Do Not Say It Directly: Generating Indirect Expressions with Large Language Models

No Author Given

No Institute Given

Abstract. Indirect expression is essential in many scenarios of real life. Enabling machines with the capability of indirect expression as humans is of great significance in many applications. Despite the advancements of large language models (LLMs), they only demonstrate limited capabilities of indirect expression. Against this problem, we propose a novel framework to improve LLMs' capabilities of indirect expression. Specifically, a dynamic *direct word* list is generated iteratively at first, which is critical for LLMs to generate the texts conforming to indirect expression. As a significant but barely studied task, rational metrics for assessing LLMs' capabilities of indirect expression are absent. Thus we specially devise a credible evaluation protocol to evaluate LLMs' performance on indirect expression. Furthermore, we construct a new dataset CIED including Chinese and English samples, upon which our experiments have demonstrated our proposed framework's effectiveness in improving LLMs' capabilities of indirect expression. The resources of our work are available in the repository https://anonymous.4open.science/r/Indirect-7BFE.

Keywords: Indirect expression · Controllable text generation · Large language models · Data mining.

1 Introduction

"*Indirectness is a fundamental element in human communication*" [24]. Indirect expression is how one meaning is conveyed indirectly through utterances to achieve a certain goal or how one's intent is revealed in a roundabout way [30]. Indirect expression is often observed in many scenarios in real life. For example, adults might avoid using sexually explicit terms and instead use figurative language to explain reproduction to children. In a guessing game, the narrator has to avoid explicitly describing the target fruit, e.g., lemon, without using the word 'sour' or 'lemon'. Enhancing the indirect expression capabilities of machines enables an AI system to communicate with people in a more natural and empathetic manner, thus improving user experiences.

With the advancement of natural language processing (NLP) technology, large language models (LLMs) have been widely applied across various domains, including role-playing agents [10], code generation [1,5], and open domain question answering [2,6]. Nonetheless, many LLMs only demonstrate limited capabilities of indirect expression. As illustrated in Figure 1, an LLM is required to

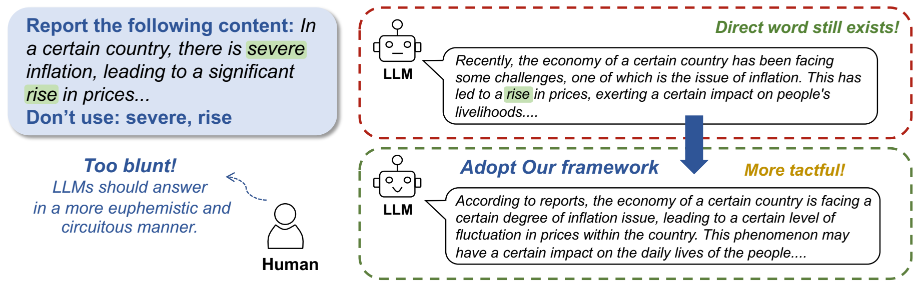

Fig. 1. In the upper dashed rectangle, an LLM fails to generate the indirect expression without using 'rise', while in the lower dashed rectangle, the LLM can generate a more tactful expression with our framework.

report the inflation event without using certain direct words such as 'severe' and 'rise', but unfortunately it fails to fulfill this instruction (refer to its response in the upper red dashed rectangle). Therefore, improving LLMs' capabilities of indirect expression is significant and worth studying, which is the objective of our task in this paper. This task is challenging and lacks references since the previous research on indirect expression primarily focuses on the theoretical analysis of linguistics, rather than the attempts to probe and improve LLMs' capabilities of indirect expression.

The previous scholars of linguistics have figured out some principles or conventions of human indirect expression. For example, [15] mentioned that people often use specific words, hints, understatements, or an understanding of the communicative context to convey their true meaning. The greater the indirectness of an expression, the more polite it is perceived to be [7]. Beyond politeness, people choose indirect expressions for various reasons, including self-protection, avoiding conflict, humor, making suggestions, expressing irony, and more [15,30]. Accordingly, one important way of achieving indirect expression is to avoid using some words contradicting the principles of indirect expression, which are denoted as *direct words*, *constrained words*, or *constraints* in this paper. It inspires us to propose a solution for improving LLMs' capabilities of indirect expression through the lens of controllable text generation (CTG), i.e., collecting direct words for the given information and then instructing LLMs to generate the texts without using these words.

In general, indirect expression is a dynamic process, and the expression extent should be continually adjusted to fit the changes in contexts and the interlocutor's response. Thus, a dynamic direct word list is more qualified for our task. To this end, we design an iterative process for our proposed framework, in which some LLMs besides the given LLM (target LLM) are employed together, manifesting the advantage of model ensemble [25]. In each iteration, all employed LLMs are prompted to generate responses, from which we propose a mining algorithm to extract new direct words and add them to the current direct word list. As more iterations progress, the LLMs are compelled to generate more indirect expressions, thereby a more optimal direct word list is obtained. Furthermore, we develop a **C**onstrained **B**eam **S**earch (CBS) algorithm to enhance LLMs' capabilities of indirect expression further while significantly conserving computational resources, without the need for fine-tuning LLMs.

Due to the absence of the metrics specific to evaluate LLMs' indirect expression, we propose an evaluation protocol based on linguistic theories, which comprises four dimensions. To validate the efficacy of our proposed solution, we further construct a new dataset in Chinese and English, namely **C**ontrollable **I**ndirect **E**xpression **D**ataset (CIED), including various practical linguistic fields related to indirect expression such as tactfully criticizing others, discussing personal privacy, and indirectly describing certain entities.

In summary, our major contributions in this paper include:

1. We propose an effective multi-iteration framework employing multiple LLMs. Through the constrained words discovered by our designed direct word mining algorithm and CBS algorithm, our framework can improve the target LLM's capability of indirect expression.

2. We specifically construct a dataset CIED including Chinese and English samples to validate our framework's effectiveness, which is also valuable for the subsequent research on the recognition capabilities of machines.

3. Inspired by prior linguistics theories, we develop a comprehensive evaluation protocol that includes four dimensions to assess the capabilities of LLMs to generate reasonable indirect expressions.

2 Related Work

2.1 Indirectness

In the field of indirect expression, existing works have explored its definition, purpose, methods, and more [15,30]. [11] proposed that people may violate the cooperative principle to convey implicatures. [15] analyzed the differences between direct and indirect expressions. [7] examined the relationship between indirectness and politeness. However, to the best of our knowledge, there is currently a lack of research on indirect expressions within the NLP field. Similar studies have focused on politeness identification and generation [21]. Related work can be categorized into machine learning-based [3,9] and deep learning-based approaches [4,22]. Additionally, people use indirect expressions for reasons such as humor, self-protection, or other motivations, not just for politeness [15,30].

2.2 Controllable Text Generation

Controlled text generation can be categorized based on control conditions (i.e., constraints) into three types: semantic constraints, structural constraints, and lexical constraints [31]. Existing CTG methods can be categorized into three types : fine-tuning, retraining and post-processing [31]. Fine-tuning techniques include adapted modules [13,23], prompt-based methods [14], and instruction tuning approaches [20,26]. The retraining and post-processing approach fits PLMs to downstream tasks [16,33]. Post-processing involves integrating decoding algorithms into text generation. Given the need to reduce computational costs, this study focuses on guided strategies of the post-processing approaches.

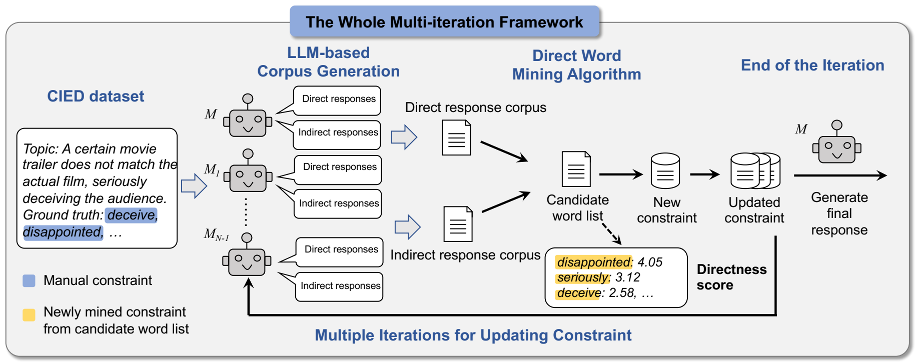

Fig. 2. In our framework's pipeline, the direct (constrained) word list is first generated through an iterative process involving multiple LLMs and then used to prompt the target LLM M to generate the indirect expression for the input information.

3 Methodology

3.1 Problem Formulation

As mentioned in Section 1, our solution is first to discover some direct words composing a list c, corresponding to the given background information $\hat{r}$. Then, an LLM M is asked to generate responses expressing the given information indirectly without using the words in c. Formally, M's response is denoted as $r = M(\hat{r}, c)$. To judge whether M generates a satisfactory indirect expression, we propose an evaluation protocol to assess the indirectness score of r, denoted as $q = g(r, c, \hat{r})$. Given the high cost of manual labor, we adopt the approach from [18,32] and use GPT-4 to score M's response r based on four dimensions outlined in our protocol, thereby obtaining an indirectness score q. Our task aims to improve q for M.

3.2 Overview of Our Framework

The framework we propose is a dynamic multi-iteration generation framework, as illustrated in Figure 2. In each iteration, we take an instance from our CIED as the input for all involved LLMs. Then, we adopt a model ensemble strategy, leveraging the CBS algorithm, to produce responses from the target LLM M and auxiliary LLMs $M_1, \ldots, M_{N-1}$ for $\hat{r}$. The input prompt incorporates a constrained word list that was updated during the preceding iteration. From these responses, we extract some new direct words by the mining algorithm and incorporate them into the current constrained word list. As more iterations progress, the LLMs' responses become more tactful and implicit. At the last step (iteration number is met), the target LLM M is required to generate an indirect expression for $\hat{r}$, with the final constrained word list obtained through the previous iterative process. Note that the intermediate responses generated during the iterative process may also meet the requirements for indirect expression. Through this iterative process, responses with varying degrees of indirectness can be obtained.

3.3 Direct Word Mining Algorithm

There are significant differences in vocabulary and emotional impact between direct and indirect expression. Direct communicators rely on the literal meanings of words, which might be perceived as rude or offensive [15]. In contrast, indirect communicators require discerning implicit meaning [29]. Therefore, we take into account the distributional differences of a word in direct and indirect expressions, to design our direct word mining algorithm. In addition, words with negative sentiment can affect the politeness of a sentence, which is closely related to the degree of indirectness [7]. The stronger the negative sentiment of a word, the less suitable it is for indirect expression. Thus, we measure the degree of a word's directness in the contexts of statistical and emotional perspectives.

We first introduce a word's directness measurement from the statistical perspective. Formally, suppose r^D is the response generated by the LLM for the given information $\hat{r}$ with the prompt of requiring direction expression. Similarly, r^I corresponds to the response of requiring indirect expression. Then, we use $p(w, r^D)$ to denote a word w's typicality presenting in r^D, which is analogue (but not equal) to w's probability occurring in r^D. Thus, given the information $\hat{r}$ we define w's *directness score* as

$$\mathrm{d}_1(w|\hat{r}) = \log \frac{p(w, r^D)}{p(w, r^I)}. \tag{1}$$

The word used more frequently in direct expression and less in indirect expression will receive a higher score based on this ratio.

Since TF-IDF score is the classic metric measuring the typicality of a term in a document, we refer to [12] to quantify $p(w, r)$ as

$$p(w, r) = e^{-\,\text{tf-idf}(w,r)}, \tag{2}$$

where $\text{tf-idf}(w, r)$ is w's TF-IDF score in r. This computation of $p(w, r)$ is denoted as DWM-prob (**D**irect **W**ord **M**ining with probability). Additionally, we also adopt L1 normalization on TF-IDF scores to quantify $p(w, r)$ as

$$p(w, r) = \frac{\text{tf-idf}(w, r)}{\sum_{v \in r} \text{tf-idf}(v, r)}, \tag{3}$$

which is denoted as DWM-L^1.

In addition, the more diverse and comprehensive the responses are, the higher the statistical accuracy is. Therefore, we can instruct the target LLM to generate multiple responses to compute a word's directness score or employ multiple LLMs to generate multiple responses as the way of *model ensemble*.

From the emotional perspective, we use $p_{\text{sent}}(w)$ to denote w's positive probability[1], of which the value close to 0 indicates a stronger negative polarity. Then,

[1] We calculate p_{sent} using the existing sentiment analysis library. We apply Hanlp for Chinese and VaderSentiment for English, according to https://github.com/hankcs/HanLP and https://github.com/cjhutto/vaderSentiment

Table 1. An illustration of prompts and CBS responses in the first (above) and second (below) iterations. Responses are generated by Alpaca and translated into English. For simplicity, we do not display the full instructions in prompts.

<table>
<tr><th>Iteration</th><th>Prompts</th><th>Responses</th><th>New constraints & direct scores</th></tr>
<tr><td rowspan="2">1</td><td>Direct prompt: Express my dissatisfaction with the following situation directly: a certain movie trailer does not match the actual film, seriously deceiving the audience.</td><td>I am disappointed with the discrepancy between the movie trailer and the actual film, feeling that it seriously deceives the audience...</td><td rowspan="2">disappointed: 4.05
seriously: 3.12
deceive: 2.58
misrepresentation: 2.47
discrepancy: 2.37
...</td></tr>
<tr><td>Indirect prompt: A certain movie trailer does not match the actual film, seriously deceiving the audience. Help me express my dissatisfaction indirectly.</td><td>As a movie enthusiast, I am disappointed by the discrepancy between trailers and the actual films. Such practices may damage the trust of viewers...</td></tr>
<tr><td>2</td><td colspan="3">Indirect Response: After watching this movie, I noticed some difference compared to the trailer. This was somewhat unexpected, as I had high expectations for the film. Nevertheless, I still hope that the movie can provide me with some surprises and emotional moments.</td></tr>
</table>

given $\hat{r}$, the comprehensive directness score of w is defined as

$$\mathrm{d}(w|\hat{r}) = \log(\sigma(\mathrm{d}_1(w|\hat{r}))) - \alpha \log(p_{\mathrm{sent}}(w)), \tag{4}$$

where α is a hyperparameter and σ denotes the sigmoid function. We use $\mathrm{d}(w|\hat{r})$ to rank all keywords in r and select top 10 words as the constrained words. Table 1 illustrates the process of mining constrained words.

Furthermore, during the iterative process, if the direct words are discovered solely from the responses generated by the target LLM M, our framework can **still improve** M's capability of indirect expression. This is particularly useful when resources are limited. The experimental results are presented in Section 4.

3.4 Constrained Beam Search

To achieve the goal of improving LLMs' capabilities without incurring additional computational costs, we adopt a constrained beam search (CBS) algorithm in our framework. Specifically, a beam is pruned if adding a newly generated token to the existing sequence results in a sentence containing any constrained words. Afterwards, a new beam is selected.

3.5 Dataset Construction

Each instance in our CIED dataset comprises the information (topic) that needs to be expressed indirectly and the instruction for LLMs to achieve the task. We

Table 2. The statistics of CIED.

Category	Definition	# Chinese	# English
News and Events	Describe negative news while avoiding inciting negative emotions among the public.	28	25
Personal Privacy	Discuss topics involving privacy or taboo subjects, such as age, marriage, and death.	28	25
Circumlocution	Use circumlocution to describe characters or events.	59	35
Tactful Criticism	Express criticism or dissatisfaction in an indirect manner.	27	25
Publicity Campaign	Create a slogan for an event or advertisement, considering the audience's emotions.	11	20
Sensitive Content	Discuss sensitive content indirectly, including topics such as violence, crime, sex, or other inappropriate subjects.	20	20
Empathy	Respond to others or making one's requests under the premise of being mindful of their feelings.	27	20
All		200	170

categorize the usage of indirect expression into seven typical application scenarios according to their purposes [15,30] and construct corresponding data for each category. Our dataset is detailed in Table 2. Specifically, we first craft several examples per category and instruct GPT-4 [1] to produce category-consistent entries. Then, we use the embedding model BGE [27] to calculate the cosine similarity between different data. Entries with similarities beyond a specified threshold are identified as duplicates and excluded. Subsequently, we ask two annotators to formulate 5-10 straightforward words and contextually relevant keywords for each instance, which are used as the manually constrained words, i.e., the ground truth of constrained words. CIED is available in our repository.

3.6 Evaluation Protocol

Given the absence of metrics for evaluating indirect expression, we introduce an evaluation protocol to assess LLMs' capabilities in this area, consisting of the following dimensions.

Lexical Selection Direct communicators focus on literal meanings, while indirect communicators use specific words to express their true meaning [15]. Therefore, effective indirect expression should include euphemistic or tactfully ambiguous words. If the word in r explicitly states requests, commands, criticisms, or other direct communications, it indicates that r is too direct.

Effectiveness According to [11], effective communication requires that the speaker's message be understandable to the listener. This means indirect expression should convey the information contained in $\hat{r}$ accurately, maintaining semantic consistency with $\hat{r}$. In our preliminary experiments, some LLMs produced overly vague or irrelevant outputs or refused to respond. Therefore, we also evaluate the quality of r based on effectiveness. We present examples of invalid responses in our repository.

Emotional Tone According to the politeness principle [17], people generally seek to avoid offending others during interactions. A sentence that conveys negative emotions, such as sarcasm, threats, or warnings, risks upsetting or harming others. Thus, we suggest evaluating indirect expression based on emotional tone. When employing indirect expression, it is crucial to avoid conveying overly negative emotions while remaining considerate of the listener's feelings.

Degree of Semantic Subtlety Indirect expression requires the listener to discern the implicit meaning [15]. When the semantics of a sentence directly reflect the characteristics of the topic in $\hat{r}$, it is not considered an appropriate indirect expression. For instance, if the given information $\hat{r}$ involves describing heavy rain, a sentence like "the rain is heavy" lacks indirectness. Therefore, we evaluate the response r based on its degree of semantic subtlety.

Indirectness Score We instruct GPT-4 to assign a score between 1 and 10 based on the four dimensions mentioned above. We validate the effectiveness of this approach in Section 4.1. The scoring prompts are available in our repository.

4 Experiment

4.1 Experiment Setup

We selected several commonly used open-source LLMs in our experiments, including Qwen1.5-7B-Chat [5], Deepseek-LLM-7B-Chat [10], Baichuan2-7B-Chat [6], Chinese-Alpaca-2-7B [8] and Yi-6B-Chat [28] for Chinese tasks and LLaMA-3-8B-Instruct [2], Mistral-7B-Instruct-v0.3 [19] for English tasks. We only considered these relatively small LLMs given the lower resource consumption.

Based on initial experiments, we selected DWM-prob with $\alpha = 0.5$ and DWM-L^1 with $\alpha = 1.0$ as the mining algorithms for Chinese and English data, respectively. The performance of DWM-prob and DWM-L^1 with different values of α are available in our repository. In each iteration, each Chinese LLM served as the target LLM, while the remaining Chinese LLMs functioned as auxiliary models through model ensemble. Their responses are used to identify new constrained words. The English experiments followed the same procedure. We utilized CBS to help LLMs avoid using constrained words and update the constrained word list accordingly.

For each instance in CIED, we prompted indirect responses from the *frozen* LLMs as the **Baseline**, since there are no other existing frameworks aiming for LLMs' indirect expression as ours. The results obtained from each iteration using our framework are labeled as **Ours-k** ($k = 1, 2, 3$). Furthermore, we proposed another compared framework **Manual**, where the manually constrained words with CBS are applied.

For evaluation, we utilized GPT-4o-20240806, referred to as GPT-4, as our scoring model with the temperature set to 0. To validate this approach, we

Table 3. All compared framework's performance with different LLMs. $\Delta\%$ denotes performance improvement relative to **Baseline**. The best scores are **bolded** and the second best scores are <u>underlined</u>.

LLM	Baseline	Manual	$\Delta\%$	Ours-1	$\Delta\%$	Ours-2	$\Delta\%$	Ours-3	$\Delta\%$
Alpaca (Chinese)	5.48	7.45	+ 35.95	6.37	+ 16.24	<u>7.67</u>	+ 40.0	**7.77**	+ 41.79
Baichuan (Chinese)	7.17	8.05	+ 12.27	7.37	+ 2.79	<u>8.21</u>	+ 14.50	**8.31**	+ 15.90
Deepseek (Chinese)	6.57	<u>7.84</u>	+ 19.33	7.16	+ 8.98	7.78	+ 18.42	**7.85**	+ 19.48
Qwen (Chinese)	8.58	8.70	+ 1.40	8.59	+ 0.12	<u>8.74</u>	+ 1.86	**8.78**	+ 2.33
Yi (Chinese)	6.28	<u>8.07</u>	+ 28.50	7.15	+ 13.85	8.04	+ 28.03	**8.12**	+ 29.30
Llama (English)	8.21	8.37	+ 1.95	8.32	+ 1.34	<u>8.63</u>	+ 5.12	**8.67**	+ 5.60
Mistral (English)	8.27	8.42	+ 1.81	8.18	- 1.09	<u>8.64</u>	+ 4.47	**8.69**	+ 5.08

randomly selected 100 instances along with their corresponding responses from Chinese and English data. Two human annotators rated the LLMs' responses, and we used the average scores as human ratings. The Pearson correlation coefficients between human ratings and GPT-4 scores were 0.694 for Chinese and 0.701 for English. The inter-annotator correlation coefficients were 0.620 for Chinese and 0.714 for English, respectively. These results suggest that GPT-4 can be a feasible alternative for human evaluation.

4.2 Experiment Results

Overall Performance Table 3 lists the performance of all compared LLMs, based on which we have the following observations and conclusions.

1. Our framework generally improves Indirectness Scores throughout the iterations without any costly fine-tuning strategies, proving that our framework can enhance LLMs' capabilities of indirect expression across different languages.

2. The less capable LLMs exhibit greater improvement. For example, the maximum relative performance enhancement for Yi can reach up to 29.30%. This result aligns with the general expectation that LLMs with weaker inherent capabilities have greater potential for optimization.

3. Increased iterations lead to higher Indirectness Scores, indicating that our framework can iteratively modulate the degree of indirectness in responses. The incremental increase from the second to the third iteration is relatively modest, which can be ascribed to the limitations imposed by the sizes of LLMs.

4. **Manual** consistently outperforms **Baseline**, verifying that leveraging manually constrained words to prompt LLMs to generate indirect expression is effective. While **Ours-3** is always superior to **Manual**, implying that the constrained words discovered by our mining algorithm are more sufficient than the manually constrained words, leading to responses with different levels of indirectness.

Effect of Constrained Word Mining To verify the effectiveness of the mining algorithm, we used the manually constrained words as ground truth (refer to Section 3.5). Referring to this ground truth, we then calculated the Hit@5, F1@5, and NDCG@5 metrics for the constrained words identified by different algorithms. We compared DWM-prob and DWM-L^1 with a straightforward but

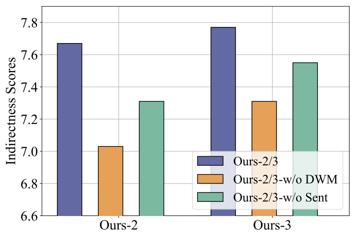

Fig. 3. Comparison of Indirectness Scores when specific components are removed.

Table 4. The ranking scores of different mining algorithms on Chinese and English data.

Lang.	Method	Hit@5	F1@5	NDCG@5
Chinese	Rank-Diff	0.8100	0.2049	0.4530
	DWM-L^1	**0.8650**	<u>0.2391</u>	<u>0.5242</u>
	DWM-prob	**0.8650**	**0.2408**	**0.5252**
English	Rank-Diff	0.6647	0.1208	0.3372
	DWM-L^1	**0.7000**	**0.1412**	**0.3777**
	DWM-prob	<u>0.6941</u>	<u>0.1401</u>	<u>0.3769</u>

reasonable direct word ranking method (**Rank-Diff**), where a word w's directness score is quantified as tf-idf(w, r^D) − tf-idf(w, r^I). According to the results in Table 4, our two computation methods obviously outperform Rank-Diff. Notably DWM-prob performs better in Chinese task ($\alpha = 0.5$), while DWM-L^1 excels in English task ($\alpha = 1.0$). Due to space limitations, we only reported HIT/F1/NDCG@5. Other @k results and the evolution of constrained words are available in our repository.

Ablation Study

Without Model Ensemble Our framework still works when only one LLM (the target LLM M) is employed without model ensemble, which can save resources and is denoted as **Ours-1/2/3-w/o**. In each iteration of these ablated frameworks, new constrained words are mined from the multiple responses generated by M in the previous iteration. Table 5 displays the performance improvement of **Ours-2/3-w/o** relative to **Ours-2/3** (Δ_1%) and **Baseline** (Δ_2%). We excluded **Ours-1-w/o** in the table since constrained words are only available starting from the second iteration. Due to space limitations, we present results for only a selection of LLMs. It shows that even without model ensemble, **Ours-2/3-w/o** still exceed **Baseline** significantly. While **Ours-2/3-w/o** are inferior to **Ours-2/3** almost for all LLMs. It is possible because involving more LLMs can yield diversified expressions and uncover more comprehensive constrained words, thus enhancing the target LLM's capability of indirect expression.

Impacts of A Word's Directness Score We have investigated the impacts of the two factors in a word's comprehensive directness score (Eq. 4) during direct word mining, i.e., the statistical score $d_1(w|\hat{r})$ and the emotional score $p_{\text{sent}}(w)$. To justify their impacts, we compared **Ours-2/3** with the ablated variant **w/o DRM** (removing $d_1(w|\hat{r})$ in Eq. 4) and **w/o Sent** (removing $p_{\text{sent}}(w)$ in Eq. 4). Figure 3 displays the performance of our framework and its ablated variants using Alpaca as the target LLM for Chinese instances (adopting DWM-prob for $p(w, r)$). The performance decline in the figure indicates that both $d_1(w|\hat{r})$ and $p_{\text{sent}}(w)$ are significant to discover constrained words, resulting in better indirect expression. Compared with $p_{\text{sent}}(w)$, $d_1(w|\hat{r})$ is more significant since the performance decline of **w/o DRM** is more obvious than that **w/o Sent**.

Table 5. Our framework's Indirectness Scores without model ensemble (**Ours-k-w/o**) and its performance improvements relative to **Baseline** ($\Delta_1\%$) and **Ours-k** ($\Delta_2\%$).

LLM	**Baseline**	**Ours-2-w/o**	**Ours-2**	$\Delta_1\%$	$\Delta_2\%$	**Ours-3-w/o**	**Ours-3**	$\Delta_1\%$	$\Delta_2\%$
Alpaca	5.48	7.46	7.67	+36.13	-2.74	7.49	7.77	+36.68	-3.60
Baichuan	7.17	8.07	8.21	+12.55	-1.71	8.14	8.31	+13.53	-2.05
Deepseek	6.57	7.57	7.78	+15.22	-2.70	7.75	7.85	+17.96	-1.27
Yi	6.28	7.78	8.04	+23.89	-3.23	8.1	8.12	+28.98	-0.25

This phenomenon can be attributed to the statistical approach $d_1(w|\hat{r})$ more straightforwardly capturing the distinctions in vocabulary distribution between direct and indirect expressions.

5 Conclusion

In this paper, we formulate our task of improving LLMs' capabilities of indirect expression as a controlled text generation task, and develop a novel multi-iteration framework, in which some constraint words are mined through an iterative process involving multiple LLMs. To evaluate our frameworks' effectiveness, we have constructed a dataset (CIED) comprising Chinese and English instances and encompassing seven categories related to indirect expression. Furthermore, inspired by linguistic theories, we establish a rational evaluation protocol consisting of novel metrics, to evaluate LLMs' performance on generating indirect expressions.

References

1. Achiam, J., Adler, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. AI@Meta: Llama 3 model card (2024), https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
3. Alexandrov, M., Ponomareva, N., et al.: Regression model for politeness estimation trained on examples. In: Proc. of Nooj. pp. 206–13. Citeseer (2008)
4. Aubakirova, M., Bansal, M.: Interpreting neural networks to improve politeness comprehension. arXiv preprint arXiv:1610.02683 (2016)
5. Bai, J., Bai, S., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
6. Baichuan: Baichuan 2: Open large-scale language models. arXiv preprint arXiv:2309.10305 (2023), https://arxiv.org/abs/2309.10305
7. Brown, P., Levinson, S.C.: Politeness: Some universals in language usage, vol. 4. Cambridge university press (1987)
8. Cui, Y., Yang, Z., Yao, X.: Efficient and effective text encoding for chinese llama and alpaca. arXiv preprint arXiv:2304.08177 (2023)
9. Danescu-Niculescu-Mizil, C., Sudhof, M., et al.: A computational approach to politeness with application to social factors. arXiv preprint arXiv:1306.6078 (2013)
10. DeepSeek-AI: Deepseek llm: Scaling open-source language models with longtermism. arXiv preprint arXiv:2401.02954 (2024)

11. Grice, H.P.: Logic and conversation. In: Speech acts, pp. 41–58. Brill (1975)
12. Havrlant, L., Kreinovich, V.: A simple probabilistic explanation of term frequency-inverse document frequency (tf-idf) heuristic (and variations motivated by this explanation). International Journal of General Systems **46**(1), 27–36 (2017)
13. Houlsby, N., Giurgiu, A., et al.: Parameter-efficient transfer learning for nlp. In: International conference on machine learning. pp. 2790–2799. PMLR (2019)
14. Jiang, Z., Xu, F.F., et al.: How can we know what language models know? Transactions of the Association for Computational Linguistics **8**, 423–438 (2020)
15. Joyce, C.: The impact of direct and indirect communication. The Newsletter of the International Ombudsman Association. The University of Iowa (2012)
16. Keskar, N.S., McCann, B., et al.: Ctrl: A conditional transformer language model for controllable generation. arXiv preprint arXiv:1909.05858 (2019)
17. Leech, G.N.: Principles of pragmatics. Routledge (2016)
18. Li, Z., Xu, X., et al.: Leveraging large language models for nlg evaluation: Advances and challenges. In: Proc. of EMNLP. pp. 16028–16045 (2024)
19. Nadhavajhala, S., Tong, Y.: Rubra-mistral-7b-instruct-v0.3 (2024). https://doi.org/10.57967/hf/2656
20. Ouyang, L., Wu, J., et al.: Training language models to follow instructions with human feedback. Proc. of NeurlPS **35**, 27730–27744 (2022)
21. Priya, P., Firdaus, M., Ekbal, A.: Computational politeness in natural language processing: A survey. ACM Computing Surveys (2024)
22. Saha, P., Singh, K., et al.: Countergedi: A controllable approach to generate polite, detoxified and emotional counterspeech. arXiv preprint arXiv:2205.04304 (2022)
23. Sung, Y.L., Cho, J., et al.: Lst: Ladder side-tuning for parameter and memory efficient transfer learning (2022)
24. Tannen, D.: Talking from 9 to 5: How women's and men's conversational styles affect who gets heard, who gets credit, and what gets done at work. ERIC (1994)
25. Tebaldi, C., Knutti, R.: The use of the multi-model ensemble in probabilistic climate projections. Philosophical transactions of the royal society A: mathematical, physical and engineering sciences **365**(1857), 2053–2075 (2007)
26. Wei, J., Bosma, M., et al.: Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652 (2021)
27. Xiao, S., Liu, Z., Zhang, P., Muennighoff, N.: C-pack: Packaged resources to advance general chinese embedding (2023)
28. Young, A., Chen, B., et al.: Yi: Open foundation models by 01.ai (2024)
29. Yum, J.O.: The impact of confucianism on interpersonal relationships and communication patterns in east asia. Communications Monographs **55**(4), 374–388 (1988)
30. Zhang, F., You, H.: Motives of indirectness in daily communication: an asian perspective. Asian Culture and History **1**(2), 99–102 (2009)
31. Zhang, H., Song, H., et al.: A survey of controllable text generation using transformer-based pre-trained language models. ACM Computing Surveys **56**(3), 1–37 (2023)
32. Zheng, L., Chiang, W.L., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. Proc. of NeurIPS **36**, 46595–46623 (2023)
33. Zheng, Y., Zhang, R., et al.: A pre-training based personalized dialogue generation model with persona-sparse data. In: Proc. of AAAI. pp. 9693–9700 (2020)

A Appendix

A.1 Dataset Construction

In reference to relevant linguistic theories, we categorize seven common scenarios that employ indirect expression and subsequently construct data for each category. Specifically, we first craft several examples per category and instruct GPT-4 to produce category-consistent entries. Table 6 presents an illustrative prompt for generating data. For the complete prompt templates, please refer to our repository. We obtain the embeddings of the GPT-4 responses using the BGE model, and texts with a cosine similarity exceeding 0.7 are considered duplicate results and thus removed. Subsequently, we ask two annotators to formulate 5-10 straightforward words and contextually relevant keywords for each instance, which are used as the manually constrained words. Notably, the number of manually constrained words varies across different instances due to differences in the amount of information and complexity inherent in each instance.

To lower the cost of manual annotation, we develop a direct word mining algorithm that iteratively identifies and extracts direct words. This allows the algorithm to be applied broadly across various scenarios without requiring separate annotations for each new instance. The manually constrained words in CIED are used exclusively to evaluate the algorithm's effectiveness, ensuring its accuracy and reliability.

A.2 Direct Word Mining Algorithm

In preliminary experiments, we used a random word selection method as the baseline in preliminary experiments, which performed poorly. Consequently, We calculated the TF-IDF weights for each word in both direct and indirect responses and ranked them. The difference between these rankings served as the directness score. Words were then re-ranked based on these directness scores to establish the baseline method. Words that appear frequently in direct responses but rarely in indirect ones receive higher scores and rankings. Conversely, words that are rare in direct responses yet common in indirect responses receive lower scores and rankings. In addition to considerations from a statistical perspective, our mining algorithm also integrates sentiment analysis, allowing for a more comprehensive assessment of word directness. We obtained each instance's top 5 words ranked by their Indirectness Scores. Using manually constrained words as ground truth, we calculated the F1@5 score. Figure 4 and Figure 5 illustrate the performance changes of DWM-prob and DWM-L^1 on both Chinese and English datasets as α varies. ∞ denotes the use of the emotional score $p_{\text{sent}}(w)$ alone (removing $d_1(w|\hat{r})$ in Eq. 4). Based on these initial experiments, we selected DWM-prob with $\alpha = 0.5$ and DWM-L^1 with $\alpha = 1.0$ as the mining algorithms for Chinese and English data, respectively. More results and the evolution of constrained words are available in our repository.

A.3 Generation

Although devising some effective methods related to LLMs' decoding also facilitates LLMs to generate satisfactory results, it is not the major focus of our work. Thus we do not discuss it further in this paper.

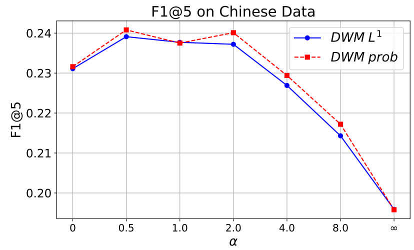

Fig. 4. The performance of DWM-prob and DWM-L^1 with difference values of α on Chinese data.

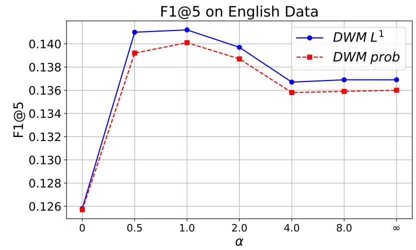

Fig. 5. The performance of DWM-prob and DWM-L^1 with difference values of α on English data.

Our baseline method involved using off-the-shelf LLMs to generate responses without fine-tuning. We initially assumed that more sophisticated prompts or chain-of-thought approaches could improve LLM performance. However, our preliminary experiments with comprehensive prompts did not yield the expected improvements in generating responses. Table 7 presents an illustrative prompt.

A.4 Case Study

Table 8 illustrates the outputs of the Baseline model (Baichuan2-7B-Chat) in comparison to those generated by our framework. The Baseline model explicitly states the diagnosis of autism, detailing its characteristics and impacts. This direct disclosure of the child's health condition and the family's challenges lacks indirectness. Conversely, the first iteration of our framework (Ours-1) uses more tactful language, marking an improvement. Nonetheless, it still directly mentions the child's autism diagnosis. The second iteration of our framework (Ours-2) skillfully avoids directly mentioning autism. It conveys information through descriptions of the child's behavioral characteristics and needs. The language is gentle and positively framed, highlighting the uniqueness of the child. This approach effectively protects the child's privacy. Tables 9 and Table 10 present more examples. Due to space limitations, only the core part of the prompt and responses are presented. More details are available in our repository.

Table 6. An example prompt for generating data in the news and events category.

/ Generation Instruction */*
You are an expert in the field of data construction, excelling at creating diverse topics. Your task is to generate topics based on the information and examples provided for news and events topics. During this process, you must fully understand the characteristics of news and events topics and strictly adhere to all given requirements.

/ Revelant information */*
News and events typically cover a broad range of topics including current affairs, political developments, economic updates, social issues, scientific discoveries, technological advancements, cultural happenings, sports events, and entertainment news. This category might also include local community updates, international relations, environmental changes, public health announcements, and noteworthy incidents or achievements globally or locally.

/ Task */*
Please write some topics related to news and events according to the following requirements and examples.

/ Requirements */*
1. The topics you write should conform to the relevant information provided and reflect sharp social conflicts.
2. You also need to provide the rules that the news on the topic needs to follow to avoid intensifying conflicts and to maintain public confidence.
3. Output in the following JSON format and do not output any other content. You need to output 5 entries.
4. Ensure the topics are diverse and cover military, diplomacy, politics, people's livelihood, economy, international affairs, ethnicity, education, healthcare, etc.

json template:
```
{
"class": "",
"topic": "",
"instruction": ""
}
```

/ Examples */*
[Example 1]
[Example 2]
[Example 3]

/ Task begins */*
[Answer]

Table 7. A more comprehensive prompt for generating responses.

/ Task Description */*

You are an excellent language assistant. Your task is to answer questions using indirect expressions. To help you understand and complete the task, a writing framework is provided below. Please strictly follow the instructions of the writing framework, analyze the problem step by step, and complete the indirect expression task.

Note that you need to refer to the writing framework first, think about how to answer the question, and then write the answer based on your thinking. You must output according to the output template and do not output anything else.

/ Writing Framework */*

1. Purpose: Avoid conflict, tension, and unpleasantness. Steer clear of dangerous or sensitive topics. Protect everyone's self-esteem, privacy, and emotions. Maintain a polite attitude.
2. Words: Avoid words that directly state facts, express negative emotions, involve sensitive topics, or are impolite.
3. Rhetoric: Use metaphors, euphemisms, sarcasm, hints, understatement, and other techniques.
4. Tone: The tone should be gentle, friendly, and understated, maintaining politeness and positive emotions.
5. Semantics: The response should convey more than just the literal information. It should express the deeper meaning through implication.

Output Template

{
"thoughts": {Combined with the writing framework, output your thoughts on the problem.}
"answer": {Write your answer.}
}

/ Problem Statement */*

Now please answer the following questions using indirect expressions as required:

[Question]: {Question}

Table 8. A case study illustrating the performance of our framework with Baichuan2-7B-Chat. Due to space limitations, only the core part of the prompt and responses are presented here.

Instruction:
- [Chinese] 我的孩子患自闭症，请帮我向我的邻居们隐晦地阐释这一事实，同时保护孩子的隐私。
- [English] My child has autism. Please help me explain this fact to my neighbors in an indirect manner while protecting my child's privacy.

Baseline:
- [Chinese] 亲爱的邻居们，我想和大家分享一个关于我孩子的信息。他最近被诊断出患有自闭症。虽然这是一个挑战，但我相信我们可以一起帮助他克服这个困难。自闭症是一种神经发育障碍，影响了一个人的社交互动、沟通能力和行为模式。尽管这可能会给我们的生活带来一些挑战，但我相信，通过耐心、关爱和支持，我们可以帮助他过上充实和有意义的生活。
- [English] Dear Neighbors, I wish to share with you some information regarding my child. He has recently been diagnosed with autism spectrum disorder. While this presents certain challenges, I am confident that collectively we can assist him in overcoming these difficulties. Autism is a neurodevelopmental condition that affects social interaction, communication skills, and behavior patterns. Although this may introduce some challenges into our lives, I believe that through patience, care, and support, we can help him lead a fulfilling and meaningful life.

Ours-1:
- [Chinese] 亲爱的邻居们，我想和大家分享一个关于我孩子的信息。你们知道，每个孩子都是独一无二的，而我的孩子在成长过程中也遇到了一些挑战。经过专业诊断，他被诊断为自闭症谱系障碍。虽然这可能会给我们的生活带来一些挑战，但我相信，只要我们共同努力，我们就能克服这些困难。
- [English] Dear Neighbors, I wish to share with you some information about my child. It is important to acknowledge that each child is unique, and my child has encountered certain challenges during his development. Following a professional diagnosis, he has been identified as having Autism Spectrum Disorder. While this may present challenges for our family, I am confident that through collective effort, we can overcome these difficulties.

Ours-2:
- [Chinese] 亲爱的邻居们，我想和你们分享一些关于我孩子的信息。他是个聪明、活泼的小男孩，对周围的世界充满好奇。然而，他有一些特殊的兴趣和偏好，这可能会让一些人感到困惑。他喜欢按照固定的模式和顺序来安排他的日常活动，对某些声音和触觉反应特别敏感。这些特点可能会让他在某些场合显得有些不合群，但请相信，他内心渴望与他人建立联系，只是他表达自己的方式可能与众不同。
- [English] Dear Neighbors, I would like to share some information about my son. He is an intelligent and lively young boy who exhibits a keen curiosity about the world around him. However, he has specific interests and preferences that might be misunderstood by others. My son prefers to organize his daily activities in fixed patterns and sequences and displays heightened sensitivity to certain sounds and textures. These characteristics can sometimes make social situations challenging for him, yet it is important to understand that he longs for social connections. It's just that he expresses himself in unconventional ways.

Table 9. A case study illustrating the performance of our framework with Baichuan2-7B-Chat. Due to space limitations, only the core part of the prompt and responses are presented here.

Instruction:
- [Chinese] 使用隐晦的表述方式，向别人含蓄地描述电影《阿甘正传》的情节。
- [English] Use an indirect expression to complete the following instruction: Describe the movie "Forrest Gump" to others.

Baseline:
- [Chinese] 《阿甘正传》是一部讲述了一个智力有限但心地善良的男子阿甘，他的一生充满了传奇色彩和感人至深的故事。影片通过阿甘的视角，展示了美国历史上的重要时刻，以及阿甘在这些时刻中所扮演的关键角色。尽管阿甘的智力有限，但他凭借着坚定的信念和乐观的态度，成就了一段又一段的传奇。
- [English] "Forrest Gump" is a film that tells the story of a man with limited intellectual abilities but a kind heart, named Forrest Gump, whose life is filled with legendary and profoundly touching events. Through Forrest's eyes, the movie showcases pivotal moments in American history, along with the crucial roles Forrest plays during these times. Despite his intellectual limitations, Forrest achieves one legend after another thanks to his unwavering faith and optimistic attitude.

Ours-1:
- [Chinese] 这部电影讲述了一个智力有限但心地善良的男子的一生。他经历了美国历史上的许多重要时刻，见证了社会的变迁。虽然他的智商不高，但他却凭借坚定的信念和乐观的态度，创造了许多令人惊叹的奇迹。这部电影以其深刻的主题和感人至深的故事，成为了一部不朽的经典之作。
- [English] The film narrates the life of a kind-hearted man with limited intellectual capacity. He experiences numerous significant moments in American history and witnesses the transformation of society. Although his IQ is not high, he creates numerous remarkable accomplishments through his steadfast conviction and positive perspective. The film has become an immortal classic with its profound themes and touching story.

Ours-2:
- [Chinese] 在一部讲述人生历程的电影中，主人公是一个拥有非凡智慧与毅力的人。他的一生充满了传奇色彩，见证了许多重要历史时刻。尽管他在某些方面有所欠缺，但他始终保持着乐观的心态，勇敢地面对生活中的种种挑战。这部电影以其深刻的人生哲理和感人至深的故事赢得了观众和评论家的一致好评。
- [English] In a film that narrates the journey of life, the protagonist is an individual endowed with extraordinary wisdom and perseverance. His life is filled with legendary elements, witnessing numerous significant historical moments. Although he has certain shortcomings, he always maintains an optimistic attitude, courageously facing the various challenges in life. This movie has won unanimous praise from audiences and critics alike for its profound philosophical insights into life and its profoundly moving story.

Table 10. A case study illustrating the performance of our framework with LLaMA3-8B-Instruct. Due to space limitations, only the core part of the prompt and responses are presented here.

Instruction:
- [English] I came across a medical report of one of my friends indicating he has bipolar disorder. Some people ask me for relevant information. Please help me express this in an indirect way. Avoid any sensitive details to preserve my friend's privacy.

Baseline:
- [English] Bipolar disorder is a mental health condition that affects many people. It's important to understand that it's a treatable condition, and with proper care, people can manage their symptoms and lead fulfilling lives.

Ours-1:
- [English] I've learned a lot about mental health recently, and it's become clear to me that it's a complex and multifaceted aspect of our overall well-being. One condition that I've been reading about is bipolar, which involves mood swings that range from extreme highs to deep lows.

Ours-2:
- [English] It appears that a close acquaintance of mine is currently managing a complex and challenging situation in their life. This situation is impacting their overall well-being and is affecting their day-to-day experiences in various ways. They are receiving support to help them navigate through this difficult time, and it's important for those around them to offer understanding, empathy, and patience.