Mitigating Hallucinations of Large Language Models in Medical Information Extraction via Contrastive Decoding

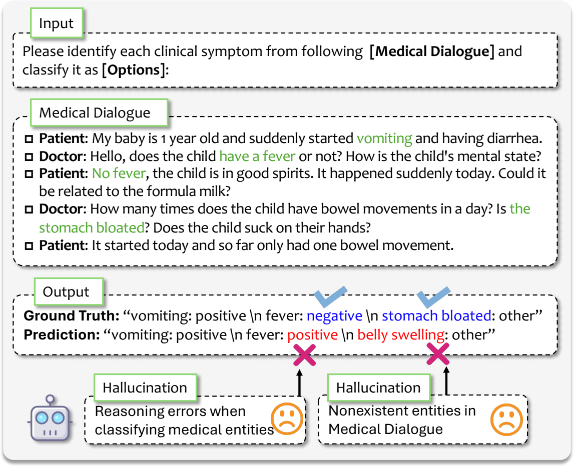
Anonymous ACL submission

Abstract

The impressive capabilities of large language models (LLMs) have attracted extensive interests of applying LLMs to medical field. However, the complex nature of clinical environments presents significant hallucination challenges for LLMs, hindering their widespread adoption. In this paper, we address these hallucination issues in the context of Medical Information Extraction (MIE) tasks by introducing ALternate Contrastive Decoding (ALCD). We begin by redefining MIE tasks as an *identify-and-classify* process. We then separate the identification and classification functions of LLMs by selectively masking the optimization of tokens during fine-tuning. During the inference stage, we alternately contrast output distributions derived from sub-task models. This approach aims to selectively enhance the identification and classification capabilities while minimizing the influence of other inherent abilities in LLMs. Additionally, we propose an alternate adaptive constraint strategy to more effectively adjust the scale and scope of contrastive tokens. Through comprehensive experiments on two different backbones and six diverse medical information extraction tasks, ALCD demonstrates significant improvements in resolving hallucination issues compared to conventional decoding methods.

1 Introduction

Medical Information Extraction (MIE), including tasks such as medical entity recognition and relation extraction, is a fundamental component of medical NLP (Hahn and Oleynik, 2020). It enables the derivation of structured knowledge from plain text, benefiting a wide array of applications, like medical knowledge graph construction (Wu et al., 2023; Xu et al., 2024), medical dialogue (Gao et al., 2023; Wu et al., 2024), and medical report generation (Liu et al., 2021). Previous MIE tasks (Yu et al., 2019; Guan et al., 2020) have been supervised, and their performance heavily depends on the quality and quantity of available training data. However, labeling medical documents requires specific knowledge which is both costly and time-consuming.


Input
Medical Dialogue
Output
Hallucination
Reasoning errors when classifying medical entities
Hallucination
Nonexistent entities in Medical Dialogue

Please identify each clinical symptom from following **[Medical Dialogue]** and classify it as **[Options]**:

- □ **Patient:** My baby is 1 year old and suddenly started vomiting and having diarrhea.
- □ **Doctor:** Hello, does the child have a fever or not? How is the child's mental state?
- □ **Patient:** No fever, the child is in good spirits. It happened suddenly today. Could it be related to the formula milk?
- □ **Doctor:** How many times does the child have bowel movements in a day? Is the stomach bloated? Does the child suck on their hands?
- □ **Patient:** It started today and so far only had one bowel movement.

Ground Truth: “vomiting: positive \n fever: negative \n stomach bloated: other”
Prediction: “vomiting: positive \n fever: positive \n belly swelling: other”

Figure 1: An example demonstrating the hallucination generated by LLMs in MIE tasks. The green font in medical dialogue indicates a high correlation with ground truth. The blue font in the output represents correct token, while the red font represents tokens with hallucination problems. These problems mainly include the presence of nonexistent entities and reasoning errors.

Recently, the remarkable zero-shot capabilities of large language models (LLMs) such as ChatGPT and GPT-4 (OpenAI, 2023) have inspired researchers to transform MIE tasks into a generation paradigm (Zhu et al., 2023). However, the medical domain is less tolerant of errors compared to other domains. While there have been attempts to apply LLMs to the medical field (Singhal et al., 2022; Sharma et al., 2023; Liu et al., 2024), there is a growing concern about the issue of hallucination (Huang et al., 2023). In the context of MIE, two types of hallucinations exist: (1) LLMs may identify medical entities that are not present in original texts, thereby fabricating facts and deviating from the original information. (2) LLMs may face reasoning errors when classifying medical entities, due to statistic biases in the pre-trained corpus. We show such a hallucination problem in Figure 1.

In this paper, we address the challenges of hallucination when applying LLMs to MIE tasks. We observe that LLMs for MIE can be conceptualized as an *identify-and-classify* process: initially identifying potential medical concept spans from the plain text, and then classifying these text spans into predefined categories (e.g., start token of a specific entity, subject of a specific relation), as shown in the 'Output' of Figure 1. The natural approach to applying LLMs is to prompt them to simultaneously complete both *identify* and *classify* steps in a unified decoding process (Lu et al., 2022; Wang et al., 2023b). We speculate that the hallucination problem may be linked to the joint next-word generation abilities of identification and classification, which could have inadvertently compromised each other's performance. Therefore, we believe that decoupling abilities of identification and classification, allowing LLMs to concentrate on specific sub-tasks, could simplify the complexity of the MIE task and potentially reduce hallucination issues (Khot et al., 2022; Bian et al., 2023).

Motivated by the aforementioned observation, we introduce ALternate Contrastive Decoding (ALCD), a straightforward decoding strategy designed to enhance the performance of LLMs on MIE tasks. In the training stage, we mask the optimization of tokens separately to decouple the identification and classification models. For instance, when fine-tuning the parameters of the identification model, classification tokens are masked to focus the model's attention solely on identification tokens, thereby ignoring its classification capability. During the inference stage, ALCD bolsters its classification/identification ability and contrasts logit predictions with another model. This contrastive decoding process alternates between classification and identification, depending on the type of the next token, which is determined by a simple rule-based judgment. Furthermore, we propose an adaptive constraint strategy to dynamically adjust the scale and scope of contrastive tokens. This allows individual samples to adapt to their unique characteristics by measuring the consistency among the three models and the level of confidence. Overall, this work makes three key contributions:

- To our knowledge, we are the first to employ contrastive decoding as a strategy to reduce hallucinations in LLMs for MIE tasks.
- We validate the broad applicability of our ALCD approach through experiments using two LLM backbones across six diverse medical tasks, such as determining causal relationships in medical concepts (Zhu et al., 2023).
- Our experimental results underscore the superiority of ALCD over eight established decoding methods. Codes will be released [1].

2 Related Work

2.1 LLMs for Medical Domain

Rapid development has been seen in directly employing general LLMs (e.g., ChatGPT (OpenAI, 2023), ChatGLM (Du et al., 2022), and Qwen (Bai et al., 2023)) to the medical domain and training medical LLMs using medical data, such as Med-PaLM (Singhal et al., 2022), clinicalGPT (Wang et al., 2023a), and MedAlpaca (Han et al., 2023). Both general LLMs and medical LLMs may suffer from hallucinations, the undesired phenomenon of LLMs generating contents not based on training data or facts when applying them to complex medical tasks. Hallucinations could be caused by multiple factors, such as imperfect representation learning or erroneous decoding (Ji et al., 2023a). Due to the high demand for reliability in the medical domain, the hallucinations are thus less tolerated. Although previous works have explored the problem of hallucination in the medical domain (Umapathi et al., 2023; Ji et al., 2023b), there is a lack of exploration in MIE task, particularly regarding the efficiency of different decoding methods for mitigating hallucination.

2.2 Contrastive Decoding

The idea of contrastive decoding for LLM has been explored in various previous works, and different decoding strategies focus on different aspects of LLM improvements. Contrastive Decoding (CD) (Li et al., 2023) is proposed to contrast output probability of large-scale expert LLMs with small-scale amateur LLMs to diminish undesired amateur behavior and improve fluency and coherence in the generated contents. Context-aware Decoding (CAD) (Shi et al., 2023) focuses on the issue of LLMs' insufficient attention to context. CAD downweights output probability associated with LLMs' prior knowledge to promote LLMs' attention to context, thus improving the faithfulness

[1] https://anonymous.4open.science/r/ALCD-8831

of the generated contents. Chuang et al. (2024) introduced DoLa, where the output next-word probability is obtained from the difference in logits between a higher layer versus a lower layer, to reduce hallucinations and enhance truthfulness in the knowledge-based question-answering tasks. Visual Contrastive Decoding (VCD) is another decoding method to mitigate object hallucinations for large vision-language models by contrasting output distributions from original and distorted visual inputs (Leng et al., 2023). Sanchez et al. (2023) adapted Classifier-Free Guidance (CFG) (Ho and Salimans, 2022) from text-to-image generation to text-to-text generation and they showed CFG can increase the LLMs' performance and adherence to various prompts, including basic prompting, chain-of-thought prompting, and chatbot prompting.

Although previous contrastive decoding strategies have been shown effective in addressing specific hallucinations in LLMs, their performance is inadequate for MIE tasks. In contrast, our ALCD effectively decouples the abilities to contrast and decode outputs, leading to notable enhancements.

3 Methodology

In this section, we introduce ALternate Contrastive Decoding (ALCD), a method specifically designed for medical information extraction tasks. Section 3.1 provides the foundational knowledge of Contrastive Decoding, while Section 3.2 delves into the details of our proposed ALCD method.

3.1 Preliminary

For generative LLMs, the common method for text generation is to predict next token in an autoregressive manner. Specifically, we denote the parameters of an LLM as θ. The model utilizes input text $\boldsymbol{x}$ and system instructions (prompts) $\boldsymbol{i}$ to generate a response $\boldsymbol{y}$. For each time step t, we have:

$$\begin{aligned} y_t \sim & \mathcal{P}_\theta(y_t|\boldsymbol{i}, \boldsymbol{x}, \boldsymbol{y}_{<t}), \\ & \sim softmax(logit_\theta(y_t|\boldsymbol{i}, \boldsymbol{x}, \boldsymbol{y}_{<t})), \end{aligned} \tag{1}$$

where y_t represents the output token at a specific time step t, and $\boldsymbol{y}_{<t}$ denotes the sequence of generated token sequence until the time step $t-1$. The common ways of the next token selection include selecting the highest probability token (greedy search), exploring multiple high-probability paths simultaneously (beam search), or sampling according to the probability distribution (e.g., nucleus sampling (Holtzman et al., 2019)).

While, in contrastive decoding, there are typically two logits, which may be obtained from different LLMs using the same input source (Li et al., 2023) or the same LLM using different input sources (Shi et al., 2023). It should be noted that they need to share the same tokenizer to keep consistency between different logits. The probability for the next token is adjusted through subtraction:

$$logit_\theta(y_t|\boldsymbol{i}, \boldsymbol{x}, \boldsymbol{y}_{<t}) - logit_{\theta'}(y_t|\boldsymbol{i}, \boldsymbol{x}, \boldsymbol{y}_{<t}). \tag{2}$$

The $logit_\theta$ and $logit_{\theta'}$ are usually generated from an LLM with high capabilities and low capabilities, respectively. For example, in CD (Li et al., 2023), $logit_\theta$ comes from a large expert LLM and $logit_{\theta'}$ comes from a small amateur LLM. Subtracting these two logits helps amplify the ground-truth tokens in $logit_\theta$ and downplay hallucinated tokens in $logit_{\theta'}$. Inspired by CD, we propose to alternately amplify or downplay the classification and identification capabilities of LLMs during the decoding process, to improve final generation results.

3.2 Alternate Contrastive Decoding

The process of our proposed ALCD is illustrated in Figure 2. We break down medical information extraction into two stages: identification and classification. In Section 3.2.1, we fine-tune LLMs separately for identification and classification. In Section 3.2.2, we utilize the decoders of three LLMs (identification, classification, and normal) together to perform MIE. As the two new LLMs are trained with Lora (Hu et al., 2021), they do not cause an excessive increase in parameter volume.

3.2.1 Decoupling with optimization masking

To effectively harness identification and classification capabilities of LLMs while minimizing interference from one another, we propose to decompose their respective abilities. Typically, it is natural to fine-tune two subtasks independently, resulting in a identification model $\mathcal{M}_{id}$ and a classification model $\mathcal{M}_{cl}$. But this method has distinct instructions and input-output formats compared to normal model $\mathcal{M}_{nl}$. It poses an issue when these models are combined during the inference step, which can lead to inconsistent input with fine-tuning step, ultimately reducing the accuracy.

In this work, we propose to optimize two capabilities separately using optimization masking during the fine-tuning process, as shown in Figure 2(Step #1). We employ the same inputs as original task for fine-tuning both $\mathcal{M}_{id}$ and $\mathcal{M}_{cl}$ models.

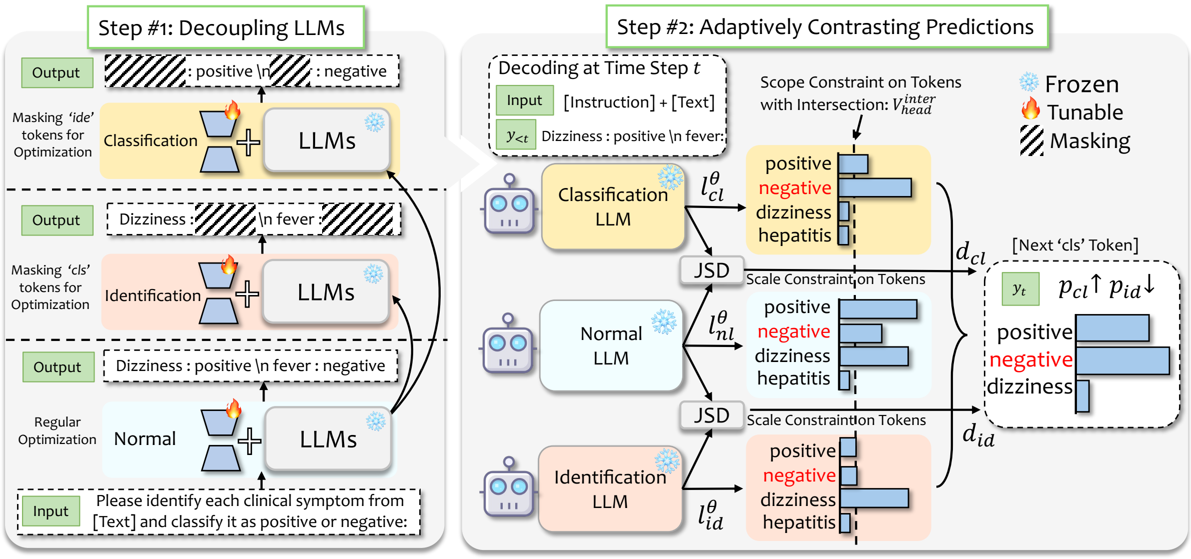


Figure 2: The overall pipeline of our proposed ALCD consists of two main steps. In step #1, our goal is to fine-tune sub-models individually in order to decouple the abilities of identification and classification. In step #2, our objective is to adaptively contrast the predictions at each time step by applying scale and scope constraints on tokens.

During fine-tuning, we selectively optimize tokens, and for instance, when optimizing parameter θ_{id} of identification model $\mathcal{M}_{id}$, we mask the tokens for classification task:

$$\max_{\theta_{id}} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \sum_{t=1, t\notin\mathcal{T}_{cl}}^{|y|} log(\mathcal{P}_{\theta_{id}}(y_t|\boldsymbol{i},\boldsymbol{x},\boldsymbol{y}_{<t})), \quad (3)$$

where $\mathcal{T}_{cl}$ represents the time step of classification tokens, which do not require optimization, and $\mathcal{D}$ denotes training dataset. On the other hand, when optimizing parameter θ_{cl} of classification model $\mathcal{M}_{cl}$, we mask the tokens for identification task:

$$\max_{\theta_{cl}} \sum_{(\boldsymbol{x},\boldsymbol{y})\in\mathcal{D}} \sum_{t=1, t\notin\mathcal{T}_{id}}^{|y|} log(\mathcal{P}_{\theta_{cl}}(y_t|\boldsymbol{i},\boldsymbol{x},\boldsymbol{y}_{<t})), \quad (4)$$

where $\mathcal{T}_{id}$ represents time step of identification tokens. By employing masking optimization, we expect to develop LLMs that possess diverse capabilities. For fine-tuning normal model $\mathcal{M}_{nl}$, we also employ formulas similar to 3 and 4, but without any masking operations. Given the constraints of computational resources, we implemented parameter-efficient fine-tuning techniques (e.g., LoRA (Hu et al., 2021)) to train these models.

3.2.2 Adaptively Contrasting the Predictions

After decoupling the capabilities, a significant challenge arises: how can we effectively harness the individual abilities of sub-models? To address this, ALCD is designed to alternate the enhancement of the classification ability of $\mathcal{M}_{cl}$ and the identification ability of $\mathcal{M}_{id}$ during LLM's inference stage, while excluding the influence of other capabilities originally present in normal model $\mathcal{M}_{nl}$. An illustration is shown in Figure 2(Step #2).

We denote $n_t \in \{cls, ide, other\}$ as the type of next token prediction, where $cls, ide, other$ indicate classification, identification, and other tokens, respectively. Generally, in order to facilitate the evaluation of text generated from LLMs, it is typically to present the output of MIE in a structured format (Lu et al., 2022). Therefore, when LLMs generate token y_t at time t, we can determine the next token based on previous tokens $\boldsymbol{y}_{<t}$ using a simple rule-based judgment: In our case, we require LLMs to utilize colon ':' to split ide and cls tokens, and each ide-cls pair is separated by a newline character '\n'. For instance, in this text: *"Dizziness: positive\n fever: negative"*, the ide tokens (*Dizziness* or *fever*) are expected to be followed by a colon and then a cls token (*positive* or *positive*). We abbreviate the representation $logit_{\theta}(\cdot|\boldsymbol{i},\boldsymbol{x},\boldsymbol{y}_{<t})$ generated by $\mathcal{M}_{nl}$, $\mathcal{M}_{cl}$, and $\mathcal{M}_{id}$ as l_{nl}^{θ}, l_{cl}^{θ}, and l_{id}^{θ}, respectively. The overall formula is as follows:

$$\begin{cases} l_{nl}^{\theta} + \alpha(l_{nl}^{\theta} + l_{cl}^{\theta} - (d_{id} * l_{id}^{\theta} + d_{cl} * l_{cl}^{\theta})), \\ \qquad\qquad \text{if } n_t = cls \\ l_{nl}^{\theta} + \alpha(l_{nl}^{\theta} + l_{id}^{\theta} - (d_{cl} * l_{cl}^{\theta} + d_{id} * l_{id}^{\theta})), \\ \qquad\qquad \text{if } n_t = ide \end{cases} \quad (5)$$

where α is a hyper-parameter and analyzed in Section 4.4. d_{id} and d_{cl} are adaptive scales proposed to measure the distance between two logit distributions: one between $\mathcal{M}_{nl}$ and $\mathcal{M}_{id}$, and the other between $\mathcal{M}_{nl}$ and $\mathcal{M}_{cl}$. We leverages Jensen-Shannon Divergence (JSD) to calculate them:

$$\begin{aligned} d_{id} &= JSD(logit_{\theta_{nl}}||logit_{\theta_{id}}), \\ d_{cl} &= JSD(logit_{\theta_{nl}}||logit_{\theta_{cl}}). \end{aligned} \tag{6}$$

Specifically, when predicting the next token in Formula (5), ALCD includes two extra components in addition to the logit l^{θ}_{nl} of the normal model. For example, if n_t is a cls token, The first component is enhancing l^{θ}_{cl}, with the motivation to utilize the classification ability of sub-model $\mathcal{M}_{cl}$. The second component involves contrasting the influence of sub-models $\mathcal{M}_{cl}$ and $\mathcal{M}_{id}$, by decreasing logit values l^{θ}_{cl} and l^{θ}_{id} through adaptive scales (d_{cl} and d_{id}). The motivation behind this is that if the outputs of $\mathcal{M}_{id}$ is more different from $\mathcal{M}_{nl}$ (e.g., larger d_{id}), indicating a stronger contrast (denoted as $-d_{id} * l^{\theta}_{id}$), which makes sure that ALCD has the potential to mitigate the hallucinations arising from identification ability. While, we subtract $d_{cl} * l^{\theta}_{cl}$ as a compensation item when utilizing $\mathcal{M}_{cl}$. Considering the strong classification ability of $\mathcal{M}_{cl}$, our objective is to aggregate the outcomes when both the $\mathcal{M}_{cl}$ and $\mathcal{M}_{nl}$ exhibit high confidence in predicting the cls token (e.g., 'negative' in Figure 2). They should be more combined as their similarity increases (i.e., smaller d_{cl} value). This ensures that ALCD can adaptively adjust the utilization of $\mathcal{M}_{cl}$.

Conversely, when the next token is an ide token, the same rule is applicable. For the next token that do not belong to either ide or cls, we solely utilize logit output l^{θ}_{nl} of normal model. By employing this alternating contrast prediction, ALCD has the capability to modify the overall probability of tokens and then harness the abilities of sub-models.

3.2.3 Scope Constraints on Tokens

In addition, it is worth noting that certain tokens may exhibit a significant discrepancy when subjected to contrastive decoding, which makes the implausible tokens receive a high score after contrast, leading to what is referred to as the false positives (Li et al., 2023; Chuang et al., 2024). In light of this, we implement a constraint that is contingent upon the confidence level:

$$\begin{aligned} &\mathcal{V}_{head}(\boldsymbol{y}_{<t}) = \{v \in \mathcal{V} : \\ &\mathcal{P}_{\theta}(v|\boldsymbol{i}, \boldsymbol{x}, \boldsymbol{y}_{<t}) \geq \beta \max_{v} \mathcal{P}_{\theta}(v|\boldsymbol{i}, \boldsymbol{x}, \boldsymbol{y}_{<t})\}, \end{aligned} \tag{7}$$

where $\mathcal{V}$ represents the output vocabulary of LLMs, v is the token of output vocabulary, and β is a hyper-parameter used to determine the max truncation rate of low-probability tokens. Instead of employing constraints with a single model in Li et al. (2023), our approach involves combining the intersection of confidence values $\mathcal{V}^{inter}_{head}$ obtained from three models (outputs of $\mathcal{M}_{nl}$, $\mathcal{M}_{id}$, and $\mathcal{M}_{cl}$). Tokens with confidence levels below a specific threshold are assigned a negative infinity value:

$$\begin{aligned} &\mathcal{V}^{inter}_{head} = \mathcal{V}^{nl}_{head} \cap \mathcal{V}^{cl}_{head} \cap \mathcal{V}^{id}_{head}, \\ &logit_{\theta}(v|\boldsymbol{i}, \boldsymbol{x}, \boldsymbol{y}_{<t}) = -\infty, \text{ if } v \notin \mathcal{V}^{inter}_{head}(\boldsymbol{y}_{<t}). \end{aligned} \tag{8}$$

By combining token constraints to enhance and contrast predictions, our proposed ALCD is able to effectively leverage capabilities of $\mathcal{M}_{id}$ or $\mathcal{M}_{cl}$, while addressing the issue of hallucinations in $\mathcal{M}_{nl}$ that arise from other capabilities in $\mathcal{M}_{cl}$ or $\mathcal{M}_{id}$.

Dataset	#Train	#Valid	#Test
CMeEE-V2	4,600	400	400
CMeIE-V2	4,600	400	400
IMCS-V2-NER	4,600	400	400
CMedCausal	2,600	400	400
IMCS-V2-SR	4,600	400	400
CHIP-MDCFNPC	4,600	400	400

Table 1: Dataset partitioning statistics.

4 Experiments

4.1 Experimental Setup

Tasks and Datasets. We apply six MIE tasks from a Chinese medical dataset named PromptCBLUE (Zhu et al., 2023) for evaluation. **CMeEE-V2** is a task of Chinese medical entity recognition. **IMCS-V2-SR** aims to normalize the patient-doctor dialogue by medical concepts. **IMCS-V2-NER** targets extracting medical concepts from dialogues. **CMedCausal** is a task of causal relation extraction for medical texts. **CHIP-MDCFNPC** refers to clinical concept finding and discrimination. **CMeIE-V2** aims to recognize and categorize the entity relation contained in medical texts. The output forms of all tasks are built with the *identify-and-classify* pattern, as mentioned in Section 1. Due to space limitations, we leave more details about the tasks to **Appendix** A.1. Since the open-source test set was not available, we used the validation set as our test set. Subsequently, we partition the training set into a new training set and validation set and ensure the validation set contains the same number of samples as the test set. Table 1 presents the dataset partitioning statistics.

Models and Baselines. To improve the learning of data, we experimented with two widely-used

Decoding Method	CMeEE-V2	CMeIE-V2	IMCS-V2-NER	CMedCausal	IMCS-V2-SR	CHIP-MDCFNPC
			ChatGLM-6B			
Greedy Search	66.48	45.60	88.37	41.01	71.55	42.58
Beam Search	66.77	45.80	88.60	41.41	71.84	42.77
Top K Sample	63.38	39.02	88.19	39.41	69.40	38.87
Nucleus Sample	64.93	41.13	88.26	40.58	69.88	41.92
CFG (Sanchez et al., 2023)	66.95	43.84	88.76	40.61	72.06	42.49
CAD (Shi et al., 2023)	66.88	44.04	88.77	40.57	72.06	42.49
CD (Li et al., 2023)	66.34	46.03	88.54	40.72	72.40	42.33
DoLa (Chuang et al., 2024)	66.46	43.78	88.96	40.47	38.68	42.92
ALCD (Ours)	**67.44**	**47.02***	**89.64**	**42.53***	**73.57***	**43.90***
			Qwen-7B-Chat			
Greedy Search	65.49	42.87	88.65	30.10	71.28	40.61
Beam Search	66.61	43.40	89.46	30.21	71.35	40.94
Top K Sample	65.71	36.34	88.83	19.55	71.04	40.19
Nucleus Sample	66.04	33.87	89.08	25.81	70.09	39.40
CFG (Sanchez et al., 2023)	65.18	39.07	88.64	12.96	71.15	40.18
CAD (Shi et al., 2023)	66.09	36.67	88.00	14.40	71.72	39.49
CD (Li et al., 2023)	65.19	35.86	88.98	14.69	70.27	39.35
DoLa (Chuang et al., 2024)	65.16	35.51	88.49	16.52	71.29	39.37
ALCD (Ours)	**68.12***	**44.89***	**90.82***	**31.68***	**72.40***	**41.91**

Table 2: Experiment results (micro F1 score↑: higher is better) on six medical datasets with the best scores highlighted **in bold**. All baselines are based on the same fine-tuned normal model, and the model-agnostic parameters for fine-tuning and inference are kept consistent, with only the specific decoding method being changed. "$*$" indicates the statistically significant improvements (i.e., two-sided t-test with $p < 0.05$) over the best baseline.

multilingual LLMs, ChatGLM-6B v1 (Du et al., 2022) and Qwen-7B-Chat v1 (Bai et al., 2023). We compared our method for mitigating hallucinations with eight decoding baselines, which can be categorized as follows: **Deterministic decoding**: 1) greedy search decoding; 2) beam search decoding; **Stochastic decoding**: 3) Top K sample decoding; 4) nucleus sample decoding; **Contrastive decoding**: 5) CFG (Ho and Salimans, 2022); 6) CAD (Shi et al., 2023); 7) CD (Li et al., 2023); 8) DoLa (Chuang et al., 2024). For the validation of Deterministic and Stochastic methods, we utilized the implementation provided by the Huggingface toolkit (Wolf et al., 2020). However, for the contrastive decoding methods, adjustments were required when applying them to MIE tasks as they were not specifically designed to tackle the hallucination problem in MIE. For CFG, we simply use logits with normal input text and logits with the last token of input text as a comparison. For CAD, we employ both normal input text and input text without classification labels to contrast the output in different contexts. For CD, we employ the normal model as the expert model and proceed with a model using only half the number of fine-tuning steps for the amateur model. DoLa is implemented following their published paper.

Implementation Details. We conducted all experiments using four NVIDIA V100 GPUs. As we fine-tuned LLMs using LoRA, the decoding process was performed using a single GPU. All experimental results were evaluated using the Micro-F1 score following Zhu et al. (2023). All hyper-parameters of baselines are set based on the optimal values found in the validation set of the corresponding works. For ALCD, we conducted a search in the validation set to determine the appropriate values for the scale of contrasting prediction α, the maximum rate of constraint β, and the step of fine-tuning. For α, we limit the search scope to the values of [0.01, 0.1, 0.2, 0.3, 0.4, 0.5]. For β, we limit the search scope to the values of [0.4, 0.45, 0.5, 0.55, 0.6, 0.65]. The fine-tuning step of the normal model remains consistent across all baselines. We employ a batch size of 8 and perform 1,000 steps to fine-tune all datasets and LLMs, except for Qwen-7B-Chat where we use 3,000 steps in CMeIE-V2, CMedCausal, and CHIP-MDCFNPC,

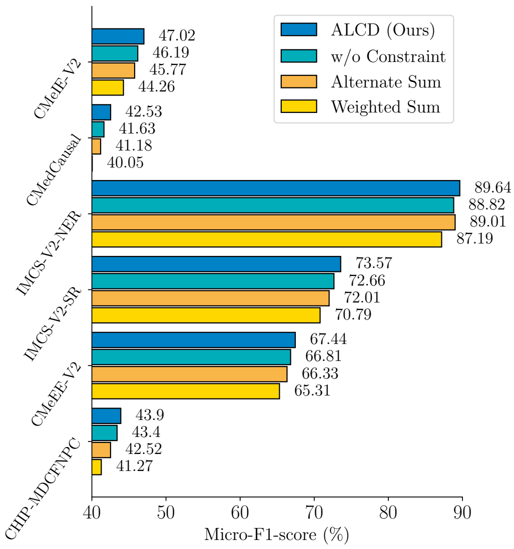


Figure 3: Ablation study on six medical datasets using ChatGLM-6B.

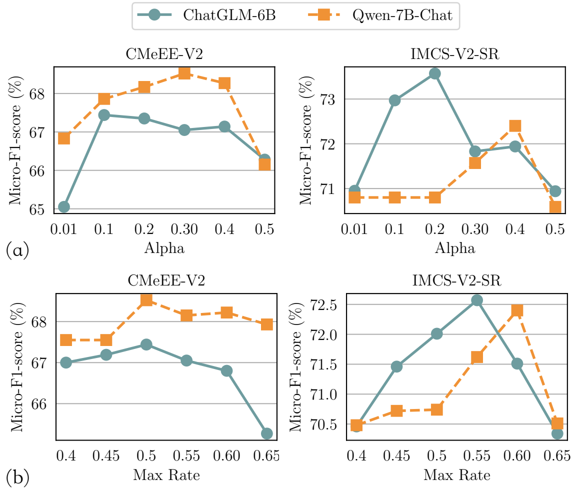


Figure 4: (a) Analysis of the scale of contrasting prediction α (in Formula 5); (b) Analysis of max rate of constraint β (in Formula 7).

due to that extra steps are required for convergence.

4.2 Main Results

In this section, we provide a comprehensive performance comparison of ALCD against other baselines on six medical datasets and two different backbone LLMs. As shown in Table 2, our proposed ALCD outperforms both contrastive decoding and non-contrastive decoding methods and the performance gap reaches the largest of 4.87% in Qwen-7B-chat on the CMedCausal dataset. Our proposed ALCD has been shown to improve performance on both ChatGLM-6B and Qwen-7B-Chat, which confirms its universality. Besides, ALCD particularly performs well on CMeEE-V2, IMCS-V2-NER, and CHIP-MDCFNPC datasets, and outperforms other baselines by a large margin. This finding aligns with our motivation as these datasets include more entity candidates, more classification labels, and thus higher difficulties for LLMs. Some contrastive decoding methods, such as DoLa, achieve much lower results on IMCS-V2-SR in the ChatGLM-6B, indicating the coupled difficulties for the medical *identify-and-classify* tasks. We find that the proposed adaptive method of DoLa predominantly selects the 2nd or 8th layer as the optimal premature layer, which suggests that DoLa's intended ability to amplify factual knowledge across different layers may not be fully aligned with the MIE tasks. We observed that the poor performance of sampling methods (Top K and Nucleus Sample) indicates that high diversity generation may not be essential for the MIE task.

4.3 Ablation Study

In this section, we analyze the effects of different components on ALCD. Specifically, we experiment with ALCD against three variants: 1) ALCD without Constraint: removing the dynamic constraints on tokens, 2) Alternate Sum: alternately summing the logits from three models instead of utilizing contrastive decoding (i.e., replacing Formula 5 with $l_{nl}^{\theta} + \alpha l_{cl}^{\theta}$, if $n_t = cls$; $l_{nl}^{\theta} + \alpha l_{id}^{\theta}$, if $n_t = ide$), 3) Weighted Sum: directly summing the logits from three models with the same weight of ALCD (i.e., replacing Formula 5 with $l_{nl}^{\theta} + \alpha(l_{cl}^{\theta} + l_{id}^{\theta})$). As depicted in Figure 3, the results confirm that incorporating token constraints enhances the performance of the normal model. Specifically, on the CMeIE-V2 dataset, the micro F1 score decreased from 47.02% to 46.19% when no constraints were utilized. Moreover, removing the alternate contrasting with either Alternate Sum or Weighted Sum resulted in performance declines, with Weighted Sum yielding the poorest overall performance. This finding highlights the effectiveness of applying alternate contrastive decoding and indicates that solely ensembling multiple LLMs for these tasks does not lead to performance improvement.

4.4 Scale of Contrasting Prediction

To investigate the effect of hyper-parameter α in Formula 5, we set different values from 0.01 to 0.5 and conduct experiments on CMeEE-V2 and IMCS-V2-SR datasets. A larger α means a larger

Dataset	Constraint in CD	Ours
CMeEE-V2	66.16	**67.44**
CMeIE-V2	46.15	**47.02**
IMCS-V2-NER	89.01	**89.64**
CMedCausal	41.88	**42.53**
IMCS-V2-SR	72.71	**73.57**
CHIP-MDCFNPC	42.72	**43.90**

Table 3: Comparison of token constraint method on all datasets using ChatGLM-6B.

scale of contrastive decoding. As shown in Figure 4(a), it can be observed that increasing the scale of contrastive decoding appropriately enhances the micro F1 score of both backbone LLMs, indicating the efficiency of our contrastive decoding method. While, excessively large values of α (e.g., exceeding 0.4), can lead to a decline in performance, which demonstrates that excessive utilization or weakening of the sub-models' ability may result in a decrease in the final effect.

4.5 Max Rate of Constraint

In this section, we examine the effect of β in Formula 7, which controls the max truncation rate of low-probability tokens for contrastive decoding. The results are shown in Figure 4(b). We observed that small β values (e.g., smaller than 0.45) have a minimal impact on the low-probability tokens, suggesting that these tokens are unlikely to significantly influence the model. We also found that the performance reaches its peak at around 0.5 and subsequently decreases with a further increase in β. This finding aligns with our analysis, as larger values of β tend to remove more false positive tokens. However, excessively large values of β can also result in the removal of true positive tokens, thereby reducing overall performance.

4.6 Comparison of Token Constraint

To further validate the effectiveness of our proposed constraint method for avoiding noisy tokens in contrastive decoding, we compare against the constraint method of CD. Specifically, we replace the token constraint related to scale and range in ALCD with a constraint employed in CD, while maintaining the alternative contrastive decoding technique unchanged. As shown in Table 3, our method consistently outperforms the 'constraint in CD' approach across all datasets. We attribute this improvement to the successful implementation of alternating adaptive token constraints on both scale and scope in our ALCD, whereas CD relies solely on a maximum value judgment.

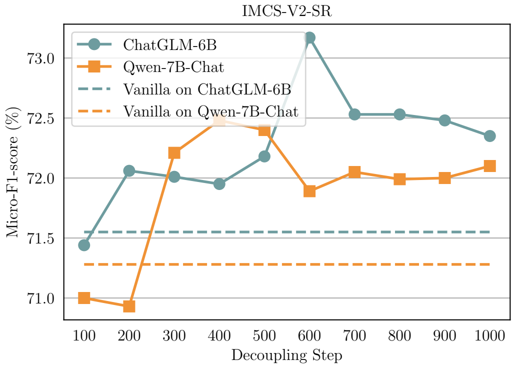


Figure 5: Analysis of varying decoupling steps during fine-tuning on IMCS-V2-SR dataset. 'Vanilla' refers to the performance of normal model using greedy search.

4.7 Affect of Decoupling Steps

To investigate how the capabilities of sub-models affect overall performance of ALCD, we conducted experiments by individually fine-tuning two subtask LLMs (i.e., $\mathcal{M}_{id}$ and $\mathcal{M}_{cl}$) with varying steps while keeping normal model (i.e., $\mathcal{M}_{nl}$) unchanged. As illustrated in Figure 5, we observed that fine-tuning on sub-models effectively enhances performance, resulting in higher micro F1 scores compared to vanilla ones with 300 steps or larger. When the number of fine-tuning steps increases, the performance rises for both LLMs, while decreases after 600 steps for ChatGLM-6B and 400 steps for Qwen-7B-Chat, respectively. We believe the reason is that excessive fine-tuning steps can potentially improve the identification capabilities of $\mathcal{M}_{cl}$ and the classification capabilities of $\mathcal{M}_{id}$, consequently compromising the desired decoupling effect between the two abilities. As a result, contrasting the predictions in ALCD fails to improve performance.

5 Conclusion

In this paper, we propose ALCD to address hallucinations of LLMs in MIE tasks. ALCD utilizes decoupled fine-tuning process to separately learn LLM's identification and classification abilities. During inference, ALCD alternately enhances these abilities while excluding other capabilities that may result in hallucinations. We also introduce adaptive scales based on distribution similarities to enable the flexible use of identification or classification abilities. Extensive experiments conducted on two backbones have demonstrated substantial enhancement achieved by ALCD in MIE tasks.

6 Limitation

Our approach aims to decouple the identification and classification abilities of LLMs in the medical information extraction tasks and leverage their respective capabilities through alternate contrastive decoding. However, this strategy leads to an increase in both fine-tuning and inference costs. In this paper, ALCD switches between identification or classification capabilities based on simple rule-based judgment, but it is worth exploring more automatic and flexible judgment methods in future work. Furthermore, we have only investigated the effectiveness of our approach in medical information extraction tasks, and expanding our ALCD framework to other medical tasks, other domains, and other language settings is an avenue for future exploration. Exploring more robust decoupling methods and contrasting decoding techniques are also potential future research directions.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. *arXiv preprint arXiv:2309.16609*.

Junyi Bian, Jiaxuan Zheng, Yuyi Zhang, and Shanfeng Zhu. 2023. Inspire the large language model by external knowledge on biomedical named entity recognition. *arXiv preprint arXiv:2309.12278*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. *Preprint*, arXiv:2309.03883.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General Language Model Pretraining with Autoregressive Blank Infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Lei Gao, Xinnan Zhang, Xian Wu, Shen Ge, and Yefeng Zheng. 2023. Dialogue Medical Information Extraction with Medical-Item Graph and Dialogue-Status Enriched Representation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13311–13321, Singapore. Association for Computational Linguistics.

Tongfeng Guan, Hongying Zan, Xiabing Zhou, Hongfei Xu, and Kunli Zhang. 2020. Cmeie: construction and evaluation of chinese medical information extraction dataset. In *Natural Language Processing and Chinese Computing: 9th CCF International Conference, NLPCC 2020, Zhengzhou, China, October 14–18, 2020, Proceedings, Part I 9*, pages 270–282. Springer.

Udo Hahn and Michel Oleynik. 2020. Medical information extraction in the age of deep learning. *Yearbook of medical informatics*, 29(01):208–220.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. MedAlpaca – An Open-Source Collection of Medical Conversational AI Models and Training Data. *Preprint*, arXiv:2304.08247.

Jonathan Ho and Tim Salimans. 2022. Classifier-Free Diffusion Guidance. *Preprint*, arXiv:2207.12598.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *Preprint*, arXiv:2106.09685.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023a. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 55(12):1–38.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023b. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2023. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. *Preprint*, arXiv:2311.16922.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive Decoding: Open-ended Text Generation as Optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

F. Liu, X. Wu, S. Ge, W. Fan, and Y. Zou. 2021. Exploring and Distilling Posterior and Prior Knowledge for Radiology Report Generation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13748–13757. IEEE Computer Society.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. 2024. When MOE Meets LLMs: Parameter Efficient Fine-tuning for Multi-task Medical Applications. *Preprint*, arXiv:2310.18339.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified Structure Generation for Universal Information Extraction. *arXiv preprint arXiv:2203.12277*.

OpenAI. 2023. GPT-4 Technical Report. *ArXiv*, abs/2303.08774.

Guillaume Sanchez, Honglu Fan, Alexander Spangher, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. 2023. Stay on topic with Classifier-Free Guidance. *Preprint*, arXiv:2306.17806.

Prabin Sharma, Kisan Thapa, Dikshya Thapa, Prastab Dhakal, Mala Deep Upadhaya, Santosh Adhikari, and Salik Ram Khanal. 2023. Performance of ChatGPT on USMLE: Unlocking the Potential of Large Language Models for AI-Assisted Medical Education. *Preprint*, arXiv:2307.00112.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Scott Wen tau Yih. 2023. Trusting Your Evidence: Hallucinate Less with Context-aware Decoding. *Preprint*, arXiv:2305.14739.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large Language Models Encode Clinical Knowledge. *Preprint*, arXiv:2212.13138.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.

Guangyu Wang, Guoxing Yang, Zongxin Du, Longjun Fan, and Xiaohu Li. 2023a. ClinicalGPT: Large Language Models Finetuned with Diverse Medical Data and Comprehensive Evaluation. *Preprint*, arXiv:2306.09968.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023b. Instructuie: multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jiageng Wu, Xian Wu, Yefeng Zheng, and Jie Yang. 2024. MedKP: Medical Dialogue with Knowledge Enhancement and Clinical Pathway Encoding. *Preprint*, arXiv:2403.06611.

Xuehong Wu, Junwen Duan, Yi Pan, and Min Li. 2023. Medical knowledge graph: Data sources, construction, reasoning, and applications. *Big Data Mining and Analytics*, 6(2):201–217.

Derong Xu, Ziheng Zhang, Zhenxi Lin, Xian Wu, Zhihong Zhu, Tong Xu, Xiangyu Zhao, Yefeng Zheng, and Enhong Chen. 2024. Multi-perspective Improvement of Knowledge Graph Completion with Large Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11956–11968.

Xin Yu, Wenshen Hu, Sha Lu, Xiaoyan Sun, and Zhenming Yuan. 2019. Biobert based named entity recognition in electronic medical record. In *2019 10th International Conference on Information Technology in Medicine and Education (ITME)*, pages 49–52.

Wei Zhu, Xiaoling Wang, Huanran Zheng, Mosha Chen, and Buzhou Tang. 2023. PromptCBLUE: A Chinese Prompt Tuning Benchmark for the Medical Domain. *arXiv preprint arXiv:2310.14151*.

A Appendix

A.1 Tasks and Datasets

In the experiments, we adopt a Chinese medical dataset, named PromptCBLUE (Zhu et al., 2023), including several common tasks. Due to limited resources, we select 6 tasks for validation. The statistics are in Table 1, and the dataset details are listed as follows:

- **CMeEE-V2**. Chinese medical name entity recognition. We consider "extracting entities from medical texts" as *identify* and "categorizing the entities" as *classify*.
- **CMeIE-V2**. Chinese medical entity relation extraction. We consider "recognizing the head and tail entities from medical texts" as *identify* and "categorizing the relation types between entities".
- **IMCS-V2-NER**. Medical entity recognition from the doctor-patient dialogue. We consider "identifying the medical entities from dialogues" as *identify* and "classifying the medical entity types" as *classify*.
- **CMedCausal**. Causal relation extraction for medical texts. We consider "recognizing the causal and effect words from medical texts" as *identify* and "categorizing the causal relation" as *classify*.
- **IMCS-V2-SR**. Medical normalization of the doctor-patient dialogue. We consider "extracting the normalized words from dialogues" as *identify* and "imputing the normalization labels" as *classify*.
- **CHIP-MDCFNPC**. Clinical concept finding and discrimination for the clinical report. We consider "extracting the clinical concepts from reports" as *identify* and "classifying the derived clinical concepts" as *classify*.