			# Slang Type		XLSE	#AvgEn	#AvgZh	#AvgExp
		No	*w/* NPST	*w/* PST	NPST	PST				
SlangDIT	Train	39,980	13,580	26,400	13,580	6,653	39,980	7.89	12.47	74.06
	Valid	1,815	1,815	0	1,815	0	1,815	7.93	12.71	76.65
	Test	1,863	1,863	0	1,863	0	1,863	7.88	12.77	77.12
	Hard Test	0	0	1,165	0	1,165	1,165	7.86	12.57	75.54

Table 1: Detailed Statistics of our SlangDIT dataset. 'NPST' and 'PST' means non-polysemous slang terms and polysemous slang terms; 'XLSE' means cross-lingual slang explanation; #: number of the corresponding item, *i.e.*, AvgEn: Average length of each utterance in English (word level); AvgZh/AvgExp: Average length of each sentence in Chinese (character level).

Slang Translation task.

Given an input sentence in the source language $X = X_1, X_2, X_3, ..., S_1, S_2, ..., S_p, ..., X_m$ where the X_i is the token and the S_i is the token belongs to a slang, the goal of the SlangDIT task is to identify the slang term $S = Y_1^s, Y_2^s, Y_3^s, ..., Y_p^s$, and then generate its explanation in a target language $E = Y_1^e, Y_2^e, Y_3^e, , ..., Y_k^e$, and finally output its translation in a target language $Y = Y_1^t, Y_2^t, Y_3^t, ..., Y_m^t$.

Formally, the probability distribution of the target output S, E, Y are defined as follows:

$$P(S, E, Y|X) = \prod_{r=1}^{R} p(Y_r^*|Y_{<r}^*, X), \qquad (1)$$

where $* \in \{s, e, t\}$ and $Y_{<r} = \{Y_1, Y_2, Y_3, ..., Y_{r-1}\}$ and $R = p + k + m$.

3 SlangDIT Dataset

In this section, we mainly introduce our SlangDIT dataset in five aspects: *Data Source* § 3.1, *Annotation Procedure* § 3.2, *Annotation Quality Assessment* § 3.3, *Dataset Statistics* § 3.4, and the introduction of *Related Datasets* § 3.5.

3.1 Data Source

Because movie subtitles contain utterances that better reflect natural conversations which usually involve slang terms (Sun et al., 2024; Chen et al., 2024), we thus choose the movie subtitle as our data source, *e.g.*, the large-scale movie subtitles (Liang et al., 2022) and OpenSubtitles (Lison and Tiedemann, 2016)[3]. Due to the dataset of Liang et al. (2022) offers the corresponding Chinese translation, we take this dataset as our choice. However, the lack of associated slang annotation and cross-lingual slang explanation makes it impossible for directly conducting research on interpretative slang translation. Therefore, we further annotate slang term and the corresponding slang term explanation.

3.2 Annotation Procedure

Since the full data are large (~28M), the annotation procedure is automatic to build the SlangDIT dataset via advanced LLMs (Qwen2.5-72b, Llama3.3-70b and GPT-4o), which includes four steps: slang judging, slang extraction, explanation generation and polysemy annotation [4].

Slang Judging. Before judging, we firstly filter the offensive and dirty sentences. Then, to improve the annotation efficiency, we utilize Qwen2.5-72b to judge whether each utterance contains any slang terms. After that, we filter the sentences that contains repetitive slang terms and we obtain ~776k out of 28M (2.8%) sentences with the slang term.

Slang Extraction. To ensure the data quality and avoid model bias, we separately utilize Qwen2.5-72b and Llama3.3-70b to extract the slang term. If both Qwen2.5-72b and Llama3.3-70b predict the same slang term and GPT-4o also approves that it is a slang term, we maintain such sentences (25k).

Explanation Generation. With the advance Chinese ability of Qwen2.5-72b, we use it to generate Chinese explanation for each slang term.

Polysemy Annotation. To fully investigate the impact of context on the slang term, we use GPT-4o to judge whether each slang term is polysemous. Consequently, we obtain 7,818 clear polysemous slang terms, and 17,258 non polysemous slang terms.

Besides, for constructing the hard testset, we randomly sample 15% instances (*i.e.*, 1,165) from the polysemous slang terms. For the remaining 85% (*e.g.*, 6,653) polysemous slang terms, we use GPT-4o to generate possible meaning for each slang term. Finally, for each meaning, we further use

[3] http://www.opensubtitles.org/

[4] The prompt used in this process are presented in Figure 7 ~ 11 of Appendix.

Datasets	SD	XLSE	Translation	Polysemy	#Slang Terms		
					Train	Valid	Test
Urban Dictionary (Ni and Wang, 2017)	✓	✗	✗	✗	371,028	-	50,000
OSD (Sun et al., 2022)	✓	✗	✗	✗	1,635	-	299
GDoS (Adams, 2012)	✓	✗	✗	✗	-	-	-
CLIX (Gluck et al., 2025)	✗	✓(English→Spanish/German)	✗	✗	278	150	200
OpenSubtitles-Slang (Sun et al., 2024)	✓	✗(English→English)	✓(English→German/French)	✗	836	-	-
SlangDIT (Ours)	✓	✓(English→Chinese)	✓(English→Chinese)	✓	20,233	1,815	3,028

Table 2: Comparison of previous slang detection dataset: Urban Dictionary, OSD, GDoS, and OpenSubtitles-Slang; (2) cross-lingual slang explanation datasets: CLIX and OpenSubtitles-Slang, and (3) our SlangDIT. 'SD' means slang detection and 'polysemy' means polysemous labeling.

GPT-4o to produce corresponding translation pairs that contains the same slang term but convey different sense. After this process, we obtain 26,400 sentence pairs for training.

3.3 Annotation Quality Assessment

To evaluate the quality of slang detection and cross-lingual slang explanation, we employ three annotators to judge whether the slang term is real and whether its explanation is correct targeting on the context over 200 randomly sampled data. Then, we measure the inter-annotator agreement. The inter-annotator agreements calculated by Fleiss' kappa are 0.685 and 0.915 for slang detection and explanation, which indicates "Substantial Agreement" and "Almost Perfect Agreement" between three annotators, respectively. The level is consistent with previous work (Liang et al., 2022) which can be considered as reliable.

3.4 Dataset Statistics

As shown in Table 1, the SlangDIT contains totally 44,823 English-Chinese utterance pairs with slang term, where each slang term has been annotated with explanation. According to slang terms, we split the dataset into 39,980 for training, 1,815 for validation, and 1,863 for testing[5]. To keep roughly the same distribution of the utterance pair and avoid the model bias, we sample 39,980/1,815/1,863 utterance pairs from the original subtitles where both Qwen2.5-72b and Llama3.3-70b judge no slang terms for training/validation/testing, respectively.

Based on the statistics in Table 1, the average numbers of tokens per utterance are about 7.8, 12.7, and 75.8 for English utterances (word level), Chinese translations (character level), and explanation (character level), respectively.

[5]Note that there is no overlap of the slang term among training, validation and testing set.

3.5 Comparison with Related Datasets

The related datasets mainly involve three research fields: slang detection, cross-lingual slang explanation, and slang translation.

In **slang detection**, there exist some dictionary-based dataset. For example, Ni and Wang (2017) construct a Urban Dictionary, which is large-scale but the quality is poor. Sun et al. (2022) manually annotate a small subset of 102 sentences from the Online Slang Dictionary (OSD). The Green's Dictionary of Slang (GDoS) (Adams, 2012) cannot be publically distributed due to copyright restrictions. Besides, OpenSubtitles-Slang (Sun et al., 2024) only provides the slang definition without context. Most of these dataset only offer the slang term and its definition, which cannot be applied to SlangDIT due to the lack of polysemous labeling, diverse context and translation.

In **cross-lingual slang explanation**, Noraset et al. (2017) propose the definition generation task. As time goes on, Zhang et al. (2023) introduce cross-lingual definitions of (general) words in English, Chinese, and French using prompt learning. Recently, Gluck et al. (2025) propose the task of cross-lingual explanations for idiomatic expressions. Sun et al. (2022) formulate the slang interpretation as the translation task, *i.e.*, the slang term and its interpretation (both in English) are fed into a translation model.

In **slang translation**, closely related to our work is Sun et al. (2022) that show the slang interpretation in the same language can result in improved translation of slang in target language. In this work, they conduct experiment in English-German/French directions with a small-scale machine-translated dataset, which is not publicly available and not target on the study of polysemous slang terms.

The resources mentioned above are extensively

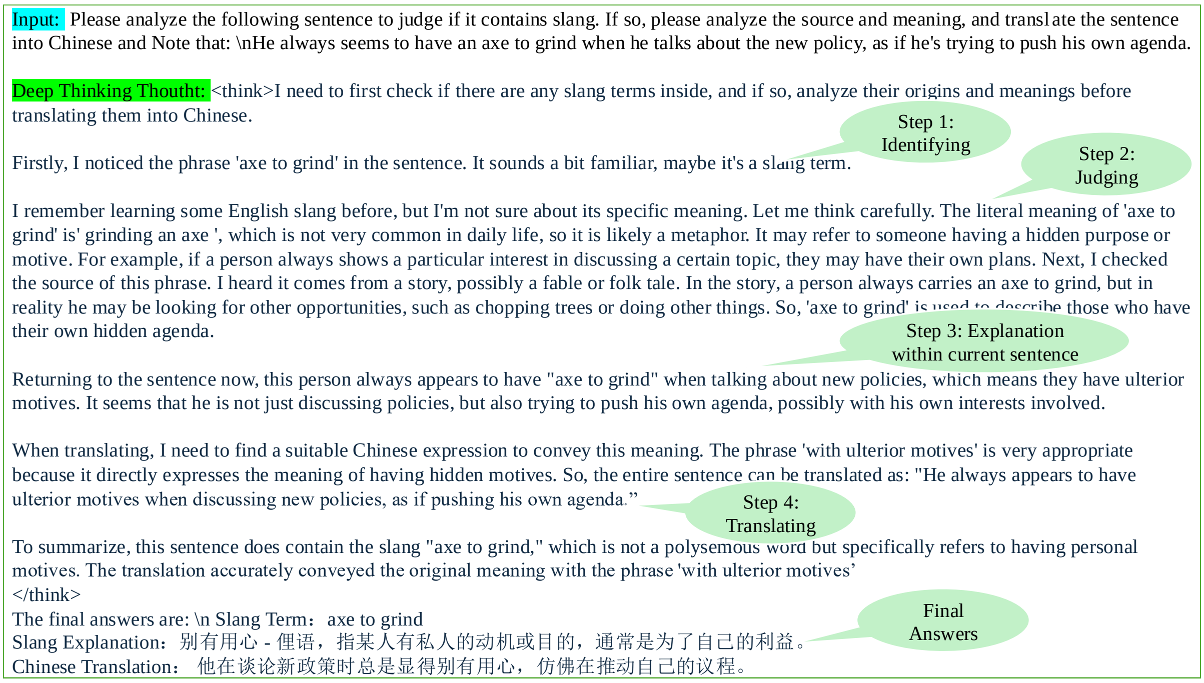

Figure 2: The generated deep thinking thought (training example) by DeepSeek-R1-Distill-Qwen-32B.

used in corresponding fields of research and they even cover some sub-tasks in SlangDIT. However, our SlangDIT is different from them in terms of complexity and diversity.

It is obvious that conducting three sub-tasks is more challenging due to the more complex scene. Furthermore, most of the above single dataset cannot be available and fails to conduct SlangDIT task. What's important, the slang term of above dataset is non-polysemous while ours is polysemous in different context which could be more difficult to interpret. Table 2 provides information on the number of available modality, state of publicly accessible, and their constituent slang terms for all the datasets. Besides, compared with two existing slang dataset, SlangDIT's quantity of English is ten-times of the annotated utterances in most of dataset (except for Urban Dictionary, whose quality is limited). More importantly, the utterance of our SlangDIT comes from movie subtitles, which is natural and diverse than existing data. Particularly, SlangDIT provides an explanation in a cross-lingual language.

4 SlangOWL Model

Backbone. We mainly utilize three LLMs as the backbones of our SlangOWL model: (1) Llama-3.1-8B-Instruct (Dubey et al., 2024); (2) Qwen2.5-7B-Instruct and (3) Qwen2.5-14B-Instruct (Yang et al., 2024).

Thought Data. To simulate cognitive process of humans in translation, we need the deep thinking thought samples to train our SlangOWL model. Given the promising reasoning ability of in existing o1 like models, we decide to use the advance DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI et al., 2025) model. Specifically, we provide four key elements: slang term, polysemy, cross-lingual explanation and translation to the DeepSeek-R1-Distill-Qwen-32B model and prompt it to generate the deep thinking thought for each training instance. The reasoning process roughly includes four steps: 1) identifies whether the sentence contains a slang, and then 2) judges whether the slang is polysemous and analyze its possible meaning. Further, 3) provides the best cross-lingual explanation of the slang term targeting on the current context. Finally, 4) according to the thought, the model offers a suitable translation. We list the generated thought example in Figure 2 and list the prompt in Figure 6 of Appendix.

5 Experiments

5.1 Experimental Setups

Comparison Models. We include three types of baselines: 1) Vanilla instructed models; 2) Vanilla reasoning models; and 3) supervised fine-tuned models without chain-of-thought (SFT w/o cot). Please refer to Appendix B for details. For more training details and inference details, please refer

Types	Models	Slang Detection			XLSE			Translation				
		P	R	F1	R-1	R-2	R-L	BLEU	Comet	GRB	CometK	GRF
Vanilla LLMs	Llama-3.1-8B-Instruct	9.44	12.96	10.61	17.43	2.78	11.79	12.69	65.17	54.24	59.96	70.02
	Qwen2.5-7B-Instruct	74.72	52.04	56.94	25.00	7.21	17.82	16.98	72.41	66.54	70.52	72.66
	Qwen2.5-14B-Instruct	66.77	75.60	70.46	35.48	10.31	25.63	17.66	73.82	68.57	72.36	75.58
	DRT-o1-8B	0.0	0.0	0.0	5.00	3.49	3.15	12.43	39.21	54.10	39.33	58.24
	DRT-o1-7B	4.72	8.94	6.12	15.88	1.13	13.01	14.94	67.62	50.31	66.47	69.54
	DRT-o1-14B	63.95	44.31	50.71	24.75	5.82	17.04	18.45	70.25	65.24	68.10	73.89
	DS-R1-D-Llama-8B	61.57	66.12	63.28	27.33	5.13	19.04	15.76	71.52	58.01	70.82	64.59
	DS-R1-D-Qwen-7B	33.67	43.06	37.79	22.21	4.13	15.55	12.89	64.61	49.17	61.65	42.05
	DS-R1-D-Qwen-14B	68.41	71.41	69.35	30.85	7.16	21.82	18.46	73.18	64.77	**72.56**	73.37
	DS-R1-D-Qwen-32B	60.90	63.33	61.19	33.04	8.28	23.21	19.12	73.69	63.40	72.47	73.39
	QwQ-32B-preview	38.92	55.45	45.42	27.07	6.34	19.63	14.07	72.87	67.98	70.72	76.84
SFT w/o cot	SFT-8B	85.84	61.22	67.03	28.88	17.85	25.48	22.91	73.17	66.39	72.50	71.35
	SFT-7B	84.19	70.67	74.74	28.93	17.87	25.53	23.65	73.70	66.98	71.12	72.71
	SFT-14B	**86.74**	74.15	78.21	29.98	18.78	26.26	24.51	73.62	69.91	71.29	76.00
Deep Thinking	SlangOWL-8B	84.35	87.30††	85.33††	54.77††	34.16††	51.28††	23.47†	73.89†	68.40††	70.72	75.08††
	SlangOWL-7B	82.59	86.49††	84.02††	58.50††	33.16††	52.51††	24.10†	74.04	68.38††	71.60†	74.64††
	SlangOWL-14B	85.17	**88.78**††	**86.47**††	**59.85**††	**34.31**††	**53.72**††	**24.94**	**74.20**†	**71.52**††	71.38	**77.52**††

Table 3: Experimental results (%) on general test set. 'XLSE' denotes cross-lingual slang explanation. 'DS-R1-D' denotes 'DeepSeek-R1-Distill'. "$\dagger$" and "$\dagger\dagger$" denote that statistically significant better than the best result of the counterpart (*e.g.*, SlangOWL-14B vs. SFT-14B) with t-test $p < 0.05$ and $p < 0.01$ hereinafter, respectively. The best and second best results are **bold** and underlined, respectively.

to Appendix A.

Metrics. For *slang detection*, we adopt P, R, and F1 as the metric following previous work (Sun et al., 2024). For *cross-lingual slang explanation*, we use ROUGE-1, ROUGE-2, ROUGE-L, where ROUGE-L evaluate n-grams overlap between the generated explanation and corresponding references (Gluck et al., 2025). We refer ROUGE-* as R-* in the Table for simplicity. For *translation*, we report reference-based and reference-free scores to evaluate the model translations. In reference-based metrics, we use BLEU (Papineni et al., 2002), and Comet (Rei et al., 2020). In reference-free metrics, we use CometKiwi (refered as CometK (Rei et al., 2022)). Here, BLEU is to evaluate n-grams overlap between the generated explanation and corresponding references, and CometScore evaluate the semantic similarity of translations against references. The CometK uses a language model to judge whether a translation conveys the semantics of the source sentence. Besides, following recent studies (Kocmi and Federmann, 2023; Wang et al., 2023, 2024) that show the strong ability of LLMs in NLP evaluation, we also use GPT-4o as evaluators in reference-based and reference-free manner, which we refer to as *GRB* and *GRF*, respectively. For calculation details and evaluation prompts, please refer to Appendix C.

5.2 Main Results

Table 3 shows the main results on the general testset where each instance contains either a non-polysemous slang term or no slang term. Table 4 presents additional results on the hard testset where each instance contains a polysemous slang term.

5.2.1 Results on the General Testset

Results on Slang Detection. Under the *vanilla* setting in Table 3, the reasoning models significantly surpasses the instructed models with the same model scale in terms of F1 score. The bigger models also show better performance than smaller ones, proving that larger models owns better memory capacity. There is no doubt that the *SFT w/o cot* consistently outperforms their counterpart, *e.g.*, the SFT-14B beats Qwen2.5-14B-Instruct, DeepSeek-R1-Distill-Qwen-14B and DRT-o1-14B, even showing better performance than QwQ-32B-preview and DeepSeek-R1-Distill-Qwen-32B. However, the best F1 score (78.21% with SFT-14B) of these models is still substantially lower than the series model of SlangOWL. It shows that the slang term is not just a simple memory task and needs deep thinking to judge whether it is a real slang term. Although o1-like models can conduct reasoning, they fail to identify the correct slang terms. Our proposed SlangOWL models have deep reasoning ability, which analyze the novel and

Models	Slang Detection			XLSE			Translation				
	P	R	F1	R-1	R-2	R-L	BLEU	Comet	GRB	CometK	GRF
Qwen2.5-14B-Instruct	49.61	55.36	51.39	39.01	10.66	27.76	12.44	67.14	67.19	63.67	75.72
DRT-o1-14B	39.25	40.86	39.65	36.98	8.71	25.61	13.76	57.75	68.32	54.02	79.16
DeepSeek-R1-Distill-Qwen-14B	56.88	64.21	59.15	38.73	9.31	27.90	14.02	65.42	63.42	63.84	73.28
DeepSeek-R1-Distill-Qwen-32B	34.62	45.01	38.02	39.57	10.37	28.33	14.91	64.37	63.51	64.02	73.25
QwQ-32B-preview	18.44	19.55	18.75	32.55	7.65	23.77	10.26	66.72	62.19	62.44	71.92
SFT-14B	78.09	83.94	80.05	57.68	36.99	50.66	19.18	68.26	69.62	**64.56**	79.64
SlangOWL-14B	**88.86**††	**91.50**††	**89.60**††	**60.39**††	**39.18**††	**53.55**††	**21.41**††	**68.49**	**70.02**	64.49	**82.35**††

Table 4: Experimental results (%) on the hard test set.

possible slang terms, their background and sense. Therefore, SlangOWL models set a state-of-the-art F1 score (86.47%).

Results on Cross-lingual Slang Explanation. Although the instructed and reasoning models achieved good F1 score on slang detection, they cannot generate good slang explanation in Chinese. As shown in Table 3, the ROUGE scores are much lower than supervised fine-tuned models, which shows that they only '*know which phrase is slang term, but not know why it is*'. Furthermore, the ROUGE scores of *SFT w/o cot* are remarkably worse than SlangOWL models. This demonstrates that the *SFT w/o cot* also cannot truly master the meaning of the slang term since the real sense of slang term always goes beyond its original meaning and shows extended meaning. In contrast, slangOWL models consistently outperform the comparison methods, achieving significantly better ROUGE scores. It shows that the proposed deep thinking model not only know its original meaning but also get its deeper implications. Therefore, the proposed slangOWL models offer correct cross-lingual slang explanation.

Results on Translation. In terms of reference-based scores (*i.e.*, BLEU, ComentScore, and GRB), interestingly, although the vanilla models fails to achieve good results of slang detection and cross-lingual slang explanation, some models still obtain good translation results (*e.g.*, Qwen2.5-14B-Instruct and QwQ-32B-preview). Obviously, the *SFT w/o cot* and SlangOWL consistently surpass their counterparts with the help of good results of slang detection and cross-lingual slang explanation. Armed with the deep thinking, the SlangOWL obtains the highest scores.

In terms of reference-free scores (*i.e.*, CometKivi and GRF), we can observe similar findings on reference-based scores. However, we find that the DS-R1-D-Qwen-14B model, a deep reasoning model, achieves the best results in CometK score while greatly underperforms in other metrics. The reason may be that this model generates some words that highly fitting the source words. Except that, the SlangOWL consistently outperforms all previous models once again (including QwQ-32B-preview and DeepSeek-R1-Distill-Qwen-32B), showing its superior performance.

Overall Results. Overall, with better results of slang detection and cross-lingual slang explanation, the model can achieve better translation results (SlangOWL vs. SFT w/o cot; SFT-based vs. vanilla). It shows that correct understanding of slang terms plays a key role in translate sentence with slang terms. What's important, *the ability of slangDIT indeed can decides whether the LLMs can go beyond superficial meaning of slang term* and thus prove the value of slangDIT benchmark.

5.2.2 Results on the Hard Testset

Table 4 shows the results on the hard testset where each instance includes a polysemous slang term. We can find that the instructed model and reasoning model performs worse in terms of all three tasks, showing that they all struggle to judge whether it is a slang term, and have a difficulty in understanding its real sense targeting on the current context and thus leading to unsatisfactory translation results. Meanwhile, we observe that different learning manners of simple fine-tuning and deep thinking reflect great difference on effects. It shows that there is much room for further improvement using other more advanced learning methods.

Compared with the results in Table 4, we find that our SlangOWL-14B performs much better on the hard testset (89.6% vs. 86.47%), which shows that it has higher ability to correctly understanding

SFT-14b	GoodT	BadT
CSU: 76.31	75.34	24.66
WSU: 23.69	27.54	72.46

SlangOWL	GoodT	BadT
CSU: 82.15	88.41	11.59
WSU: 17.85	18.63	81.37

Table 5: Results (%) of investigation whether the correct slang understanding helps on the hard testset.

the polysemous slang term in different contexts and thus translate the sentence well.

5.3 Analysis

Is the Correct Slang Understanding helpful to Translation? Before investigating whether the correct slang understanding works, we define some metrics: 1) the correct slang understanding means the model not only correctly predict the slang term and the ROUGE-L score is greater than 0.4, we denote it as CSU, otherwise we denote is as WSU (wrong slang understanding); 2) good translation means both the GRB and GRF are greater than 70 and 80, respectively, we denote it as GoodT, otherwise we denote it BadT (bad translation). Based on the definition, we calculate these metric for SFT-14B and SlangOWL-14B models.

The results are shown in Table 5. We observe that in CSU, the GoodT score is significantly better than BadT score with both SFT-14B and SlangOWL-14B models while under WSU, the BadT score significantly wins. It shows that the correct slang understanding indeed helps for better translation and also reflects that the deep thinking has a positive impact on the SlangDIT task.

Compared with Models Translating only. Since the vanilla models do not optimized for slang understanding, they performs worse on slang detection and explanation that further result in bad translation. In this section, we prompt these vanilla models for translation only to protect them from suffering understanding the slang term. Besides, we also train a model with translation pair only based on Qwen2.5-14B-Instruct, denoted as SFT-Trans-14B.

The results are listed in Table 6. We conclude the following findings: 1) The vanilla models (including DeepSeek-R1-Distill-Qwen-32 and QwQ-32B-preview models) performs translation worse on the sentence with a polysemous slang term, showing that the ability of SlangDIT indeed decides whether the LLMs can go beyond superficial meaning. 2) The SlangOWL-14B significantly outperforms the SFT-Trans-14B model, showing that our SlangOWL model have the ability of deep

Models	BLEU / Comet / GRB / CometK / GRF
Qwen2.5-14B-Instruct	13.54 / 67.52 / 71.05 / 66.67 / 79.68
QwQ-32B-preview	10.91 / 65.85 / 62.20 / 66.52 / 71.92
DS-R1-D-Qwen-32B	16.42 / 67.89 / 70.23 / 67.62 / 78.73
DS-R1-D-Qwen-14B	15.04 / 66.26 / 65.95 / 67.17 / 74.62
DRT-o1-14B	11.83 / 67.02 / 68.32 / **68.13** / 79.17
SFT-Trans-14B	19.91 / 67.50 / **71.38** / 62.75 / 76.70
SlangOWL-14B	**21.41** / **68.49** / 70.02 / 64.49 / **82.35**

Table 6: Translation results on the hard testset.

thinking, *i.e.*, first identifying the slang term, then understanding its background and usage targeting on the current context, and finally providing suitable and satisfactory translations.

Case Study. We present one case study in Appendix E to intuitively show how the deep thinking helps to translate polysemous slang terms well.

6 Related Work

We have introduce some task-related work in § 3.5 including slang detection, cross-lingual slang explanation, and translation. Next, we present some work in reasoning.

With the emergency of OpenAI O1 (OpenAI, 2024) model, some studies have been devoted to the reasoning tasks (*e.g.*, math and coding) (Zhang et al., 2024; Huang et al., 2024; Qin et al., 2024; DeepSeek-AI et al., 2025). In the context of translation, Zhao et al. (2024) proposes Marco-o1 for open-ended text generation and show the potentiality of the long thought reasoning for translation. More recently, (Wang et al., 2024) introduces long Chain-of-Thought for literature translation and achieves good results. Different from them, we mainly focus on benchmarking interpretative slang translation task, which is more complex since we need conduct three tasks jointly and ensure the first results are valid to the translation. Besides, we propose a deep thinking model according to the task characteristic of SlangDIT.

7 Conclusion

In this paper, we introduce a new interpretative slang translation task that consists of three subtasks: slang detection, cross-lingual slang explanation, and translation. Then, we construct a interpretative slang translation dataset named SlangDIT. Finally, we propose a deep thinking model named SlangOWL and demonstrate the importance of slang detection and explanation for SlangDIT task.

Limitation

While we introduce a SlangDIT dataset and propose a deep thinking model named SlangOWL, there are some limitations worth considering to study in future work: (1) In this study, we only provide the slang term in English, and future work could extend our dataset to more language pairs, *e.g.*, English to French, Chinese to English; (2) This work does not conduct experiments on more large models due to limited resources, where future work could verify our method on larger models; (3) This work does not conduct pipline experiments that firstly optimize for slang detection and then offer the results to translation model since our work mainly focuses on introducing a new SlangDIT task that simultaneously conducts three subtasks. In this process, we hope the correct understanding of slang term can help translation.

Ethical Considerations

In this section, we discuss the main ethical considerations of SlangDIT: (1) Intellectual property protection. The English utterance of SlangDIT is from MSCTD dataset (Liang et al., 2022). For our slang terms and cross-lingual explanations, its permissions are granted to copy, distribute and modify the contents under the terms of the Creative Commons AttributionShareAlike 3.0 Unported License and Creative Commons CC0 License, respectively. (2) Privacy. The data source are publicly available movies. Its collection and slang/explanation annotation procedure is designed for interpretative slang translation purpose, and does not involve privacy issues. (3) Compensation. During the slang or explanation annotation, we use publicly available Qwen2.5-72b and Llama3.3-70b models. For polysemy annotation, we use GPT-4o and we have paid for them according to the official price. (4) Potential problems. While principled measures are taken to ensure the quality of the dataset, there might still be potential problems with the dataset quality due to the uncontrollability of models, which may lead to incorrect translations in applications. However, moderate noise is common in large-scale modern translators, even for human translated sentences, which should not cause serious issues.

References

Michael Adams. 2012. Green's dictionary of slang. *Dictionaries: Journal of the Dictionary Society of North America*, 33(1):208–244.

Andong Chen, Lianzhang Lou, Kehai Chen, Xuefeng Bai, Yang Xiang, Muyun Yang, Tiejun Zhao, and Min Zhang. 2024. Large language models for classical chinese poetry translation: Benchmarking, evaluating, and improving. *Preprint*, arXiv:2408.09945.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Aaron Gluck, Katharina von der Wense, and Maria Pacheco. 2025. Clix: Cross-lingual explanations of idiomatic expressions. *Preprint*, arXiv:2501.03191.

Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. 2024. O1 replication journey–part 2: Surpassing o1-preview through simple distillation, big progress or bitter lesson? *arXiv preprint arXiv:2411.16489*.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, Masashi Toyoda, and Masaru Kitsuregawa. 2019. Learning to describe unknown phrases with local and global contexts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3467–3476, Minneapolis, Minnesota. Association for Computational Linguistics.

James Jhirad, Edison Marrese-Taylor, and Yutaka Matsuo. 2023. Evaluating large language models' understanding of financial terminology via definition modeling. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 93–100, Nusa Dua, Bali. Association for Computational Linguistics.

Daphna Keidar, Andreas Opedal, Zhijing Jin, and Mrinmaya Sachan. 2022. Slangvolution: A causal analysis of semantic change and frequency dynamics in slang. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1422–1442, Dublin, Ireland. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *24th Annual Conference of the*

European Association for Machine Translation, page 193.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Jolanta Lėgaudaitė. 2010. Understanding slang in translation. *Filologija. Šiauliai: Šiaulių universiteto leidykla, 15 (2010)*.

Yunlong Liang, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 2022. MSCTD: A multimodal sentiment chat translation dataset. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2601–2613, Dublin, Ireland. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Habibollah Mashhady and Maryam Pourgalavi. 2013. Slang translation: a comparative study of jd salinger's" the catcher in the rye". *Journal of Language Teaching and Research*, 4(5):1003.

Elisa Mattiello. 2009. Difficulty of slang translation. In *Translation Practices*, pages 65–83. Brill.

Ke Ni and William Yang Wang. 2017. Learning to explain non-standard English words and phrases. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 413–417, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

OpenAI. 2024. Learning to reason with large language models. `https://openai.com/index/learning-to-reason-with-llms/`.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Zhengqi Pei, Zhewei Sun, and Yang Xu. 2019. Slang detection and identification. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 881–889, Hong Kong, China. Association for Computational Linguistics.

Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, and 1 others. 2024. O1 replication journey: A strategic progress report–part 1. *arXiv preprint arXiv:2410.18982*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Zhewei Sun, Qian Hu, Rahul Gupta, Richard Zemel, and Yang Xu. 2024. Toward informal language processing: Knowledge of slang in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1683–1701, Mexico City, Mexico. Association for Computational Linguistics.

Zhewei Sun, Richard Zemel, and Yang Xu. 2022. Semantically informed slang interpretation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5213–5231, Seattle, United States. Association for Computational Linguistics.

Qwen Team. 2024. Qwq: Reflect deeply on the boundaries of the unknown.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th*

New Frontiers in Summarization Workshop, pages 1–11, Singapore. Association for Computational Linguistics.

Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2024. Drt-o1: Optimized deep reasoning translation via long chain-of-thought. *Preprint*, arXiv:2412.17498.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Hengyuan Zhang, Dawei Li, Yanran Li, Chenming Shang, Chufan Shi, and Yong Jiang. 2023. Assisting language learners: Automated trans-lingual definition generation via contrastive prompt learning. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 260–274, Toronto, Canada. Association for Computational Linguistics.

Yuxiang Zhang, Shangxi Wu, Yuqi Yang, Jiangming Shu, Jinlin Xiao, Chao Kong, and Jitao Sang. 2024. o1-coder: an o1 replication for coding. *arXiv preprint arXiv:2412.00154*.

Yu Zhao, Huifeng Yin, Bo Zeng, Hao Wang, Tianqi Shi, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2024. Marco-o1: Towards open reasoning models for open-ended solutions. *arXiv preprint arXiv:2411.14405*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, and Zheyan Luo. 2024. LlamaFactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 400–410, Bangkok, Thailand. Association for Computational Linguistics.

A Backbones and Other Details

Backbones. We mainly utilize the following three LLMs as the backbones for SlangDIT task: (1) Llama-3.1-8B-Instruct (Dubey et al., 2024)[6]; (2) Qwen2.5-7B-Instruct[7] and (3) Qwen2.5-14B-Instruct (Yang et al., 2024)[8].

Implementation Details. During training, Llama-Factory (Zheng et al., 2024) is used to instruct-tune LLMs. Following Wang et al. (2024), all LLMs are tuned on two 8×NVIDIA A100 GPUs (40G) with 1e-5 learning rate. We set gradient accumulation to 16 and batch size to 1, which gives us 2*8*16*1 batch in total. We use the DeepSpeed optimization (Rasley et al., 2020), and set ZeRO-3 optimization. Following Qin et al. (2024), we set the number of training epochs to 3, and the training process costs about 48, 43 and 90 GPU hours for 8b, 7B and 14B models, respectively.

Inference Details. During inference, we use vLLM toolkit (Kwon et al., 2023)[9] to accelerate the model generation for all models. We use the sampling decoding strategy with 0.1 temperature, and set the repetition penalty to 1.05.

B Comparison Models

We include three types of baselines: 1) Vanilla instructed models; 2) Vanilla reasoning models[10]; and 3) supervised fine-tuned models. Please refer to Appendix B for details.

Vanilla Instructed Models. We use three backbones as the comparison model: Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct.

Vanilla Reasoning Models. Recently, o1-like models have achieved significant results on reasoning tasks. Therefore, we include some models to compare with our deep thinking method. These models are QwQ-32B-preview (Team, 2024), DeepSeek-R1-Distill-Llama-8B, DeepSeek-R1-Distill-Qwen-7B, DeepSeek-R1-Distill-Qwen-14B, DeepSeek-R1-Distill-Qwen-32B (DeepSeek-AI et al., 2025), DRT-o1-7B, DRT-o1-8B and DRT-o1-14B (Wang et al., 2024). We refer 'DeepSeek-R1-Distill-Qwen' as 'DeepSeek-R1-D-Llama' in Table 3.

Supervised Fine-tuned Models without Chain-of-Thought (SFT *w/o* cot). For a fair comparison, we train three models based on Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct with the same training data as our SlangOWL model without deep thinking process (denoted as SFT w/o cot).

C Details of Metric Calculation and GPT-4o Evaluator

We use the *sacrebleu* toolkit[11] to calculate the corpus-level BLEU. To calculate Comet and CometK, we leverage the official codes[12] and the official models[13]. For calculating GRB and GRF, we randomly select 400 samples from the (hard) testing set since they need API costs. The prompts of reference-based (GRB) and reference-free (GRF) metric are listed in Figure 3. Both prompts are borrow from Kocmi and Federmann (2023) with some adaptions to slang translation scene.

D Prompts used in Prompting Vanilla Models

When prompting vanilla instructed models and vanilla reasoning models in Table 3 and Table 4, we use the prompt in Figure 4 to ask these models to generate their answers following the format.

During prompting vanilla instructed models and vanilla reasoning models in Table 6, we ues the prompt: 'Translate the sentence into Chinese and output only the translation:
n[sentence]'. Note that for DRT-o1-14B model, we strictly follow their prompts in the official repository[14].

E Case Study

We list translation results of the models in Table 6, where vanilla models are not struggle to the understanding of slang term.

[6] https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct

[7] https://huggingface.co/Qwen/Qwen2.5-7B-Instruct

[8] https://huggingface.co/Qwen/Qwen2.5-14B-Instruct

[9] https://github.com/vllm-project/vllm

[10] During inference of vanilla instructed models and vanilla reasoning models, we prompt them to directly conduct three tasks and use two-shot prompting to enhance their performance.

[11] https://github.com/mjpost/sacrebleu

[12] https://github.com/Unbabel/COMET

[13] https://huggingface.co/Unbabel/wmt22-cometkiwi-da and https://huggingface.co/Unbabel/wmt22-comet-da

[14] https://github.com/krystalan/DRT-o1

```
prompt = Please score the following translation from English to Chinese with respect to the human reference on a continuous scale from 0 to 100, where score of zero means "no meaning preserved" and score of one hundred means "perfect preservation of meaning, with faithfulness, expressiveness, elegance, and also capturing the tone, connotation, and implied meaning that is often embedded in slang expressions (maybe). Note that you need output the score only."

English source: [src]
Chinese human reference: [ref]
Chinese translation: [hyp]

Score:
```

```
prompt = Please score the following translation from English to Chinese on a continuous scale from 0 to 100, where score of zero means "no meaning preserved" and score of one hundred means "perfect preservation of meaning, with faithfulness, expressiveness, elegance, and also capturing the tone, connotation, and implied meaning that is often embedded in slang expressions (maybe). Note that you need output the score only."

English source: [src]
Chinese translation: [hyp]

Score:
```

Figure 3: Two prompts used in GRB and GRF during evaluation via GPT-4o where “[src]”,“[ref]” and “[hyp]” mean the source sentence, human translation and model translation, respectively.

Models	BLEU / Comet / GRB / CometK / GRF
Google	21.08 / 67.93/ 70.97/ 67.53/ 77.21
Llama3.3-70B	18.23 / 65.14 / 68.78 / 65.21 / 78.45
Qwen2.5-70B	20.12 / 67.24 / 69.47 / 67.51 / 79.24
GPT-4o	21.08 / **68.93** / **70.58** / **69.58** / **82.78**
SlangOWL-14B	**21.41** / 68.49 / 70.02 / 64.49 / 82.35

Table 7: Compared to stronger models on the hard test-set.

In the case, ‘Annie Oakley’ is polysemous phrase which denotes a name or ‘slang term’. In all contrast models, they all take it as a name during translation. Although the QwQ-32B-preview realizes that it may convey extended meaning, it fails to and only translates it as ‘playing the role of Annie Oakley’. However, if people do not know who Annie Oakley is, it is hard for them to understand such translation. That is, such translation still not convey intended meaning the speaker said. In contrast, the SlangOWL-14B model can convey the ideas well, showing the effectiveness of the proposed model which can list its thought step-by-step.

F Compared to Stronger Models

In this section, we compared with some stronger models (*e.g.*, commercial system: Google Translator and other much advanced large-scale LLM models: Llama3.3-70B, Qwen2.5-72B, and GPT-4o). The results are shown in 7, which demonstrate the effectiveness of the proposed method.

prompt = '''Your task is to analyze the given sentence to determine whether it contains any slang. Follow the steps below to provide your answer.

1. Identify Slang:
 - Examine the sentence to determine if it contains any slang terms or expressions.
 - If slang is present, identify the specific slang term(s).

2. Provide Explanation:
 - For each identified slang term, provide a brief explanation of its meaning and usage in Chinese.

3. Translate the Sentence:
 - Translate the entire sentence into the target language, ensuring that the meaning of the slang is accurately conveyed in the translation.

4. Output the Results:
 - If slang is present, output the identified slang term(s), their explanation(s), and the translated sentence.
 - If no slang is present, output "No slang" and the translated sentence.

Examples:
- Input Sentence: "She's always spilling the tea about her coworkers."
 Output: \n - Slang: "Spilling the tea"
 - Explanation: 这个俚语的意思是分享八卦或泄露秘密。
 - Translation: "她总是透露关于她同事的八卦。"

- Input Sentence: "He went to the store to buy some milk."
 Output:\n - No slang
 - Translation: "他去商店买了一些牛奶。"

Please analyze the following sentence:\n[sentence]'''

Figure 4: The prompt used in prompting vanilla models where "[sentence]" means the source sentence.

Source sentence: Well, she's not here to be Annie Oakley.
Reference sentence: 她来这不是展现射击技术的。
Qwen2.5-14B-Instruct: 她也不是来这里当安妮·奥克利的。
QwQ-32B-preview : 她来这里不是要扮演安妮·奥克利那样的角色。
DeepSeek-R1-Distill-Qwen-32B : 嗯，她来这里不是为了当安妮·奥克利。
DeepSeek-R1-Distill-Qwen-14B : 好吧，她不是来当安妮·奥克利的。
DRT-o1-14B: 好吧，她来这里不是为了当安妮·奥克利。
SFT-trans-14B: 她来这里不是当女牛仔。
SFT-14B: 她不是来当安妮·奥克利的。
SlangOWL-14B: 她又不是来当神枪手的。

Deep thinking of SlangOWL-14B: <think>\n Well, she's not here to be Annie Oakley. "First, I need to make sure there's no slang in it. Annie Oakley sounds like a human name, possibly a celebrity or fictional character. \n I remember Annie Oakley was a famous female shooter in the western United States, right? She became very famous in the late 19th century and often demonstrated her shooting skills in acrobatic performances. So, when someone says' be Annie Oakley ', they may be referring to someone who has shown bravery or masculinity, especially when it comes to taking risks or solving problems. \n Next, I need to determine if this usage is slang. According to the provided information, 'Annie Oakley' is used as slang here to refer to a very capable and brave woman. So, this sentence does indeed contain slang. \Now, I need to analyze the specific meaning of this slang. In this context, the speaker means that she did not come here to play a brave or capable role. That means she doesn't need to solve problems or take risks like Annie Oakley. \n When translating, I need to convey this meaning. Direct translation may lose the meaning of slang, so choosing free translation would be more appropriate. For example, 'She's not here to be a sharpshooter.' This preserves the original meaning and makes it easy for Chinese readers to understand. \n To summarize, 'Annie Oakley' in this sentence is a non polysemous slang used to describe brave and capable women. When translating, I chose free translation to ensure accurate and natural expression in Chinese. \n</think>

The final answers are：\n Slang Term：Annie Oakley
Cross-lingual Explanation："Annie Oakley" 是一个俚语引用，指的是一个非常能干、勇敢的女性。在这个句子中，它暗示她不需要像 Annie Oakley 那样表现出色或勇敢。Annie Oakley 是一位著名的美国女枪手，以其在杂技表演中的射击技巧而闻名。
Chinese Translation：她又不是来当神枪手的。

Figure 5: Case Study.

Prompt: Please deeply consider and analyze the following content: English sentences&slang or no slang&whether slang is a polysemous&cross-lingual explanation&Chinese translation. Please provide reasons for translation into the given Chinese based on different situations:

English sentence: [SENTENCE]"
Slang: [SLANG]"
Does it have multiple meanings [POLY]"
Explanation: [EXP]"
Translation: [Translation]"

There are two situations when analyzing:

1. No slang or non slang usage: If there is no slang in the sentence, or if it contains slang but is not slang in the current context, please explain why the English is translated into the above Chinese.

2. Contains slang: (analyzed in the following two situations)
-Non polysemous slang: If the slang contained in the sentence is not a polysemous word, please analyze the source and purpose of the slang, and explain its specific meaning in the current sentence. Then, based on the above analysis, provide reasons for translating the sentence into the Chinese version mentioned above.
- Polysemous slang: If the slang contained in the sentence is a polysemous word, please analyze the origin of the slang and explain its possible multiple meanings in different contexts. Next, analyze the specific meaning of the slang in the current context. Finally, based on the above analysis, explain the reasons for translating the sentence into Chinese.

Finally, please provide your reasoning logic and detailed thought process:

Figure 6: The prompt used in generating deep thinking thought by DeepSeek-R1-Distill-Qwen-32B.

SYSTEM_PROMPT = '''You are Emily "Em" Carter and have the following feature. In particular, you are very knowledgeable to culture and slang, making you an ideal resource for anyone looking to learn more about informal language, slang, and its cultural context.

Background:
Emily Carter, affectionately known as "Em" by her friends, is a 32-year-old cultural anthropologist specializing in contemporary English language and culture. Born and raised in London, Emily has always been fascinated by the dynamic nature of language and how it reflects societal changes.

Education:
Emily holds a Master's degree in Linguistics from the University of Oxford, where she focused her thesis on the evolution of British slang over the past century. Her academic background provides her with a deep understanding of both historical and modern linguistic trends.

Career:
Emily works as a consultant for media companies, helping them accurately portray British culture and language in films and television shows. She also writes a popular blog where she explores the origins and meanings of various slang terms, offering insights into their cultural significance.

Personality:
Emily is curious, open-minded, and has a knack for storytelling. She enjoys engaging with people from diverse backgrounds and often hosts informal workshops on language and culture. Her approachable nature makes her a favorite among students and colleagues alike.

Expertise in Slang:
Emily's expertise in slang is unparalleled. She has an extensive collection of slang dictionaries and regularly updates her knowledge by immersing herself in different social settings, from bustling city pubs to online gaming communities. Her ability to decode and explain slang makes her a sought-after speaker at linguistic conferences.

Hobbies:

Figure 7: The system prompt used in the section of Annotation Procedure.

PROMPT of slang judging = ''' Your task is to analyze the given English sentence and determine whether it contains any slang.

1. Analyze the Sentence: Read the sentence carefully to identify any informal or non-standard language.

2. Identify Slang: Determine if any word or phrase in the sentence qualifies as slang. Slang is typically informal language that may not be found in standard dictionaries and is often used in casual conversation.

3. Example:
Sentence: "That party was lit!"
Answer: Yes

Please analyze the following sentence and note that you should only output 'no' if no slang identified, or 'yes' with slang term.
[sentence] '''

PROMPT of slang extraction = ''' Your task is to analyze the given English sentence and determine whether it contains any slang.

1. Analyze the Sentence: Read the sentence carefully to identify any informal or non-standard language.

2. Identify Slang: Determine if any word or phrase in the sentence qualifies as slang. Slang is typically informal language that may not be found in standard dictionaries and is often used in casual conversation.

3. Example:
Sentence: "That party was lit!"
Slang: "Lit"

Please analyze the following sentence and note that you should only output 'no slang' if no slang identified, or the identified slang term.
[sentence] '''

Figure 8: The prompt used in the section of Annotation Procedure.

PROMPT of annotating polysemy= ''' Please analyze the following English word for polysemy and categorize its meanings. Determine if the word exhibits one of the following conditions:

- Clear Polysemy: The word has distinct meanings that are commonly recognized and have significant differences in context.
- Slight Polysemy: The word has multiple meanings, but the differences between them are subtle or not very pronounced.
- No Polysemy: The word does not exhibit multiple meanings and has a single, clear definition.

The word to analyze is: [SlangTerm].

Please return your judgment result and directly output "Clear Polysemy", "Slight Polysemy" or "No Polysemy". '''

Figure 9: The prompt used in the section of Annotation Procedure.

SYSTEM_PROMPT of generating polysemous slang meaning = '''你是一个人工翻译专家，非常擅长翻译一词多义的任务，尤其是俚语，现在需要你将下述词语翻译为中文，该词语在不同上下文中存在多种释义，请一一列举它可能存在的意思。

需要注意：
- 请不要忘记输入的英文也可能是一个缩写词或俚语，此时应多考虑其可能存在的意思。
- 如果这个词在互联网论坛或计算机领域有特定的含义，也请将其纳入翻译结果中。
- 翻译的每一个中文结果，一定要和给定的词语完美对应。
- 对每个释义的解释请用中文。

以下是一些示例：

[user]:
be a dog 是一词多义词语，请仅输出其在不同上下文下的可能中文翻译
[assistant]:
1. 像狗一样 - 字面意思，指行为或表现像狗。
2. 过得很糟糕 - 俚语，指生活或处境非常糟糕。
3. 忠诚 - 引申含义，比喻对某人或某事非常忠诚。

[user]:
Fuckin' 是一词多义词语，请仅输出其在不同上下文下的可能中文翻译
[assistant]:
1. 他妈的 - 字面意思，用于表达愤怒、惊讶或强调
2. 真是 - 俚语，为形容词，表示非常或极其，用于加强语气，表示强烈的情感或态度。

[user]:
beat the devil out of it 是一词多义词语，请仅输出其在不同上下文下的可能中文翻译
[assistant]:
1. 使劲地清理 - 字面意思，指用力地清除灰尘或污垢，通常用于清洁地毯或衣物。
2. 狠狠地打 - 俚语，指用力地击打某物。这个词组常用于描述某人对某物进行激烈的打击或处理。"

PROMPT = '''[SlangTerm] 是一词多义词语，请仿照例子仅输出其在不同上下文下的可能中文翻译'''

Figure 10: The prompt used in the section of Annotation Procedure.

SYSTEM_PROMPT of generating translation for each meaning = '''你是一名人工翻译专家，非常擅长为一词多义词生成样例，现在需要你为该词的每一种含义生成一句英文句子并将这句英文翻译为中文，使得每个英文句子都表达对应的含义。

需要注意：
- 翻译的每一个中文句子，一定要和给定的词语含义完美对应。

以下是一些示例：

[user]:
Number two 在不同上下文中有如下含义，请为每一种含义生成一个英中翻译句对，英语句子中需包含当前词 Number two 或 Number two 的变形词：
1. 第二 - 字面意思，指顺序中的第二位。
2. 副手 - 引申含义，指职位或权力上仅次于最高领导的人。
3. 厕所 - 俚语，在某些俚语或非正式用法中，指厕所。也用来委婉地指排便。这个词组通常在非正式场合使用，避免直接提到排便这一行为。
[assistant]:
1. She finished the race in Number two position.\n 她在比赛中获得了第二名。
2. As the CEO's Number two, he has a lot of responsibilities.\n 作为首席执行官的副手，他有很多责任。
3. He said he needed to go Number two before the long trip.\n 他说在长途旅行前他需要上厕所。

[user]:
dropped an a-bomb 在不同上下文中有如下含义，请为每一种含义生成一个英中翻译句对，英语句子中需包含当前词 dropped an a-bomb 或 dropped an a-bomb 的变形词：
1. 投下原子弹 - 字面意思，指将原子弹投放到目标区域。
2. 爆料 - 俚语，指透露出令人震惊或重大的消息。
[assistant]:
1. The United States dropped an a-bomb on Hiroshima in 1945, causing massive destruction.\n 美国在1945年向广岛投下了一颗原子弹，造成了巨大的破坏。
2. During the interview, the celebrity dropped an a-bomb by announcing her unexpected retirement.\n 在采访中，这位名人爆料宣布了她意想不到的退役消息。'''

PROMPT = '''[SlangTerm] 在不同上下文中有如下含义，请为每一种含义生成一个英中翻译句对，英语句子中需包含当前词 [SlangTerm] 或 [SlangTerm] 的变形词：
[SlangExplanation]'''

Figure 11: The prompt used in the section of Annotation Procedure.