DeRea: Improving Idiom Translation with Detect-Retrieve-Arbitrate Reasoning

Rongqing Jiang[1] **Xuebo Liu**[1*] **Shengxin Liu**[2] **Yutong Wang**[1]
Min Zhang[3] **Shimin Tao**[3] **Daimeng Wei**[3] **Min Zhang**[1]
[1]Institute of Computing and Intelligence, Harbin Institute of Technology, Shenzhen
[2]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
[3]Huawei Translation Services Center
{jiangrongqing,wangyutong}@stu.hit.edu.cn
{liuxuebo,sxliu,zhangmin2021}@hit.edu.cn
{zhangmin186,taoshimin,weidaimeng}@huawei.com

Abstract

Idiom translation remains a formidable challenge for Large Language Models (LLMs), as the constraints of static parametric memory and the noise in sentence-level retrieval often lead to literal misinterpretations. To address this, we propose DeRea, a detect-retrieve-arbitrate framework. The system employs a preference-aligned detector to identify idiomatic spans by reasoning over semantic conflicts between literal and contextual meanings. Subsequently, an idiom-centric translator invokes a fine-tuned embedding model to efficiently retrieve canonical definitions from an external knowledge base. The translator then utilizes a dual-path arbitration mechanism to select the optimal rendering by weighing the retrieval-augmented translations against direct translation. To evaluate our framework, we introduce LoMI, a high-difficulty benchmark with low data contamination. Experimental results demonstrate that DeRea significantly enhances performance across various model scales, improving GPT-5-mini by over 5.2 points in both idiomatic quality and consistency according to LLM-based metrics. Furthermore, evaluations on an emerging slang dataset from Urban Dictionary validate the potential of our approach in handling novel and evolving linguistic data. Our code is available at https://github.com/jrongqing/DeReA.

SRC	The boy was foaming at the mouth , throwing things around.
REF	男孩/boy 极度/extremely 愤怒/angry 乱扔/throwing around 东西/things
GPT5[1]	男孩/boy ~~口吐白沫/white foam coming out of his mouth~~ 乱扔/throwing around 东西/things
Detect	foaming at the mouth
Retrieve	foam at the mouth: be extremely angry
Arbitrate	极度愤怒: more contextually appropriate 口吐白沫: conflict with context
Translate	男孩/boy 极度/extremely 愤怒/angry 乱扔/throwing around 东西/things

Table 1: An example of our approach. DeRea employs a Detect-Retrieve-Arbitrate pipeline to successfully rectify erroneous translations generated by GPT-5.

1 Introduction

Idiomatic expressions are non-compositional units where figurative meanings diverge from literal interpretations. Although Large Language Models (LLMs) have significantly advanced Machine Translation (MT), the accurate interpretation of idiomatic expressions remains hampered by the inherent over-compositionality of the transformer architecture (Dankers et al., 2022; Mi et al., 2025). Even more challenging is the increase of informal idiomatic expressions catalyzed by the internet, whose semantics evolve rapidly with shifting social contexts (Wang et al., 2024b; Zheng et al., 2024). Consequently, increasing LLM size alone fails to solve idiom translation errors. (Liu et al., 2023a,b; Wuraola et al., 2024; Pang et al., 2025).

Recent research has transitioned from traditional parameter optimization toward two primary paradigms: reasoning enhancement and external knowledge integration. The first paradigm leverages the reasoning of LLMs to conduct autonomous semantic inference for mitigating literal errors (He et al., 2024; Liang et al., 2025; Wang et al., 2024a). However, single models struggle to align detection with translation and are restricted by internal knowledge when encountering novel expressions. The second paradigm adopts Retrieval-Augmented Generation (RAG) to leverage external knowledge bases (Li et al., 2024a; Donthi et al., 2025), yet its efficacy is frequently hindered by retrieval noise. Since idioms occupy narrow spans within sentences, coarse-grained sentence-level retrieval often introduces irrelevant information that misleads the model. To address this, a more robust

*Xuebo Liu is the corresponding author.

[1]gpt-5-2025-08-07

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6603–6621
July 2-7, 2026

paradigm should leverage semantic detection to extract idiomatic expressions, followed by idiom-centric RAG for targeted translation enhancement, as shown in Table 1.

To this end, we propose DEREA: a detect-retrieve-arbitrate reasoning process for idiomatic expression translation that emulates the cognitive workflow of human translators. Recognizing that intrinsic LLM detection is often hindered by morphological variations and the rapid evolution of language, we first develop a preference-aligned detector. By leveraging the semantic conflict between an idiom's literal and contextual meanings to construct preference data, we optimize the detector's capabilities via reinforcement learning. This detector identifies whether an input sentence contains idiomatic expressions and explicitly extracts them from the text. Following idiom detection, the system generates translation candidates through both direct and RAG-based paths, then employs comparative arbitration to select the optimal translation.

Furthermore, to ensure a robust and unbiased evaluation, we constructed a high-difficulty idiom translation benchmark alongside a test set curated from recent Urban Dictionary entries to capture evolving linguistic trends. Extensive experiments across various scales demonstrate that DEREA consistently surpasses all baselines. For lightweight models, our framework yields substantial improvements in both COMET and LLM-based metrics, while for large-scale models such as Qwen3-32B and GPT5-mini, it significantly enhances idiomatic semantic accuracy and consistency.

Our contributions are summarized as follows:

- We introduce LoMI, a high-difficulty idiom translation benchmark with a low contamination rate.
- We propose DEREA, a detect-retrieve-arbitrate reasoning process for idiom translation. Experimental results show that our method consistently enhances idiom translation performance across various model scales.
- We leveraged Urban Dictionary to construct an Emerging Slang benchmark, validating the potential of our approach in handling novel and evolving linguistic data.

2 Related Works

MT of Idiomatic Expressions Idiomatic expressions, particularly idioms, pose significant linguistic and cognitive challenges to machine translation as their meanings cannot be directly derived from constituent words (Lin, 1999; Tzou et al., 2017). Consequently, Transformer-based LLMs frequently struggle to model context-dependent semantic extensions, leading to literal translation errors (Castaldo and Monti, 2024; Donthi et al., 2025). Although Chain-of-Thought (CoT) prompting has demonstrated efficacy in enhancing reasoning across various domains including translation (Zhang et al., 2024; Zhao et al., 2024; Wang et al., 2025a; Liang et al., 2025), existing approaches often rely on a single model to concurrently handle self-retrieval and translation. More critically, idioms proliferate across diverse sociocultural origins, exhibiting rapid evolution and complex morphological variations (Liu et al., 2023a; Wuraola et al., 2024; Pang et al., 2025), rendering static single-model inference inadequate for capturing their dynamic nature. We address this limitation by leveraging multi-model collaboration (Zhuge et al., 2024; Wu et al., 2024; Zhang et al., 2025; Wang et al., 2025c, 2026) to mimic the human cognitive workflow of "detect-retrieve-arbitrate".

RAG of LLM-based MT RAG has become a pivotal paradigm for grounding LLMs in external knowledge, thereby mitigating hallucinations inherent in static training data (Lewis et al., 2020; Xu et al., 2024; Asai et al., 2024; Chang et al., 2025; Rao et al., 2025). In modern machine translation, frameworks primarily leverage prompt-level RAG—utilizing translation memories and in-context learning—to supplement models with domain-specific or stylistic cues (Gao et al., 2023; Li et al., 2025; Baek et al., 2023; Li et al., 2024a; Chen et al., 2025; Wang et al., 2025b). This necessitates balancing the model's internal parametric memory with external non-parametric evidence (Li et al., 2024b). However, these methods often rely on sentence-level retrieval, which can introduce significant noise when the global semantics of a retrieved pair conflict with the local, non-compositional meaning of an idiom (Moslem et al., 2023). Furthermore, standard RAG often leads to "over-correction" or knowledge interference, where unreliable external data overshadows the model's intrinsic linguistic capabilities (Yao and Fujita, 2024). Unlike traditional monolithic RAG pipelines, DEREA addresses these gaps by explicitly detecting semantic conflicts between literal and contextual meanings. By integrating a fine-grained

retrieval and arbitration mechanism, our approach ensures that the translation process is selectively guided by contextually verified knowledge rather than unfiltered references.

3 Methodology

To address single-model limitations, we propose DEREA (Figure 1), a detect-retrieve-arbitrate framework that mimics human cognition by decomposing idiom translation into specialized stages supported by external knowledge. The formal inference algorithm is detailed in Appendix A.1.

While conventional models typically identify idioms based on rote memorization, they often falter when faced with rapidly evolving expressions. Inspired by how human translators detect unfamiliar idioms by identifying the semantic conflict between a phrase's literal meaning and its context, we develop a preference-aligned detector that transcends the boundaries of predefined model memory. To mitigate RAG-induced noise and fluency issues, we implement a dual-path arbitration mechanism that critically selects the optimal rendering from both direct and RAG-enhanced candidates.

3.1 Preference-Aligned Idiom Detection

Formally, given a source sentence x, we train M_d to sequentially generate a triplet $y = (r, l, i)$. Here r denotes the detection CoT, l is a classification label such that $l \in \{\text{YES}, \text{NO}\}$ and i denotes the specific idiomatic span identified in the sentence. In this work, we simplify the task by focusing on the detection of a single candidate idiom per instance. The model identifies non-compositional expressions by analyzing the semantic discontinuity between literal and contextual meanings, thereby transcending the limitations of static parametric memory. The specific prompt is detailed in the Appendix A.2.

If the detection token l is NO, the model yields a literal translation; otherwise, it proceeds to the subsequent retrieval stage.

Answer-Guided SFT Data Construction Given the flexibility of idiom distribution and the complexity of morphological variations, we first construct Supervised Fine-Tuning (SFT) data with an "answer-guided" strategy. Specifically, we provide the same model used for SFT stage with the source sentence x along with its ground truth label l^* and idiomatic spans i^*. The model is then prompted to "reason backward" to generate a coherent reasoning chain r^* that explains the semantic conflict.

We utilize Few-shot prompting to stimulate the model's capability for "semantic conflict detection." For instance, in the sentence "When the share market crashed his fingers were burnt...", the reasoning path r^* demonstrates: (1) interpreting "burn fingers" as physical harm, (2) analyzing the financial context (stock market crash), (3) highlighting the semantic incongruity in the current context–the subject suffered financial loss rather than physical harm–thereby classifying it as a non-compositional expression. We denote the policy of the detector model M_d as π_θ, the training objective is to maximize the likelihood of the reasoning chain and the detection result:

$$L_{\text{SFT}}(\theta) = -\mathbb{E}_{(x,y^*)\sim D_{\text{SFT}}}\big[\log \pi_\theta(y^* \mid x)\big]. \quad (1)$$

Negative-Mining DPO Data Sampling However, SFT alone often leads to over-sensitivity in complex linguistic contexts. To enhance detection precision, we employ Direct Preference Optimization (DPO) using a preference dataset $D_{\text{DPO}} = \{(x, y^*, y^-)\}$, where y^* is the gold standard response and y^- is the rejected sample.

To construct the preference dataset D_{DPO}, we first employ the base model to generate candidate outputs $y = (r, l, i)$ for each source sentence. By assessing the consistency between the generated y and the gold standard $y^* = (r^*, l^*, i^*)$, we sample erroneous responses as rejected data y^- according to the following hierarchical criteria: (1) Judgment Errors: The model misclassifies the presence of idioms; (2) Extraction Errors: The model correctly identifies the presence of an idiom but fails to accurately extract the specific spans or canonical forms.

By contrasting these structured, string-verified failures against the ground truth, the DPO objective is defined as:

$$L_{\text{DPO}}(\theta) = -\mathbb{E}_{(x,y^*,y^-)\sim D_{\text{DPO}}}\left[\log \sigma\left(\beta\Delta\right)\right], \quad (2)$$

where the implicit reward difference Δ between the chosen and rejected sequences is given by:

$$\Delta = \log \frac{\pi_\theta(y^* \mid x)}{\pi_{\text{ref}}(y^* \mid x)} - \log \frac{\pi_\theta(y^- \mid x)}{\pi_{\text{ref}}(y^- \mid x)}. \quad (3)$$

Here, π_{ref} is the reference SFT model, β is the temperature parameter.

3.2 Idiom-centric Retrieval-Augmented Translation

Idiom retrieval is conceptualized as a specialized tool invoked to provide the translation model M_t

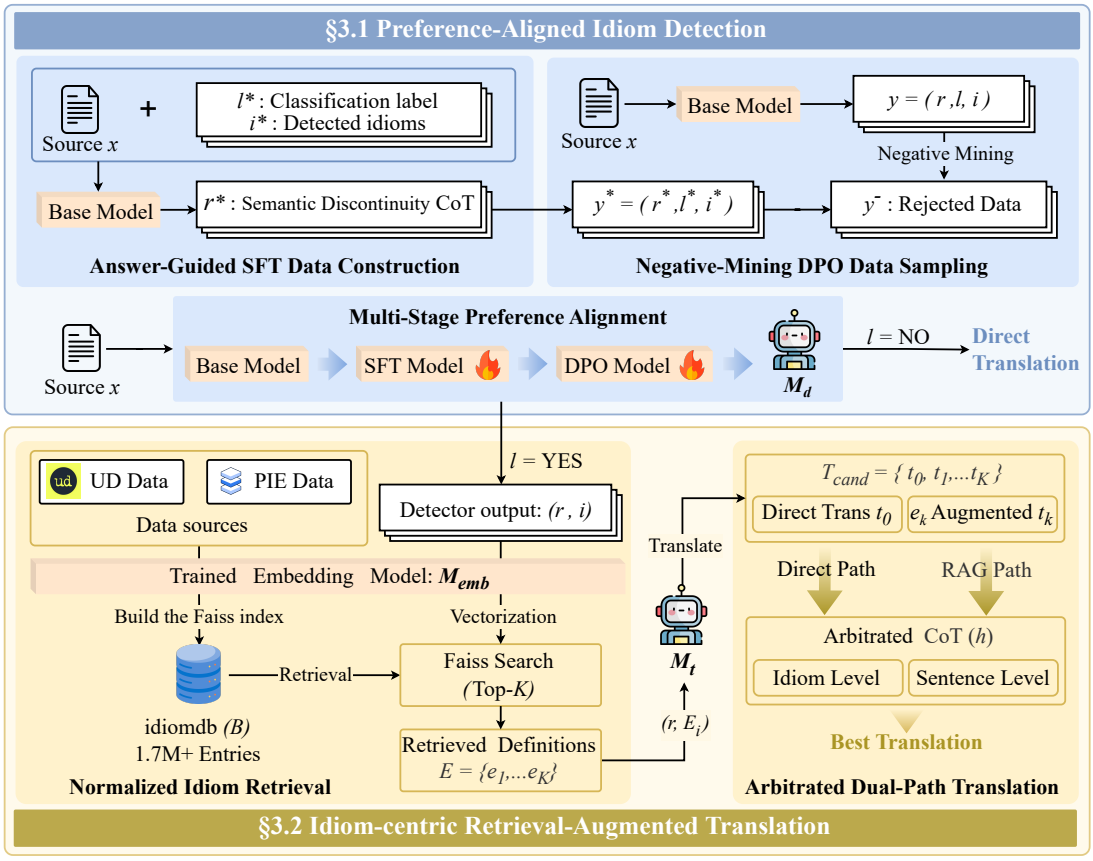

Figure 1: An overview of the proposed framework DEREA.

with precise linguistic evidence. To support this tool, we first construct a large-scale idiom knowledge base (B) by integrating expert-curated items from the PIE Dataset (Zhou et al., 2021) with approximately 2.5 million raw entries from Urban Dictionary[2]. To ensure the quality of retrieval evidence, we employ an LLM-based distiller to filter and transform these raw entries into a structured "one idiom, one definition" format.

Normalized Idiom Retrieval Formally, given the detector output (r, i), we leverage the representation space of M_{emb} to perform a similarity search for i within B. This yields the Top-K definitions $E = \{e_1, e_2, \ldots, e_K\}$, providing targeted external guidance for the translation stage.

However, as the idioms identified by the detection model often exhibit morphological variations and are embedded within contextual fragments, general embedding models frequently fail to map these non-standard surface forms back to their canonical entries in the knowledge base. To address these morphological and contextual discrepancies, we construct an augmented dataset consisting of various idiom variants. Leveraging this dataset, we fine-tune M_{emb} via:

$$\mathcal{L}_{\text{CL}} = -\mathbb{E}_{(v,i)\sim\mathcal{D}_{\text{aug}}}\left[\log \frac{e^{s(v,i)/\tau}}{\sum_{j\in N} e^{s(v,j)/\tau}}\right] \quad (4)$$

where τ is the temperature parameter, N denotes the mini-batch containing one positive and multiple negative samples, and $s(v, i)$ represents the cosine similarity between the embeddings of variant v and idiom i. For high-performance retrieval, B is indexed via a Faiss (Johnson et al., 2019) vector database. This architecture enables efficient mapping of identified idiom candidates to their canonical forms through semantic similarity search, significantly enhancing the precision and efficiency of subsequent translation augmentation.

Arbitrated Dual-Path Translation To address the potential "over-correction" of standard RAG methods, we adopt an arbitrated dual-path strategy.

Given the input source sentence x, the translator first leverages its internal parametric knowledge to produce a literal translation t_0. Subsequently, utilizing E, the translator generates K distinct aug-

[2] https://www.kaggle.com/datasets/therohk/urban-dictionary-words-dataset

Label	Idiom / Sense	Source Sentence	Reference Translation (ZH)
With_Idiom	**until the cows come home**: for an indefinite time	You can keep on trying to convince **till the cows come home**, but I won't change my views.	你可以一直试图说服我直到天荒地老/till the cows come home，但我不会改变我的观点。
Without_Idiom	–	You can keep on trying to convince me **forever**, but I won't change my views.	你可以一直/keep on…forever 试图说服我，但我不会改变我的看法。

Table 2: Representative examples from the LoMI benchmark.

mented translations, where t_k is augmented by its corresponding definition e_k. Formally, the generation of the candidate pool T_{cand} is defined as:

$$t_0 \sim P_{M_t}(\cdot \mid x) \tag{5}$$

$$t_k \sim P_{M_t}(\cdot \mid x, e_k), \quad \forall k \in \{1, \ldots, K\} \tag{6}$$

$$T_{\text{cand}} = \{t_0, t_1, \ldots, t_K\} \tag{7}$$

Based on this pool, M_t conducts an arbitration reasoning analysis. Guided by the reasoning chain r from the detector, the model generates a comprehensive critique h for each candidate in T_{cand}. This critique involves a comparative study of semantic accuracy and sentence fluency.

The final decision process is modeled as follows:

$$h \sim P_{M\text{t}}(\cdot \mid x, r, E, T\text{cand}) \tag{8}$$

$$t^* = Arbiter(M_{\text{t}}, h) \tag{9}$$

where h serves as the reflective CoT that weights the merits of literal interpretation against retrieval-augmented candidates. The optimal translation t^* is selected by the model acting as an Arbiter, which evaluates the candidates based on the reasoning trajectory h. This selection logic ensures that the model resolves semantic deviations identified in t_0. If the RAG-augmented candidates fail to provide a more accurate idiomatic mapping, the model maintains the linguistic naturalness of the baseline t_0. The specific prompt governing this prompting-based arbitration is detailed in Appendix A.2.

4 LoMI Benchmark

The vast pre-training scale of LLMs often leads to benchmark contamination, where memorization-driven scores mask their true generalization capabilities for novel idiomatic expressions. To address this, we introduce **LoMI** (**Lo**w **M**emorization **I**diom Translation), a high-difficulty benchmark designed for rigorous and contamination-resistant evaluation. We utilize the PIE corpus (Zhou et al., 2021) as our data source, which contains 1,197 idioms and 5,170 associated context sentences.

4.1 Dataset Filtering and Selection

To identify truly challenging entries and ensure a low-contamination benchmark, we employ a dual-metric filtering mechanism consisting of a memorization index and a translation quality indicator. The memorization index leverages the lexical rigidity of idioms to assess whether they are stored in a model's parametric memory. Following established methodologies for probing linguistic knowledge (Haviv et al., 2023; Liu et al., 2024), we conduct a token-level completion task to evaluate the model's internal memorization of idiomatic expressions. Specifically, we provide the model with a truncated idiom prefix and assess its ability to predict the canonical final token. Simultaneously, the translation quality indicator evaluates contextual understanding by using Qwen3-8B to generate direct translations, which are then assessed by GPT-5 to verify if the figurative meaning is correctly conveyed. By aligning these metrics, we categorize samples by difficulty, prioritizing those where the model demonstrates both a lack of memorization and a failure to translate correctly. Since our raw source provides standard idiomatic forms, our classification is not affected by surface variations. Following this process, we selected 109 unique idioms across 693 context sentences. For each idiomatic sentence, we also curated a semantically equivalent, idiom-free version for contrastive evaluation.

4.2 Reference Construction

To guarantee high-quality ground truth, we utilize a knowledge-guided generation process. We provide GPT-5 with the source sentence along with the idiom's explicit figurative meaning to guide the creation of reference translations. This ensures the nuances of the idiomatic expressions are accurately preserved in the target languages. Subsequently, we implement a self-correction mechanism where the model scrutinizes the initial outputs to identify and rectify unnatural phrasings. The resulting LoMI benchmark comprises 1,386 high-quality parallel

pairs across Chinese, French, and German. This total is composed of the 693 selected idiomatic sentences and their corresponding idiom-free control sentences, facilitating a robust assessment of idiom-specific translation performance (see Table 2 for representative examples).

5 Experiments

5.1 Setup

Data For detection model training, we utilize a GPT-5-refined subset of the MAGPIE corpus (Haagsma et al., 2020), selecting 35,536 high-quality instances from the original 44,451 entries. For contrastive data construction, we utilize GPT-5 to generate idiom-free instances for each entry. To ensure experimental integrity and prevent data leakage, we curate a training subset of simple idioms from the PIE dataset for the embedding model, strictly ensuring no overlap with the LoMI test set. Regarding the construction of knowledge base B described in Sec 3.2, we use Qwen3-32B[3] as the backbone for distillation and structuring. A manual evaluation of the knowledge base's interpretational error rate is provided in Appendix A.3.

Evaluation We conduct all experiments on the LoMI benchmark. We employ COMET[4] as the primary sentence-level evaluation metric. In addition, we employ Claude Sonnet 4.5[5] as an expert judge to evaluate idiom translation performance across two specific dimensions: (1) Semantic Fidelity (LLM_q): given the idiom's gold meaning, this metric assesses whether the translation accurately captures the figurative semantics of the idiom; (2) Contextual Consistency (LLM_c): evaluates whether the translated expression fits naturally within the target context and maintains coherent discourse flow. Both dimensions are scored on a 10-point scale and scaled by a factor of 10 in our results tables to align with COMET metrics. Detailed prompts are provided in Appendix A.2.

Detection Model We instantiate M_d using Qwen3-8B[6] as the backbone. We evaluate its performance against three detection baselines: (a) XLM-RoBERTa-base[7]: representing traditional encoder-based detection models fine-tuned on the MAGPIE corpus; (b) Qwen3-8B base: prompting the LLM to identify idioms directly without fine-tuning; (c) Qwen3-8B Fine-tuning: the model fine-tuned only for classification without the CoT reasoning path. Detailed results of detection performance are summarized in Table 3.

Method	P (%)	R (%)	F1 (%)
XLM-R fine-tuning	87.3	67.4	76.1
Qwen base	84.3	64.8	73.3
Qwen fine-tuning	**89.7**	73.2	80.6
Qwen SFT (CoT)	84.3	87.6	85.9
Qwen DPO (CoT)	86.9	**87.7**	**87.3**

Table 3: Experimental results of Detection Models. "Qwen" denotes Qwen3-8B.

Translation Model For the retrieval component, we instantiate M_{emb} using Qwen3-Embedding-0.6B[8] and set the retrieval size to $K = 3$. The translator M_t is evaluated across various LLMs including Qwen3-8B, Llama3-8B-Instruct[9], Qwen3-32B, and GPT-5-mini[10]. We compare DEREA against several representative translation methods: (a) Direct Translation: the base LLM directly translates the source sentence without any specific prompting strategy; (b) Sentence-level RAG: a standard RAG approach where the entire source sentence is used as a query to retrieve relevant knowledge; (c) SlangOWL (Liang et al., 2025): a single-model reasoning baseline that prompts the LLM to self-retrieve and analyze potential idioms. Detailed specifications and hyperparameter settings are provided in Appendix A.4.

5.2 Main Results

Results on Detection As shown in Table 3, our detection model achieves significant improvements over all baselines. Notably, the introduction of DPO training effectively boosts Recall (84.3% $\rightarrow$ 86.9%) while maintaining high Precision (87.6% $\rightarrow$ 87.7%), demonstrating that the model becomes more sensitive to idioms while reducing false positives on non-idiomatic sentences.

Results on Translation As shown in table 4, for lightweight models, DEREA yields substantial gains. Our method achieves the highest scores

[3] https://huggingface.co/Qwen/Qwen3-32B
[4] https://github.com/Unbabel/COMET
[5] claude-sonnet-4-5-20250929
[6] https://huggingface.co/Qwen/Qwen3-8B
[7] https://huggingface.co/FacebookAI/xlm-roberta-base
[8] https://huggingface.co/Qwen/Qwen3-Embedding-0.6B
[9] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[10] gpt-5-mini-2025-08-07

System	En ⇒ Zh			En ⇒ De			En ⇒ Fr			All		
	COMET	LLM_q	LLM_c	COMET	LLM_q	LLM_c	COMET	LLM_q	LLM_c	COMET	LLM_q	LLM_c
Llama3-8B-Instruct												
Direct	76.0	41.3	42.1	73.5	43.8	45.8	73.7	46.3	48.2	74.4	43.8	45.4
Sentence-RAG	76.0	44.4	45.9	73.3	45.8	46.8	73.7	48.5	50.3	74.3	46.2	47.7
SLANGOWL	76.1	46.8	48.4	69.9	49.6	50.6	73.8	53.8	56.6	73.3	50.1	51.9
DEREA	**78.8**	**55.0**	**56.9**	**74.0**	**55.0**	**58.1**	**75.2**	**58.9**	**62.5**	**76.0**	**56.3**	**59.2**
Qwen3-8B												
Direct	81.4	52.5	52.7	75.1	45.9	47.3	75.1	47.7	50.2	77.2	48.7	50.1
Sentence-RAG	81.7	52.3	52.9	73.2	41.4	43.3	73.5	42.9	44.9	76.1	45.5	47.0
SLANGOWL	82.8	58.2	58.2	74.4	47.2	49.9	75.2	50.3	52.5	77.5	51.9	53.5
DEREA	**85.0**	**65.8**	**66.6**	**77.3**	**55.0**	**58.1**	**77.8**	**57.3**	**59.5**	**80.1**	**59.4**	**61.4**
Qwen3-32B												
Direct	83.7	60.9	62.7	77.7	54.2	56.6	77.4	57.0	60.1	79.6	57.4	59.8
Sentence-RAG	84.5	64.2	65.6	78.0	55.7	58.9	78.4	58.0	61.7	80.3	59.3	62.1
SLANGOWL	84.8	69.0	69.1	79.4	61.8	65.3	79.9	64.3	67.7	81.4	65.0	67.4
DEREA	**85.1**	**70.7**	**72.6**	**80.1**	**65.2**	**67.5**	**80.3**	**66.8**	**69.5**	**81.8**	**67.6**	**69.9**
GPT5-mini												
Direct	85.2	70.3	70.1	82.1	67.4	67.1	81.6	67.3	67.8	83.0	68.3	68.3
Sentence-RAG	**85.6**	73.3	74.6	82.7	69.1	68.1	**82.9**	71.6	70.4	83.7	71.3	71.0
SLANGOWL	85.1	74.2	71.8	83.0	71.2	70.0	82.7	70.9	70.6	83.6	72.1	70.8
DEREA	**85.6**	**75.9**	**75.1**	**83.5**	**72.8**	**73.5**	82.4	**73.0**	**72.0**	**83.8**	**73.9**	**73.5**

Table 4: Main results for English to various languages translation. LLM_q and LLM_c represent semantic fidelity and contextual consistency scores, respectively. Statistical significance test results are provided in Appendix A.5.

Variant	COMET	LLM_q	LLM_c	Avg
w/o Arbitrate (h)	77.3	53.1	55.8	62.1
w/o Direct (t_0)	78.4	55.7	58.3	64.1
w/o Detector CoT (r)	79.3	56.9	59.4	65.2
Top1-retrieval	79.7	57.6	60.4	65.9
DEREA	**80.1**	**59.4**	**61.4**	**67.0**

Table 5: Ablation results on the LoMI benchmark using Qwen3-8B as the backbone.

across all three language pairs and evaluation metrics. Remarkably, the DEREA-enhanced 8B model matches or even surpasses the performance of the direct translation from the much larger 32B model. For large-scale models (Qwen3-32B, GPT-5-mini), while the improvement in COMET scores is relatively modest due to their inherently strong baseline translation capabilities, DEREA demonstrates significant advantages in LLM-based evaluations. Specifically, when applied to GPT-5-mini, our framework achieves scores of 73.9 in idiomatic semantic quality and 73.5 in semantic consistency, outperforming the SlangOWL which scores 72.1 and 70.8, respectively. Beyond translation performance, the efficiency of the idiom detection and retrieval modules is critical for real-world applications. We conduct a detailed efficiency analysis in Appendix A.6.

5.3 Analysis

Ablation and Strategy Analysis To verify the necessity of our collaborative reasoning and multi-candidate selection mechanisms, we conduct experiments on LoMI across four ablation variants (see Table 5). (1) w/o Arbitrate (h), which removes all selection processes and provides detected idioms directly in a simple prompt; (2) w/o Direct (t_0), forcing the model to rely solely on RAG by excluding literal translations from the candidate pool; (3) w/o Reasoning (r), which omits CoT during detection. (4) Top-1 Retrieval (K=1). Experimental results reveal that omitting the arbitration mechanism (h) incurs the most significant performance decline across all metrics. This underscores the pivotal role of h in path balancing—filtering out irrelevant retrieved noise while preserving the model's intrinsic linguistic capabilities. The absence of either the direct translation path (t_0) or the reasoning chain (r) also consistently degrades performance, demonstrating that "semantic conflict analysis" is essential for accurate idiom identification and that forced reliance on RAG can lead to knowledge interference. Finally, the Top-1 retrieval variant proves insufficient, as multiple definitions are often required to cover the diverse morphological and contextual nuances of real-world idioms.

Backbone	Method	Figurative	Literal	Avg
Llama3-8B -Instruct	Direct	74.4	**85.0**	79.7
	sRAG	74.3	83.4	78.9
	SLANGOWL	73.3	82.5	77.9
	DEREA	**76.0**	84.5	**80.4**
Qwen3-8B	Direct	77.2	**87.6**	82.4
	sRAG	76.1	86.6	81.4
	SLANGOWL	77.5	87.4	82.5
	DEREA	**80.1**	87.5	**83.8**

Table 6: Robustness evaluation on full LoMI. All results are averaged across three languages.

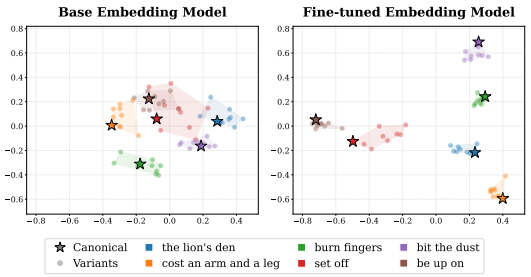

Figure 2: Semantic space alignment of idiom canonical forms and contextual variants: PCA-based visualization of Base vs. Fine-tuned Qwen3-Embedding-0.6B.

Robustness Analysis on Idiom-free Samples To assess resilience against retrieval noise, we evaluate DEREA on LoMI's idiom-free subsets (Table 6). While baselines like SlangOWL exhibit significant performance degradation and model-dependent instability, DEREA maintains performance nearly identical to direct translation (e.g., 84.5 vs. 85.0 on Llama3-8B). Specifically, the dual-path arbitration mechanism acts as a critical safety net, effectively shielding the final output from "over-correction" by favoring the more fluent literal path when retrieved definitions are contextually irrelevant. These results demonstrate that our framework avoids over-correction and ensures consistent stability across various backbones.

Balancing Retrieval Accuracy and Efficiency We evaluate embedding models using Recall@K and time (s) to balance accuracy and latency. As shown in Table 7, our fine-tuned Qwen3-0.6B significantly outperforms both fuzzy matching and the base model. This gain is visually confirmed in Figure 2, where fine-tuning effectively clusters contextual variants around canonical forms, whereas the base model exhibits a scattered distribution. Notably, retrieval gains plateau at $K = 3$, with $K = 5$ offering only marginal improvement. Furthermore, while the Qwen3-8B base model shows competitive accuracy, its superior performance is offset by increased inference latency, justifying our choice of a fine-tuned lightweight retriever.

Retrieval Method	R@1	R@3	R@5	time (s)
Fuzzy Matching	54.8	65.0	66.0	**0.23**
Qwen3-0.6B (Base)	62.7	72.0	72.0	0.27
Qwen3-8B (Base)	**79.6**	87.3	89.8	1.13
Qwen3-0.6B (ft)	78.1	**90.1**	**91.0**	0.27

Table 7: Retrieval accuracy of different embedding methods. "ft" denotes our fine-tuned version. "R@K" denotes Recall@K score.

Backbone	Method	COMET	LLM_q	LLM_c	Avg
Llama3-8B -Instruct	Base	66.6	32.8	33.8	44.4
	SLANGOWL	59.6	32.6	34.6	42.3
	sRAG	64.6	33.4	35.6	44.5
	DEREA	**67.8**	**37.7**	**38.0**	**47.8**
Qwen3-8B	Base	73.8	44.2	46.4	54.8
	SLANGOWL	73.8	44.1	46.3	54.7
	sRAG	73.9	45.7	48.4	56.0
	DEREA	**75.3**	**50.7**	**53.1**	**59.7**
Qwen3-32B	Base	76.5	44.7	48.7	56.6
	SLANGOWL	76.3	48.4	51.9	58.9
	sRAG	76.3	51.1	53.9	60.4
	DEREA	**77.1**	**53.5**	**54.7**	**61.8**
GPT5-mini	Base	75.8	53.0	56.4	61.7
	SLANGOWL	75.9	54.7	57.8	62.8
	sRAG	76.5	56.1	57.6	63.4
	DEREA	**78.5**	**59.6**	**61.0**	**66.4**

Table 8: Translation performance on the Emerging Slang dataset.

Superiority in Emerging Slang Translation To verify that DEREA is not constrained by fixed model parameters and can adapt to evolving linguistic trends, we evaluate its performance on newly emerged slang. Following the methodology of Mei et al. (2024), we curated a specialized benchmark of 200 emerging idiomatic expressions uploaded to Urban Dictionary after June 2025. This ensures that the data is absent from the training corpora of current LLMs and presents a significant translation challenge. Example sentences from the dictionary served as English sources and high-quality Chinese references were constructed by providing GPT-5 with standard slang definitions, followed by rigorous manual review and expert refinement.

Results on the emerging slang dataset in Table 8 reveal that performance for all baseline models drops significantly. Notably, deep reasoning methods yield negligible improvements as they are limited by the model's internal parametric knowledge.

System	LoMI (Fig.)	LoMI (Lit.)	WMT24
Direct	77.2	87.6	84.6
sRAG	76.1	86.6	83.9
SlangOWL	77.5	87.4	84.3
DEREA	80.1	87.5	84.4

Table 9: Robustness evaluation (COMET scores) on idiomatic and standard datasets.

Source	Experts (Avg.)	Grok-4	Mean
WMT24 Ref.	8.73	8.30	8.63
LoMI Ref. (Ours)	8.87	9.10	**8.93**

Table 10: Blind quality evaluation of reference translations (1–10 scale).

System	Idiomaticity	Adequacy
SlangOWL	6.75	6.87
DEREA	**7.18**	**7.36**

Table 11: Expert scores for model outputs, confirming alignment with automated metrics.

Method	Metric	Claude 4.5	Grok-4	GPT-5	Std. (σ)
SlangOWL	LLM_q	6.18	5.86	5.88	0.18
	LLM_c	6.42	7.51	6.90	0.55
DEREA	LLM_q	**6.99**	**6.58**	**6.79**	0.20
	LLM_c	**7.19**	**8.10**	**7.41**	0.47

Table 12: Multi-judge robustness and stability analysis across 200 samples.

Sentence-level RAG shows marginal gains but is hindered by retrieval noise. In contrast, DEREA provides substantial and stable improvements, even for GPT-5-mini. This proves that our framework effectively bypasses the knowledge cut-off of LLMs by seamlessly integrating an updatable knowledge base. Appendix A.7 presents a detailed case study of DEREA in challenging scenarios.

Robustness on Standard Datasets To evaluate the generalization of DEREA across diverse domains, we conduct additional experiments on the WMT24[11] English-to-Chinese test set (n = 997). As shown in Table 9, since WMT24 is a standard multi-domain dataset with few idiomatic expressions, specialized models generally exhibit a slight performance decrease compared to the direct translation baseline. However, DEREA only experiences a negligible decline of 0.2 COMET points, maintaining competitive performance in standard translation tasks. This consistency with our results on the idiom-free subset of LoMI (Section 5.3) further demonstrates the robustness of our approach.

Validation of Data and Evaluation Reliability To verify the quality of LLM-generated references and automated metrics, we conduct a two-fold validation. First, we assess reference fidelity by comparing 100 sampled LoMI references against 100 WMT24 gold standards. Three bilingual experts and one LLM judge (Grok-4[12]) scored these on a scale of 1–10. The specific evaluation prompts used for the LLM judge are detailed in Appendix A.2. As shown in Table 10, LoMI references achieved a mean score of 8.93, exceeding the WMT24 benchmark and confirming professional standards. Second, we conduct a blind human evaluation on 200 model outputs. Experts scored DEREA and the strongest baseline, SlangOWL, on idiomaticity and adequacy. Table 11 shows DEREA maintains a clear lead. Crucially, these expert scores align with the trends of our automated LLM-based metrics.

Stability Analysis Across Multiple LLM Judges To assess the robustness of our automated evaluation, we re-evaluate 200 samples using Claude 4.5 Sonnet, Grok-4, and GPT-5 as independent judges. Table 12 shows that DEREA consistently outperforms SlangOWL across all model families (Anthropic, xAI, OpenAI). The low variance in LLM_q ($\sigma \leq 0.20$) indicates strong consensus on idiom-specific quality. The relatively higher variance in LLM_c reflects inherent differences in how models calibrate full-sentence metrics like fluency or tone, yet the performance trend remains stable across all judges, confirming the reliability of our multi-model evaluation framework.

6 Conclusion

In this paper, we propose DEREA, a collaborative reasoning framework that effectively enhances idiom translation across various model scales. By decoupling detection, retrieval, and arbitration, our approach surmounts the parameter bottlenecks of single-model and mitigates retrieval noise in standard RAG. Experimental results on our newly introduced LoMI and Emerging Slang benchmarks demonstrate that DEREA significantly improves translation accuracy and contextual consistency. Our framework offers a robust, scalable paradigm for handling evolving linguistic phenomena, ensuring high translation fidelity for novel expressions.

[11] https://www2.statmt.org/wmt24

[12] grok-4.1-2025-11-17

Limitations

Despite its effectiveness, our work has several limitations: First, this study focuses on English as the source language. Our current work is limited to the construction of english idiom knowledge bases; however, the framework is inherently language-agnostic and can be generalized to other languages by constructing language-specific idiom knowledge bases. Second, the multi-stage pipeline increases computational overhead. Multiple model calls and the processing of additional tokens result in higher inference latency and operational costs compared to direct translation. Third, reliance on evaluation metrics. Although we employ LLM-based metrics and COMET to assess translation quality, these automated metrics may not fully capture the profound cultural nuances of idiomatic expressions.

Acknowledgments

This work was supported in part by Guangdong S&T Program (Grant No. 2024B0101050003), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2024A1515011491), and Shenzhen Science and Technology Program (Grant Nos. ZDSYS20230626091203008, KJZD20231023094700001, KQTD20240729102154066). We would like to thank the anonymous reviewers and meta-reviewer for their insightful suggestions.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jinheon Baek, Soyeong Jeong, Minki Kang, Jong Park, and Sung Hwang. 2023. Knowledge-augmented language model verification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1720–1736, Singapore. Association for Computational Linguistics.

Antonio Castaldo and Johanna Monti. 2024. Prompting large language models for idiomatic translation. In *Proceedings of the 1st Workshop on Creative-text Translation and Technology*, pages 32–39, Sheffield, United Kingdom. European Association for Machine Translation.

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, and 1 others. 2025. Main-rag: Multi-agent filtering retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2607–2622.

Guanhua Chen, Yutong Yao, Lidia S. Chao, Xuebo Liu, and Derek F. Wong. 2025. SGIC: A self-guided iterative calibration framework for RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28357–28370, Vienna, Austria. Association for Computational Linguistics.

Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. Can transformer be too compositional? analysing idiom processing in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

Sundesh Donthi, Maximilian Spencer, Om B. Patel, Joon Young Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2025. Improving LLM abilities in idiomatic translation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 175–181, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *ArXiv preprint*, abs/2312.10997.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Adi Haviv, Ido Cohen, Jacob Gidron, Roei Schuster, Yoav Goldberg, and Mor Geva. 2023. Understanding transformer memorization recall through idioms. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 248–264, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih,

Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Bryan Li, Jiaming Luo, Eleftheria Briakou, and Colin Cherry. 2025. Leveraging domain knowledge at inference time for LLM translation: Retrieval versus generation. In *Proceedings of the 4th International Workshop on Knowledge-Augmented Methods for Natural Language Processing*, pages 91–106, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024a. Translate meanings, not just words: Idiomkb's role in optimizing idiomatic translation with language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18554–18563. AAAI Press.

Xingxuan Li, Ruochen Zhao, Yew Ken Chia, Bosheng Ding, Shafiq Joty, Soujanya Poria, and Lidong Bing. 2024b. Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yunlong Liang, Fandong Meng, Jiaan Wang, and Jie Zhou. 2025. Slangdit: Benchmarking llms in interpretative slang translation. *ArXiv preprint*, abs/2505.14181.

Dekang Lin. 1999. Automatic identification of non-compositional phrases. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 317–324, College Park, Maryland, USA. Association for Computational Linguistics.

Emmy Liu, Aditi Chaudhary, and Graham Neubig. 2023a. Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15095–15111, Singapore. Association for Computational Linguistics.

Liangxin Liu, Xuebo Liu, Derek F. Wong, Dongfang Li, Ziyi Wang, Baotian Hu, and Min Zhang. 2024. Selectit: Selective instruction tuning for llms via uncertainty-aware self-reflection. In *Advances in Neural Information Processing Systems*, volume 37, pages 97800–97825. Curran Associates, Inc.

Xuebo Liu, Yutong Wang, Derek F. Wong, Runzhe Zhan, Liangxuan Yu, and Min Zhang. 2023b. Revisiting commonsense reasoning in machine translation: Training, evaluation and challenge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15536–15550, Toronto, Canada. Association for Computational Linguistics.

Lingrui Mei, Shenghua Liu, Yiwei Wang, Baolong Bi, and Xueqi Cheng. 2024. SLANG: New concept comprehension of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12558–12575, Miami, Florida, USA. Association for Computational Linguistics.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2025. Rolling the DICE on idiomaticity: How LLMs fail to grasp context. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7314–7332, Vienna, Austria. Association for Computational Linguistics.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Jianhui Pang, Fanghua Ye, Derek Fai Wong, Dian Yu, Shuming Shi, Zhaopeng Tu, and Longyue Wang. 2025. Salute the classic: Revisiting challenges of machine translation in the age of large language models. *Transactions of the Association for Computational Linguistics*, 13:73–95.

Jun Rao, Zepeng Lin, Xuebo Liu, Xiaopeng Ke, Lian Lian, Dong Jin, Shengjun Cheng, Jun Yu, and Min Zhang. 2025. APT: Improving specialist LLM performance with weakness case acquisition and iterative preference training. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20958–20980, Vienna, Austria. Association for Computational Linguistics.

Yeh-Zu Tzou, Jyotsna Vaid, and Hsin-Chin Chen. 2017. Does formal training in translation/interpreting affect translation strategy? evidence from idiom translation. *Bilingualism: Language and cognition*, 20(3):632–641.

Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. 2025a. Drt: Deep reasoning translation via long chain-of-thought. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6770–6782.

Yutong Wang, Siyuan Xiong, Xuebo Liu, Wenkang Zhou, Liang Ding, Miao Zhang, and Min Zhang. 2026. Agentdropoutv2: Optimizing information flow in multi-agent systems via test-time rectify-or-reject pruning. *Preprint*, arXiv:2602.23258.

Yutong Wang, Jiali Zeng, Xuebo Liu, Fandong Meng, Jie Zhou, and Min Zhang. 2024a. TasTe: Teaching large language models to translate through self-reflection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6144–6158, Bangkok, Thailand. Association for Computational Linguistics.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2025b. DelTA: An online document-level translation agent based on multi-level memory. In *The Thirteenth International Conference on Learning Representations*.

Zeqiang Wang, Jiageng Wu, Yuqi Wang, Wei Wang, Jie Yang, Jon Johnson, Nishanth Sastry, and Suparna De. 2024b. Revealing COVID-19's social dynamics: Diachronic semantic analysis of vaccine and symptom discourse on Twitter. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3383–3394, Miami, Florida, USA. Association for Computational Linguistics.

Zhexuan Wang, Yutong Wang, Xuebo Liu, Liang Ding, Miao Zhang, Jie Liu, and Min Zhang. 2025c. AgentDropout: Dynamic agent elimination for token-efficient and high-performance LLM-based multi-agent collaboration. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24013–24035, Vienna, Austria. Association for Computational Linguistics.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.

Ifeoluwa Wuraola, Nina Dethlefs, and Daniel Marciniak. 2024. Understanding slang with LLMs: Modelling cross-cultural nuances through paraphrasing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15525–15531, Miami, Florida, USA. Association for Computational Linguistics.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *ArXiv preprint*, abs/2401.11817.

Chengyuan Yao and Satoshi Fujita. 2024. Adaptive control of retrieval-augmented generation for large language models through reflective tags. *Electronics*, 13(23):4643.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xiong-Hui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. 2025. AFlow: Automating agentic workflow generation. In *The Thirteenth International Conference on Learning Representations*.

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*, abs/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*, abs/2411.14405.

Jonathan Zheng, Alan Ritter, and Wei Xu. 2024. Neo-bench: Evaluating robustness of large language models with neologisms. *ArXiv preprint*, abs/2402.12261.

Jianing Zhou, Hongyu Gong, and Suma Bhat. 2021. PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing. In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 33–48, Online. Association for Computational Linguistics.

Mingchen Zhuge, Wenyi Wang, Louis Kirsch, Francesco Faccio, Dmitrii Khizbullin, and Jürgen Schmidhuber. 2024. GPTSwarm: Language agents as optimizable graphs. In *Forty-first International Conference on Machine Learning*.

A Appendix

A.1 Algorithm: DEREA Inference Pipeline

This section provides the formal algorithmic description of the DEREA inference pipeline (see Algorithm 1), detailing the sequential execution of reasoning-based detection, canonical knowledge retrieval, and arbitrated translation.

A.2 Prompt Templates

In this section, we detail the prompt templates utilized for the Detector and Translator modules, as well as the LLM-as-a-judge evaluation component within the DEREA framework.

Detector Prompts The Detecor is tasked with identifying idiomatic expressions by analyzing semantic incongruity between literal and contextual meanings.

System Prompt

** Core Method
You will act as a rigorous language analyst. Your core methodology is to objectively determine whether an expression is an "Unconventional Expression" by logically comparing its "literal meaning" with its "contextually inferred meaning".
** Key Definition: Unconventional Expression: An "Unconventional Expression" refers to a word or phrase whose overall meaning cannot be directly inferred from the literal meanings of its individual components. Such expressions are typically figurative, idiomatic, or culturally specific, and are highly dependent on context. They mainly include idioms, slang ...
** Specific Requirements

1. Unconventional expressions in a sentence may involve variations in person, tense, abbreviations, etc. Please identify and normalize them to their canonical forms.
2. Conventional usages of unconventional expressions should also be taken into consideration.
3. When both literal and figurative meanings are explainable, the final judgment should be made by selecting the interpretation that best fits the given context.

User Prompt

** Example sentence containing an unconventional expression:
"When the share market crashed his fingers were burnt from all the investments that he had made."
Example Output:
***Thought process:
"his fingers were burnt": The literal meaning is "his fingers were physically burned", but in the given context—"when the share market crashed, he suffered losses due to his investments"—"burn fingers" is used to indicate that the subject incurred financial losses as a result of poor investment decisions. The literal interpretation clearly does not fit the context, so it is identified as an unconventional expression.
Other expressions such as "the share market crashed" and "the investments" align with their literal meanings.
***Output: ## YES ## 1.burn fingers
Example sentence without an unconventional expression:
"She is reading a book."
Example Output:
***Thought process:
The sentence has a clear structure, and expressions such as "reading a book" are common phrases with no hidden meanings. All expressions in the sentence are used in their literal sense, and there is no unconventional expression present.
***Output: ## NO
Please follow the system-provided methodology and refer to the examples above to process the unconventional expressions in the following English sentence.
English text: {sentence}
Thought process & Output
Please follow the output format below:
***Thought process:
***Output: ## YES ## <Unconventional Expression>
or
***Thought process:
***Output: ## NO

Literal Translation The "literal translation" serves as a zero-shot baseline, generated using the following prompt:

Algorithm 1 DEREA Inference Pipeline

Require: Source sentence x; Detection model M_d; Retrieval model M_{emb}; Translation model M_{t}; Knowledge Base B.

Ensure: Final Translation t^*.

Phase 1: Preference-Aligned Idiom Detection

$(r, l, i) \leftarrow M_d(x)$ ▷ Generate reasoning r, label l, and spans i

if $l =$ NO **then**

  $t_0 \sim P_{M_{\mathrm{t}}}(\cdot \mid x)$ **return** t_0 ▷ Yield direct translation if no idiom detected

end if

Phase 2: Idiom-centric Retrieval-Augmented Translation

// i. Normalized Idiom Retrieval

$E \leftarrow \text{FaissSearch}(M_{emb}(i), B, K)$ ▷ Retrieve Top-K definitions $E = \{e_1, \ldots, e_K\}$

// ii. Arbitrated Dual-Path Translation

$t_0 \sim P_{M_{\mathrm{t}}}(\cdot \mid x)$ ▷ Generate literal baseline t_0

$T_{\text{cand}} \leftarrow \{t_0\}$

for each $e_k \in E$ **do**

  $t_k \sim P_{M_{\mathrm{t}}}(\cdot \mid x, e_k)$ ▷ Generate RAG-augmented candidates

  $T_{\text{cand}} \leftarrow T_{\text{cand}} \cup \{t_k\}$

end for

$h \sim P_{M_{\mathrm{t}}}(\cdot \mid x, r, E, T_{\text{cand}})$ ▷ Generate reflective critique h based on r

$t^* = \arg\max_{t \in T_{\text{cand}}} P_{M_{\mathrm{t}}}(\text{Select}(t) \mid h, T_{\text{cand}})$ ▷ Select optimal translation

return t^*

Literal Translation

Translate the following English sentence into `{target_lang}`. English: `{sentence}` `{target_lang}`:

SlangOWL Prompt This refers to a training-free prompting scheme. To ensure a fair comparison, the backbone model used for SlangOWL is the same as the one used in our proposed method (as specified in the main results table). The specific prompt used for SlangOWL:

SlangOWL Prompt

'As a translation expert, please translate the sentence I provide to you into `{lang}`: The sentence may contain unconventional expressions. Please first detect whether the sentence contains any unconventional expressions. If they exist, provide an explanation of the unconventional expression before translating the sentence; if they do not exist, translate the sentence directly.

Translator Prompts The Translator integrates retrieval results and detection reasoning to produce the final translation.

System Prompt

You are a translation assistant whose task is to analyze translation background information and ultimately provide the optimal Chinese translation.
You will be given the following translation background information:

1. The original English sentence.
2. The literal translation result from the model.
3. The analysis method and detection result for unconventional expressions within the sentence.
4. Dictionary entries retrieved from RAG (Retrieval-Augmented Generation) (Unconventional expressions: their figurative meanings).

User Prompt

Input Information:

1. Original English Sentence: {en}
2. Analysis Method and Detection Result for Unconventional Expressions in the Sentence: {think}

3. Retrieved Dictionary Entries: {rag_info}
4. Translation 0: Model's Literal Translation Result: {direct}

Output Format: Based on the retrieved dictionary entries, generate three separate translations: Translation 1, Translation 2, and Translation 3.
Systematically analyze the translation quality of unconventional expressions and the overall sentence quality for Translation 0, Translation 1, Translation 2, and Translation 3, and finally determine which translation will be selected as the final output:
Output Example:
*** RAG Translation Results ***

1. Translation 1: Based on dictionary entry '...', translated as: "..."
2. Translation 2: Based on dictionary entry '...', translated as: "..."
3. Translation 3: Based on dictionary entry '...', translated as: "..."

*** Translation Quality Analysis ***

1. Translation 0: Result: ... Analysis of unconventional expression quality: ... Analysis of overall sentence quality: ...
2. Translation 1: Result: ... Analysis of unconventional expression quality: ... Analysis of overall sentence quality: ...
3. Translation 2: Result: ... Analysis of unconventional expression quality: ... Analysis of overall sentence quality: ...
4. Translation 3: Result: ... Analysis of unconventional expression quality: ... Analysis of overall sentence quality: ...

*** Final Translation ***
(Output only the final translation result in this section; no other information is permitted)

LLM As Judge Prompts This section presents the prompt templates used to guide LLM-based metrics in evaluating translation quality and contextual consistency.

Translation Quality

/* Task prompt */ Please evaluate the translation quality of specific expressions from English sentences into German sentences.
/* Evaluation Criteria */ /** Translation Quality Score **/ Based on the common meaning of the given expression, judge whether the translation conveys the actual intended meaning, tone, and nuance.
Score from 1 to 10, applying a strict and wide distribution:

1. **Low Range (1-4):** For translations that are wrong, omitted, or overly literal (word-for-word) causing confusion.
2. **Mid Range (5-7):** For translations that convey the basic meaning but lack idiomatic flavor or refinement.
3. **High Range (8-10):** For translations that are not only accurate but also culturally adapted and elegant.

Important: Do not hesitate to give extreme scores (1 or 10) to truly differentiate translation quality.
/* Test Data */ Evaluate the following translation:

1. English sentence: {source}
2. Idiom in the sentence: {idiom}
3. Idiom Sense: {sense}
4. {lang} translation: {translation}

Evaluation (Output only the integer score between 1 and 10; do not output any other information):

Contextual Consistency

/* Task prompt */ Please evaluate the translation quality of specific expressions from English sentences into German sentences.
/* Evaluation Criteria */ /** Contextual Consistency Score **/ Judge whether the translated expression is consistent with the meaning in the context and whether it integrates smoothly and logically into the sentence.
Score from 1 to 10, applying a strict and wide distribution:

1. **Low Range (1-4):** The expression does not fit the context logically, breaks the sentence flow, or feels forcefully inserted (awkward or jarring).
2. **Mid Range (5-7):** The sentence is generally fluent and conveys the meaning, but the integration of the expression

feels slightly stiff or unnatural.
3. **High Range (8-10):** The context is completely natural and coherent; the expression is embedded seamlessly and reads like a native sentence without any forced integration.

Important: Do not hesitate to give extreme scores (1 or 10) to truly differentiate how well the expression fits the context.
/* Test Data */ Evaluate the following translation:
1. English sentence: {source}
2. Idiom in the sentence: {idiom}
3. Idiom Sense: {sense}
4. {lang} translation: {translation}

Evaluation (Output only the integer score between 1 and 10; do not output any other information):

LLM Judge Evaluation Prompt We utilize Grok-4 as an automated judge to assess the quality of idiom translations. Below is the detailed system prompt and instruction template used for the reference fidelity evaluation:

Evaluation

System Prompt: You are an expert linguistic evaluator specializing in translation quality assessment between English and Chinese.
Instruction: Please evaluate the quality of the following idiom translation based on three criteria:
Accuracy: Does it capture the exact figurative meaning of the source idiom?
Fluency: Is the target language natural and idiomatic?
Contextual Fit: Does the translation suit the provided sentence context?
Source Sentence: sentence Target Idiom: idiom Candidate Translation: translation
Provide a score between 1 and 10, where 10 indicates a professional-grade translation.
Return your response in the format: `Score: , Reasoning: .`

A.3 Knowledge Base Quality Assurance

To ensure the semantic integrity and translation utility of the distilled knowledge base B, we conducted a rigorous manual inspection. We employed a random sampling approach, selecting 100 idiom-definition pairs from the processed repository to cover both expert-curated items from the PIE dataset and noise-heavy entries from Urban Dictionary. Two independent reviewers with near-native English proficiency evaluated these entries based on three criteria: wrong-sense errors (literal or incorrect figurative meanings), incomplete definitions (missing critical nuances), and hallucinations (AI-generated artifacts). This process verifies whether the distillation pipeline effectively filters out the conversational noise and personal anecdotes inherent in raw social media data while preserving the core semantic interpretations required for accurate translation.

The manual evaluation revealed a high level of reliability for the Qwen3-32B distillation pipeline, with zero instances of wrong-sense errors or hallucinations identified in the sampled set. This confirms that the model successfully captured the precise linguistic essence of the idiomatic expressions. We observed a marginal 2% rate of incomplete definitions, primarily occurring in highly polysemous slang where the distillation prioritized the most prevalent contemporary sense over secondary meanings. Overall, the qualitative analysis shows that the pipeline effectively transforms informal, context-dependent raw snippets—such as those often found in Urban Dictionary—into concise, dictionary-style definitions. These structured entries provide high-quality linguistic evidence that supports the downstream translation model M_t in resolving idiomatic ambiguities.

A.4 Models & Hyperparameters

Detector We utilize Qwen3-8B as the backbone for the detection model. In the SFT stage, we fine-tune the model for 3 epochs using DeepSpeed ZeRO-Stage 3 optimization, with a batch size of 16 and a learning rate of $5.0\mathrm{e}{-6}$. In the DPO stage, we apply Low-Rank Adaptation (LoRA) with a rank of 8 and $\beta = 0.1$. The model is trained for 300 steps with a batch size of 32 and a learning rate of $5.0\mathrm{e}{-6}$. During the inference phase for generating SFT and DPO data, we set the temperature to 0.7 and top-p to 0.95 to encourage diversity. For the final testing phase, we use temperature = 0 to ensure deterministic and stable detection results.

Embedding Model The embedding model M_{emb} is initialized with Qwen3-0.6B and fine-tuned using the *Sentence Transformers* framework. To optimize the representation space for idiom variants,

we employ MultipleNegativesRankingLoss for contrastive learning. The model is trained for 5 epochs with a batch size of 256 and a learning rate of $3.0\text{e}{-5}$, incorporating a warmup ratio of 0.1. To ensure hardware efficiency and training stability, we utilize bf16 precision and a non-duplicate batch sampling strategy to enhance the quality of in-batch negatives.

Translator We evaluate it in a training-free manner using four state-of-the-art LLMs: Qwen3-8B, Qwen3-32B, Llama-3-8B-Instruct, and GPT-5-mini. For the three open-source models, we set the inference temperature to 0.7 and top-p to 0.95. For GPT-5-mini, the temperature is set to 0.7.

A.5 Statistical Significance Testing

To verify the statistical reliability of the performance improvements, we perform Paired Bootstrap Resampling with $M = 10,000$ iterations. We compare DEREA against the SlangOWL baseline using Qwen3-8B as the translation backbone across the full test set ($n = 693$).

As summarized in Table 13, the results demonstrate that DEREA's gains are highly significant. The frequency-based p-value ($p < 0.0001$) indicates that our framework outperformed the baseline in every resampling trial. Furthermore, the precise analytical p-values for COMET and our proposed LLM metrics (LLM_q, LLM_c) are all below 10^{-12}. This provides definitive empirical evidence that the observed improvements are robust across different test items and are not the result of random variance or specific data distributions.

Metric	Delta (Δ)	p-value (Freq.)	p-value (Precise)
LLM_q	+0.7864	< 0.0001	4.62×10^{-20}
LLM_c	+0.8716	< 0.0001	4.66×10^{-19}
COMET	+0.0217	< 0.0001	3.16×10^{-12}

Table 13: Statistical significance results using Paired Bootstrap Resampling ($M = 10,000$).

A.6 Inference Latency Analysis

To evaluate the computational efficiency of our multi-stage pipeline, we measure the end-to-end inference latency of DEREA and various baselines. All benchmarks are conducted on a single NVIDIA A800 GPU using the Qwen3-8B model scale. As detailed in Table 14, DEREA's latency (38.96 ms/item) is higher than that of single-pass models due to the sequential execution of detection, retrieval, and arbitration.

System	Latency (ms/item)
Base (Direct)	4.33
SlangOWL	16.59
Sentence-level RAG	20.20
DEREA	38.96

Table 14: Average inference latency benchmarked on a single NVIDIA A800 GPU.

We contend that this overhead is a necessary trade-off for the gains in idiomatic fidelity and translation adequacy. The multi-stage design ensures that idioms are explicitly identified and grounded in retrieved knowledge rather than being processed by simple pattern matching. Future optimization efforts will focus on streamlining the "translate-and-select" process within the arbitration phase to reduce token consumption and time latency without sacrificing performance.

A.7 Case Study

To qualitatively evaluate the performance of DEREA, we present several translation examples in Table 15 and Table 16, comparing our framework against various baselines including direct translation, SlangOWL, and srag. These cases involve highly non-compositional neologisms and slang—such as "Astrojumbotron" and "Foolaxy"—which pose significant challenges for traditional translation models.

C.1	**sentence**: bro pulled a full astrojumbotron last night at Coldplay **idiom**: Astrojumbotron **sense**: When your secret corporate affair gets accidentally exposed on a stadium jumbotron, typically during a Coldplay concert, turning your private mess into public entertainment. **reference**: 昨晚那哥们在 Coldplay 演唱会上当众把见不得人的办公室婚外情全暴露了个底儿掉 / completely exposed the secret affair ***Translations generated based on the above context:*** **###direct**: 哥们儿昨晚在Coldplay的演唱会上彻底玩了个“astrojumbotron” / literal transliteration **###slangOWL**: 哥们昨晚在Coldplay演唱会上彻底闹出了一场像Astroworld那样的大混乱 / chaos like Astroworld (hallucination) **###srag**: 哥们儿昨晚在Coldplay演唱会上彻底上演了个“Astrojumbotron”，把他的秘密办公室恋情意外地在场馆大屏上曝光了 / secret affair exposed on screen **###DeRea**: 哥们儿昨晚在Coldplay的演唱会上直接被场馆大屏当众把秘密恋情曝光了 / secret affair publicly exposed on screen
C.2	**sentence**: Stop being a Butternut Chud—you literally just repeated everything I said. **idiom**: Butternut Chud **sense**: A person who mindlessly repeats others' words or ideas without independent understanding or reflection; a humorous or mildly insulting term for unthinking imitation. **reference**: 别当复读机了 / stop being a repeater, 你刚才就是把我说的话一字不差又念了一遍。 ***Translations generated based on the above context:*** **###direct**: 别当个“Butternut Chud” / literal transliteration ——你简直就把我说的每句话都重复了一遍。 **###SlangOWL**: 别当个“Butternut Chud” / literal transliteration ——你明明只是把我说的话全都重复了一遍。 **###srag**: 别当个“Butternut Chud” / literal transliteration ——你简直就把我说的每句话都重复了一遍。 **###cot**: 别当个复读机 / stop being a repeater ——你字面上把我说的所有话都重复了一遍。
C.3	**sentence**: Scrolling for hours just gives me cheap dopamine, but finishing a tough workout gives me that unmatched feeling of expensive dopamine. **idiom**: Expensive dopamine **sense**: Satisfaction derived from effortful activities involving delayed gratification, leading to long-term fulfillment and purpose. **reference**: 刷手机刷上好几个小时只会让我得到廉价的即时快感，但咬牙完成一组高强度训练却能带来那种无可比拟、发自内心的深层满足感 / deep satisfaction from within. ***Translations generated based on the above context:*** **###direct**: 长时间刷手机只是带来廉价的多巴胺，而完成一次艰苦的锻炼才会让我感受到那种无与伦比的“昂贵”多巴胺快感 / “expensive” dopamine (literal). **###SlangOWL**: 长时间刷屏只会给我廉价的多巴胺快感，而完成一次艰苦的锻炼却能带给我那种无与伦比、仿佛“昂贵多巴胺”般的满足感 / “expensive dopamine”-like satisfaction. **###srag**: 刷手机几个小时只会给我廉价的多巴胺，而完成一次艰苦的锻炼却会带来那种无与伦比的昂贵多巴胺的感觉 / feeling of expensive dopamine. **###cot**: 长时间刷手机只会给我廉价的多巴胺，而完成一次艰苦的锻炼却能带来那种由努力换来的、无与伦比的珍贵满足感 / precious satisfaction earned through effort.

Table 15: Case studies of translations for non-compositional expressions with English annotations.

C.4	**sentence**: I didn't notice I'd been tapping my foot during the meeting until halfway through — that's pure radio cognition. **idiom**: Radio Cognition **sense**: The sudden conscious awareness of an action or thought that had been occurring subconsciously. **reference**: 开会开到一半我才发现自己一直在抖脚，这就是典型的事后才意识到的自动反应 / automatic reaction realized after the fact. ***Translations generated based on the above context:*** **###direct**: 我直到开会进行到一半才发现自己一直在抖脚——这完全是被收音机/广播影响的表现 / influenced by physical radio (wrong). **###SlangOWL**: 我直到开会进行到一半才发现自己一直在抖脚——这纯属被收音机（音乐）无意识带动的反应 / reaction driven by radio music (wrong). **###srag**: 我在会议上一直无意识地跺脚，直到进行到一半才发现——这就是典型的"Radio Cognition" / literal transliteration. **###cot**: 我开会到一半才发现自己一直在跺脚——这就是典型的事后觉察 / typical after-the-fact awareness.
C.5	**sentence**: Her room was a mess for weeks—she wasn't just lazy; she was in full foolaxy. **idiom**: Foolaxy **sense**: A persistent state of extreme laziness marked by lack of motivation, energy, or willingness to act. **reference**: 她的房间好几个星期都乱得一塌糊涂——她不只是懒，而是彻底进入了一种极度摆烂、什么都不想干的状态 / a state of total "letting it rot" and lack of motivation. ***Translations generated based on the above context:*** **###direct**: 她的房间好几个星期都乱糟糟的——她不只是懒，而是彻底傻到家了 / stupid to the extreme (wrong). **###SlangOWL**: 她的房间乱了好几周——她可不只是懒；她完全是在彻底耍蠢 / playing stupid (wrong). **###srag**: 她的房间好几周都乱成一团——她不仅仅是懒；她完全处于 foolaxy 状态 / literal transliteration. **###cot**: 她的房间好几个星期都乱糟糟的——她不只是懒，而是处于极度的懒惰状态 / state of extreme laziness.

Table 16: Additional case studies comparing DEREA against baseline models with semantic annotations.