Prompt-based Generation of Natural Language Explanations of Synthetic Lethality for Cancer Drug Discovery

Ke Zhang[1,2,3], **Yimiao Feng**[1,4], **Jie Zheng**[1,5*]

[1] School of Information Science and Technology, ShanghaiTech University, Shanghai, China

[2]Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai, China

[3]University of Chinese Academy of Sciences, Beijing, China

[4]Lingang Laboratory, Shanghai, China

[5]Shanghai Engineering Research Center of Intelligent Vision and Imaging, ShanghaiTech University, Shanghai, China

{zhangke1, fengym, zhengjie}@shanghaitech.edu.cn

Abstract

Synthetic lethality (SL) offers a promising approach for targeted anti-cancer therapy. Deeply understanding SL gene pair mechanisms is vital for anti-cancer drug discovery. However, current wet-lab and machine learning-based SL prediction methods lack user-friendly and quantitatively evaluable explanations. To address these problems, we propose a prompt-based pipeline for generating natural language explanations. We first construct a natural language dataset named NexLeth. This dataset is derived from New Bing through prompt-based queries and expert annotations and contains 707 instances. NexLeth enhances the understanding of SL mechanisms and it is a benchmark for evaluating SL explanation methods. For the task of natural language generation for SL explanations, we combine subgraph explanations from an SL knowledge graph (KG) with instructions to construct novel personalized prompts, so as to inject the domain knowledge into the generation process. We then leverage the prompts to fine-tune pre-trained biomedical language models on our dataset. Experimental results show that the fine-tuned model equipped with designed prompts performs better than existing biomedical language models in terms of text quality and explainability, suggesting the potential of our dataset and the fine-tuned model for generating understandable and reliable explanations of SL mechanisms.

Keywords: Explainability, Text Mining, Pre-trained Language Model, Knowledge Graph

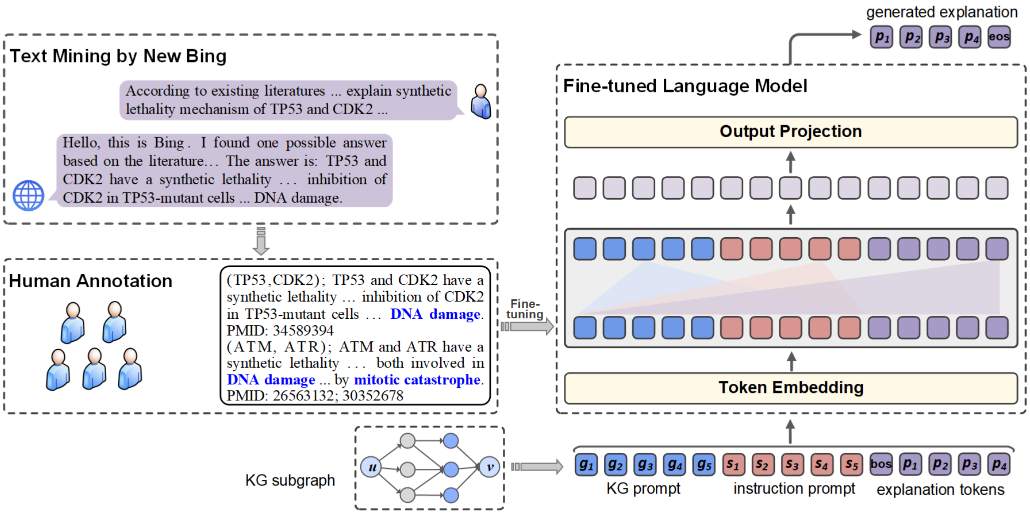
1. Introduction

Synthetic lethality (SL) is a genetic interaction where a single gene mutation allows cell survival, but simultaneous mutations in two genes lead to cell death (Kaelin, 2005). Targeting an SL partner gene can selectively kill cancer cells with a specific mutation, offering a potential treatment for undruggable mutant genes (Jariyal et al., 2020). While many SL gene pairs have been discovered through biological screening techniques (Huang et al., 2020) and computational methods (Wang et al., 2022b), their clinical applications are limited due to unclear mechanisms. Hence, understanding these mechanisms is crucial for developing anticancer drug targets.

Previous computational studies explaining SL mechanisms can be categorized into statistical methods and machine learning methods (Wang et al., 2022b). Statistical methods are usually based on some hypotheses and require considerable prior knowledge (Lee et al., 2018; Magen et al., 2019; Lee et al., 2021), while machine learning methods mainly rely on knowledge graphs (KGs) (Wang et al., 2021b; Liu et al., 2022; Zhang et al., 2023). KGs provide explicit and accurate knowledge, enabling the generation of explanations through graph-based reasoning (Pan et al., 2023). However, when semantic information about SL is sparse on the knowledge graph, interpretations of the SL mechanisms are limited. Furthermore, graph-based explanations are typically in the form of paths or subgraph structures, but such explanations may not be intuitive enough for downstream users. In contrast, natural language models inherently contain rich domain knowledge, as they are pre-trained on a large number of biomedical texts, and natural language explanations are more user-friendly (Wei et al., 2023; Wang et al., 2021a). Therefore, we aim to combine text-based knowledge and KGs to generate natural language explanations for SL mechanisms. Herein, we expect that, for any SL gene pairs, the natural language explanation for the SL mechanism must involve the functional connections of the two genes and how these connections lead to cancer cell death.

The task of natural language generation is common in many domains, such as explainable recommendation and question answering (QA) (Liu et al., 2023a; Luo et al., 2022). A basic requirement for this task is the availability of natural language

[*] Corresponding authors



According to existing literatures ... explain synthetic lethality mechanism of TP53 and CDK2 ...

Hello, this is Bing. I found one possible answer based on the literature… The answer is: TP53 and CDK2 have a synthetic lethality … inhibition of CDK2 in TP53-mutant cells ... DNA damage.

(TP53,CDK2); TP53 and CDK2 have a synthetic lethality … inhibition of CDK2 in TP53-mutant cells … **DNA damage**.
PMID: 34589394
(ATM, ATR); ATM and ATR have a synthetic lethality … both involved in **DNA damage** ... by **mitotic catastrophe**.
PMID: 26563132; 30352678

Figure 1: Pipeline of explanation collection by New Bing and explanation generation by a fine-tuned language model.

datasets. For instance, in recommendation systems, we can naturally access certain user review datasets, serving as explanation texts (Yan et al., 2023). For biomedical QA, abstracts of published articles can be easily collected for building the QA dataset (Jin et al., 2019). Based on these datasets, language models can be trained to generate natural language explanations or answers (Li et al., 2023, 2021). However, for the task of training large language models to generate natural language SL explanations, there is no available dataset that explains SL mechanisms systematically, since the causation of SL-based cancer could be complex and diverse, and it requires deep domain knowledge to summarize the mechanisms. Therefore, it is challenging to acquire high-quality explanatory datasets.

Pre-trained language models (PLMs) have been successfully applied to many downstream Natural Language Processing (NLP) tasks (Wei et al., 2023). The downstream tasks mainly benefit from the learned representations. Specifically, through pre-training Transformer's decoder modules, Generative Pre-trained Transformer (GPT) exhibits superior performance on language generation tasks, such as machine translation and text summarizations (Li and Liang, 2021). Recently, general-purpose pre-trained large language models based on GPT-3 (Brown et al., 2020) and GPT-4 (OpenAI, 2023) have shown excellent performance across various domains, especially ChatGPT[1] and New Bing[2]. As New Bing incorporates search capabilities, it is capable to provide users with both answers and related sources. Users can assess the reliability of the answers based on the provided search sources. Therefore, we can harness these innovative generation tools for text mining.

Recently, there have been growing interests in applying pre-trained large language models in biomedical domain (e.g., MedGPT (Kraljevic et al., 2021), BioGPT (Luo et al., 2022) and Med-PaLM (Singhal et al., 2023)). However, since fine-tuning on large pre-trained models requires sufficient instances, it is challenging to directly conduct fine-tuning in few-shot learning. Prompt learning has recently emerged as a new research direction (Liu et al., 2023b). Using task-specific prompts, prompt learning can imbue pre-trained models with a wealth of knowledge that aligns with a specific task, thereby enabling the models to better adapt to specific domain tasks with a few data. Prompts can be categorized as continuous prompts (i.e., learnable embeddings) and discrete prompts (i.e., sequences of words) (Liu et al., 2023b; Wu et al., 2023b). The discrete prompts can be either fixed for all instances or personalized. For the task of explanation generation, personalization is very important to reflect the characteristics of specific input instances, thus we consider designing prompt templates and then generating personalized prompts for different SL gene pairs.

In this study, we propose a prompt-based framework to generate natural language explanations for SL mechanisms. Our main contributions contain a natural language explanation dataset named NexLeth[3] and fine-tuning of pre-trained biomedical language models for new explanation generation. This is the first study that explains SL mechanisms in the form of natural language. The overall pipeline is depicted in Figure 1. We first design prompts and employ New Bing to automatically mine literature on SL mechanisms and summarize answers into com-

[1] https://chat.openai.com/
[2] https://www.bing.com/new

[3] https://github.com/JieZheng-ShanghaiTech/NexLeth

Algorithm 1: Text mining by using New Bing

```
Input: gene pairs $Q_{SL}$, template QueryPrompt
GenepairsCollection $Q_{nb} \leftarrow \{\}$
AnswerCollection $M_{nb} \leftarrow \{\}$
CitationCollection $R_{nb} \leftarrow \{\}$
$Q_{query} \leftarrow Q_{SL}$
$Q_{fail} \leftarrow \{\}$
for $i \leftarrow 1$ to 5 do
    for $j \leftarrow 1$ to $len(Q_{query})$ do
        gene pair $(u, v) = Q_{SL}[j]$
        (answer,citation)$\leftarrow$New
          Bing(QueryPrompt$(u, v)$)
        if FailureAnswersCheck(answer) then
            if KeyWordsCheck(answer) then
                $Q_{nb} \leftarrow$ (u,v)
                $M_{nb} \leftarrow$ answer
                $R_{nb} \leftarrow$ citation
            else
                $Q_{fail} \leftarrow$ (u,v)
            end
        else
            $Q_{fail} \leftarrow$ (u,v)
        end
    end
    $Q_{query} \leftarrow Q_{fail}$
end
```

prehensible explanations. The curated dataset contains 707 explanations, corresponding key features and literature. NexLeth aids biologists in delving deeper into synthetic lethality and offers a standard for evaluating the explainability of existing explanation methods. Leveraging this dataset, we then fine tune PLMs to generate natural language for SL mechanisms. To facilitate a better understanding of domain context and task, we transform KG subgraphs into natural language prompts in a rule-based manner and combine them with instruction prompts to construct novel personalized prompts. Experiments show that the fine-tuned model employing our designed prompts outperforms existing baseline models in terms of both text quality and answer explainability.

2. Dataset

The first goal of our task is to construct a natural language dataset for SL explanations. Therefore, in this section, we first introduce the procedure of mining the natural language explanations for SL mechanisms, and then we present statistical analyses of the built dataset.

2.1. Data collection

Text mining using New Bing Using New Bing as a text mining tool based on GPT-4, our objective is to explore literature and condense SL mechanisms into comprehensible explanations for specific gene pairs. SynLethDB (Wang et al., 2022a) is a database containing numerous SL gene pairs with their associated literature. We selected 1,556 SL gene pairs from this database as the subjects for our explanations. Chosen based on either biological experiments or text mining, these gene pairs are supported by studies offering comprehensive insight into their genetic interactions, making them ideal candidates for clear SL mechanism summarization. For New Bing, we crafted a distinct prompt template to guide its queries. In a single-turn dialogue, New Bing was tasked with locating and condensing the SL mechanism of a gene pair based on available literature. This text mining approach is detailed in Algorithm 1.

We then designed two post-processing steps to automatically filter the results. The first step excludes gene pairs failing to return query results, and the second step removes answers with insufficient explainability, since some sentences mention gene functions but do not describe their role in cell death. To do this, we set an SL keyword collection (e.g. "*cell* growth", "*essential for*", etc.). For answers that do not contain any keywords from this collection, we remove the answers and corresponding gene pairs. To ensure the completeness of our dataset, we re-attempted queries and post-processing for the gene pairs associated with failed responses. This is because network conditions and chat sessions can occasionally result in unsuccessful responses. By repeating this step, we aimed to obtain the most comprehensive set of answers possible. After five iterations, we obtained the final data collection denoted as M_{nb}, comprising 683 qualified explanatory answers. Each answer in M_{nb} has several corresponding citation literatures, denoted as R_{nb}.

Human annotations Using dataset collected via New Bing, we further annotated the extracted information in terms of explainability. Our objective was to validate the accuracy of generated sentences and the consistency of the SL mechanisms with their source literature. Drawing inspiration from evaluation methods for explainable recommendation (Li et al., 2023), we marked key phrases related to SL mechanisms within the answers. These annotations help assess a language model's explanatory capability. We recruited five experts with specialization in cancer and SL to annotate the 683 answers in the dataset. The annotation procedures are shown in the Appendix A.1. Further, these experts summarized explanatory sentences for newly identified SL pairs from the literature. The resulting dataset, NexLeth, comprises 707 gene pairs and their explanations. These explanations are divided into two categories: M_{fact} and M_{hypo}. M_{fact} are derived from existing studies, detailing validated SL

Table 1: Statistics of our dataset

Statistics		M_{fact}	M_{hypo}
# answers (gene pairs)		145	562
maximum length of answers		110	94
minimum length of answers		24	18
average length of answers		47	53
sources of gene pairs in SynLethDB	individual reports	131	137
	large-scale screening	14	425

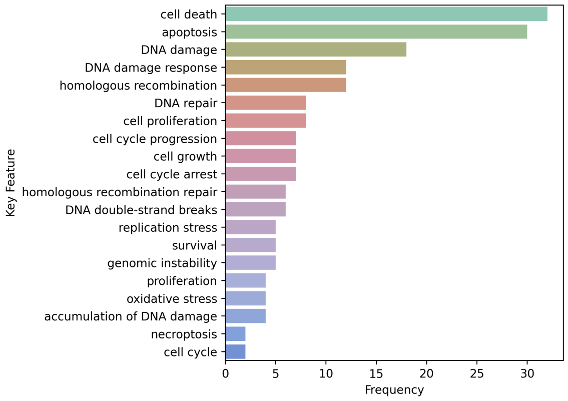


Figure 2: Distribution of the number of distinct key features in M_{fact}, ranked by the frequency of occurrence in all explanations.

mechanisms and annotated with key features and citations. In contrast, M_{hypo} serves as hypothetical explanations and provides possible evidence for further wet-lab validation.

2.2. Dataset statistics

Table 1 shows an overview of our dataset's statistics. Specifically, according to the gene pair sources in SynLethDB, we categorize the sentences within M_{fact} and M_{hypo} into two types, i.e., large-scale screening and individual reports. It can be seen that most gene pairs correspond to M_{fact} are reported individually, making their SL mechanisms easier to summarize. Some gene pairs are discovered through large-scale screening, and further literature mining uncovers in-depth studies about their mechanisms. In contrast, most gene pairs in M_{hypo} are identified from large-scale screenings, indicating limited further research discussing their SL mechanisms. For the remaining gene pairs in M_{hypo}, we attempted to use ChatGPT or New Bing to summarize their mechanisms again according to the source reports, but the Chatbots failed to give satisfactory answers. We noted that most individual reports did not explicitly describe SL mechanisms in the abstracts or offered a few assumptions, and some did not mention the names of gene pairs in abstracts. Therefore, it is challenging to extract clear and comprehensive SL mechanisms from these reports.

Next, we visualized the distribution of the number of distinct key features within M_{fact}, as shown in Fig. 2. We can see that most explanatory answers include phrases such as "*cell death*" and "*apoptosis*", which are closely related to synthetic lethality. In addition, DNA-related features are also frequently mentioned such as "*DNA repair*" and "*DNA Damage*", reflecting that DNA instability may be a common factor leading to cell death.

3. Methods

The second goal of our task is to generate a natural language explanation $P_{u,v}$ for SL gene pairs (u, v), so as to explain the biological mechanisms behind the SL interaction. The SL gene pair can be provided by an SL prediction model. During both training and testing stages, given only the names of two genes that are predicted to have the SL relationship, our trained model will generate explanations for them. Therefore, our method is compatible with any SL predictive model. In this section, we first introduce an approach to extending the training dataset, then we present a novel prompt template and the prompt-based fine-tuning and inference for explanation generation.

There are many new fine-tuning paradigms equipped with prompt learning, including fixed-prompt fine-tuning models, learning prompt-fixed models, and learning both prompts and fine-tuning models (Liu et al., 2023a). Here, we choose the first paradigm, i.e. fixing prompts and fine-tuning pre-trained biomedical language models.

3.1. Data augmentation via ChatGPT

In our dataset, the explanations in M_{fact} have been validated by existing literature, thus they can be used for both training and inference. However, the sentences in M_{fact} may be insufficient to effectively fine-tune a generative language model, therefore we augment M_{fact} by rephrasing the sentences with ChatGPT (Dai et al., 2023).

We first design a prompt template and ask ChatGPT to rephrase the explanations in M_{fact} for several times. The template is shown in Appendix A.2. Suppose $s \in \mathrm{M}_{\text{fact}}$ is an original explanation, and $\mathcal{R}_s = \{r_0, ..., r_N\}$ represents a collection of all distinct rephrased sentences for s. To ensure the consistency between the original and rephrased sentences, we apply the cosine similarity scores as a measurement of similarity, which is commonly used in NLP tasks:

$$\cos(s, r_i) = \frac{f_s \cdot f_{r_i}}{\|f_s\|_2 \|f_{r_i}\|_2}, \quad (1)$$

where f_s and f_{r_i} are embedding vectors for s and r_i respectively. Sentences are deemed to be qualified if the similarity score is greater than 0.5.

The qualified sentences after rephrasing are taken as new instances and added into $\mathrm{M}_{\mathrm{fact}}$. Moreover, although $\mathrm{M}_{\mathrm{hypo}}$ explanations lack empirical validation, they have been deemed reasonable through rigorous annotation, hence they can also be utilized as part of the training instances.

3.2. Discrete prompt

PLMs such as BERT-liked models (Devlin et al., 2019) and GPT-liked models (Radford et al., 2018) are typically trained on large-scale general-purpose corpora, which may lack biomedical domain knowledge. Although pre-trained biomedical language models (e.g., BioBERT (Lee et al., 2020) and BioGPT (Luo et al., 2022)) contain more comprehensive domain knowledge, their ability to transfer to specific tasks may be inadequate. To better adapt the pre-trained models to our task, we first design an instruction prompt shown in Table 2. The prompt template explicitly tells the model our task, i.e., generating explanatory text for synthetic lethal mechanisms. Given an SL gene pair (u, v), the template is filled with the names of the two genes, thereby forming a personalized prompt for (u, v). We denote the word sequence of this personalized instruction prompt as $S_{u,v}$ and the token representation of the sequence as $\mathbf{S}_{u,v} = [\mathbf{s}^1_{u,v}, ..., \mathbf{s}^n_{u,v}]$, where n is the number of tokens in $S_{u,v}$. The token representation $\mathbf{s}^i_{u,v}$ is calculated as

$$\mathbf{s}^i_{u,v} = \mathbf{W}_t \mathbf{T}^i_{u,v}, \quad (2)$$

where $\mathbf{W}_t \in \mathcal{R}^{d \times |\mathcal{V}|}$ is the token embedding matrix, d is the dimension of latent space and $|\mathcal{V}|$ is the vocabulary size. $\mathbf{T}^i_{u,v}$ represents the i-th token index vector.

Table 2: Two types of prompt templates

instruction prompt	explain the synthetic lethality mechanism between $\underline{u}$ and $\underline{v}$:
KG prompt	$\underline{u}$ and $\underline{v}$ may share common functions, including function 1, function 2.

Different from plain texts, KGs contain rich structured semantic information. Therefore, fusing domain-specific KGs and language models can assist in learning graph tasks or text tasks (Ju et al., 2022; Pan et al., 2023). To enrich the knowledge contained in the prompt and further characterize different gene pairs' prompts, we introduce a KG-based prompt template. The prompt template is designed based on an SL predictive model named KR4SL (Zhang et al., 2023), which constructs an SL KG and conducts knowledge reasoning on the KG for predicting SL gene pairs. Meanwhile, the model provides a subgraph structure on the KG for each predicted gene pair and takes the subgraph as an explanation for the prediction. A subgraph consists of several paths, with all paths starting from one gene in a pair and ending at the other. For explanatory subgraphs of most gene pairs, the paths within subgraphs follow the same schema, i.e.,

$$\text{gene } u \longrightarrow \mathcal{K} \longrightarrow \text{gene function 1} \longrightarrow \text{gene } v, \quad (3)$$

where $\mathcal{K}$ is a set of SL partner genes of gene u and $v \notin \mathcal{K}$. This schema can be cast as a rule among most subgraphs.

Based on the rule, we assume that both gene u and gene v might be involved in gene function 1. As such, among the nodes representing gene functions in an explanation subgraph, we randomly sample k nodes and convert our assumption into a prompt template shown in Table 2. An example can be found in Appendix A.2.

For gene pair (u, v), the KG prompt template is filled with gene names and the sampled nodes and becomes a personalized KG prompt for this gene pair. We thus denote the word sequence of the personalized KG prompt as $G_{u,v}$ and the token representation as $\mathbf{G}_{u,v} = [\mathbf{g}^1_{u,v}, ..., \mathbf{g}^m_{u,v}]$, where m is the number of tokens of $G_{u,v}$. Then we concatenate the KG prompt and the instruction prompt as a new prompt $\mathbf{D}_{u,v} = [\mathbf{g}^1_{u,v}, ..., \mathbf{g}^m_{u,v}, \mathbf{s}^1_{u,v}, ..., \mathbf{s}^n_{u,v}]$.

3.3. Prompt-based fine-tuning of a PLM

We take a pre-trained GPT model as our backbone. Suppose $E_{u,v}$ is the explanation word sequence for gene (u, v), and $\mathbf{E}_{u,v} = \mathbf{e}^1_{u,v}, ..., \mathbf{e}^L_{u,v}$ is the token representation of $E_{u,v}$. During the training stage, we concatenate the representations of prompt tokens and explanation tokens as the input, i.e. $\mathbf{H}^0_{u,v} = [\mathbf{d}^1_{u,v}, ..., \mathbf{d}^{n+m}_{u,v}, \mathbf{e}^1_{u,v}, ..., \mathbf{e}^L_{u,v}]$. After passing through all the hidden layers of the pre-trained model, the hidden representation is finally mapped into a vocabulary space by an output projection layer. For instance, in the predicted token sequence, the probabilities of all possible tokens at the j-th position are calculated as

$$p^j_{u,v} = \sigma(\mathbf{W}_o \mathbf{h}^j_{u,v} + \mathbf{b}_o), \quad (4)$$

where $\mathbf{W}_o \in \mathcal{R}^{|\mathcal{V}| \times d}$ and $\mathbf{b}_o$ are weights. $\sigma(\cdot)$ is softmax function. $\mathbf{h}^j_{u,v}$ is the hidden representation at the j-th position.

We fine-tune the token representation layer and the output projection layer, i.e. $\mathbf{W}_t$, $\mathbf{W}_o$ and $\mathbf{b}_o$ are learnable parameters during training. We adopt negative log-likelihood (NLL) as the loss function.

$$\mathcal{L} = \frac{1}{|\mathcal{C}|} \sum_{(u,v)\in\mathcal{C}} \frac{1}{L_{u,v}} \sum_{j=1}^{L_{u,v}} -\log p_{u,v}^{|D_{u,v}|+j} . \quad (5)$$

Here $\mathcal{C}$ is the set of gene pairs for training, $L_{u,v}$ is the number of explanation tokens for (u,v), and $p_{u,v}^{|D_{u,v}|+j}$ is the probability of ground truth token at position $|D_{u,v}|+j$. Note that we take an offset with the length of prompt tokens $|D_{u,v}|$, since we only calculate the loss on the explanation tokens at the end of sequence.

3.4. Text generation

During the inference stage, given a test gene pair, the prompt is denoted as D of length $|D|$, and the model generates a word sequence P^* with the maximum log-likelihood:

$$P^* = \arg\max_{P\in\hat{\varepsilon}} \sum_{j}^{L} \log p^{|D|+j}, \quad (6)$$

where $\hat{\varepsilon}$ is the set of all possible generated sequences, and L is the length of the output sequence. We adopt the nucleus (top-p) sampling strategy to decide each predicted token (Holtzman et al., 2020). Specifically, at each generation step, we choose the smallest possible set of words as the set of prediction words, whose cumulative probability exceeds a specified probability, then the probabilities are redistributed among the set of words. For each predicted word, we append it to the input sequence to form a new sequence, and the new input is fed into the model for the next generation. Such process is repeated until the model generates end-of-sequence token $< eos >$.

4. Experiments

4.1. Evaluation settings

Benchmark models We assess the performance of the following four biomedical pre-trained models on our dataset, all of which are pre-trained using auto-regressive approach:

- BioGPT is initialized using GPT-2$_{\text{Medium}}$ with 347 million parameters (Luo et al., 2022), pre-trained on a large collection of PubMed abstracts.
- BioGPT-Large is an upscale model of BioGPT, it leverages GPT-2$_{\text{XL}}$ as the backbone and has 1.5 billion parameters (Radford et al., 2019).
- BioMedLM is pre-trained on the PubMed abstract dataset using GPT-Neo, it contains 2.7 billion parameters (Venigalla et al., 2022).
- PMC-LLaMA, taking LLaMA as the original pre-trained model, is fine-tuned on the PubMedQA and MedMCQA datasets and tested on the USMLE dataset, the model has 7 billion parameters (Wu et al., 2023a).

Evaluation scenarios We test and fine-tune the above models under two scenarios:

- Zero-shot inference: Models are only used for inference without training. We test all the models in this scenario.
- Parameter-efficient fine-tuning: LoRA (Low-Rank Adaptation) (Hu et al., 2022) enables efficient fine-tuning of large pre-trained models without learning all parameters, which greatly saves the computational cost. We utilize PEFT LoRA (Mangrulkar et al., 2022) to fine-tune token embedding and output projection layer of a model.

Metrics For text generation tasks, the evaluation metrics usually measure the text quality, i.e., the relevance between generated texts and ground truth texts based on n-gram. Here we choose two typical metrics, i.e., BLEU (n=1,4) (Papineni et al., 2002) and ROUGE (n=1,2) (Sennrich et al., 2016). Higher text quality scores mean that the generated explanations are more similar to the ground truth explanations. However, the two metrics may not reflect real explainability. For instance, the models may generate many repeated sentences for different gene pairs, resulting in high text similarity scores. However, if the key phrases associated with SL mechanisms specific to gene pairs are omitted, it can lead to low explainability. Therefore, to evaluate such explainability of generated texts, we adopt three metrics following Li et al. (2023):

- USR (Unique sentence ratio) calculates the ratio of unique sentences among all generated sentences.
- FCR (Feature coverage ratio) calculates the ratio of distinct features among all generated sentences:

  $$\text{FCR} = \frac{N_f}{|\mathcal{F}|} \quad (7)$$

  where N_f is the number of unique features of all generated sentences and $|\mathcal{F}|$ is the number of unique features of all ground truth explanations.
- FMR (Feature matching ratio) in Li et al. (2023) is denoted as whether the generated explanations include the ground truth features. Since each ground truth explanation in our dataset may be annotated with more than one feature, we modify the metric as:

  $$\text{FMR} = \frac{1}{|\mathcal{C}_{test}|} \sum_{(u,v)\in\mathcal{C}_{test}} \frac{\sum_{f_{u,v}\in\mathcal{F}_{u,v}} \mathbb{I}(f_{u,v} \in P_{u,v})}{|\mathcal{F}_{u,v}|} \quad (8)$$

  where $\mathcal{C}_{test}$ contains all test gene pairs, $\mathcal{F}_{u,v}$ is a set of features for test gene pair (u,v) and $P_{u,v}$ is the predicted explanation. $\mathbb{I}(x) = 1$ when x is true, otherwise $\mathbb{I}(x) = 0$.

Table 3: Performance comparison of zero-shot inference and fine-tuning. B-n, R-nF, R-nR, R-nP are short for BLEU-n, ROUGE-nF1, ROUGE-nRecall, ROUGE-nPresision respectively, where n is used for n-grams. The best performance in each column is bolded, and the second-best performance is underlined.

Method	Explainability			Text quality							
	FCR	FMR	USR	B-1	B-4	R-1F	R-1R	R-1P	R-2F	R-2R	R-2P
zero-shot inference											
BioGPT-D1	0.11	0.01	1.0	3.58	0.08	14.21	10.68	26.16	1.39	1	2.82
BioGPT-D2	0.16	0.02	1.0	6.94	0.45	16.98	13.05	28.76	2.83	2.12	5.36
BioGPT-Large-D1	0.23	0.05	1.0	15.55	0.92	20.45	26.83	17.25	2.97	4.33	2.3
BioGPT-Large-D2	0.21	0.05	1.0	16.21	1.27	20.75	26.99	17.55	3.6	5.23	2.88
BioMedLM-D1	0.19	0.04	1.0	17.71	1.2	21.23	23.57	20.91	3.67	4.44	3.43
BioMedLM-D2	0.18	0.07	1.0	18.99	2.02	20.79	22.03	21.79	4.6	5.33	4.52
PMC-LLaMA-D1	0.20	0.06	1.0	19.37	1.49	22.63	23.51	22.82	3.91	4.41	3.67
PMC-LLaMA-D2	0.18	0.05	1.0	19.69	1.72	22.51	23.89	22.21	4.47	5.1	4.17
fine-tuning											
BioGPT-D1	0.18	0.06	0.53	11.48	2.08	11.95	12.53	12.84	3.04	3.52	2.97
BioGPT-D2	0.27	0.11	0.94	19.79	3.42	23.33	24.63	24.27	5.84	6.97	5.45
BioMedLM-D1	**0.29**	0.13	1.0	**27.69**	9.14	**33.03**	**34.13**	33.55	12.43	**13.89**	11.82
BioMedLM-D2	**0.29**	**0.14**	1.0	27.14	**9.22**	31.95	32.06	**34.11**	**12.79**	13.81	**12.82**

Table 4: Effects of data augmentation and our designed prompts. B-n, R-nF are short for BLEU-n, ROUGE-nF1 respectively, n is used for n-grams.

Method	Explainability		Text quality			
	FCR	FMR	B-1	B-4	R-1F	R-2F
BioMedLM-D2	0.29	0.14	27.14	9.22	31.95	12.79
BioMedLM-D2 (woaug)	0.19	0.08	15.0	3.04	17.95	4.76
BioMedLM (woprompt)	0.24	0.06	19.62	4.15	22.95	5.26

4.2. Implementation details

Explanations in original M$_{\text{fact}}$ are first split by 25/20/100 for training/validation/test. Then we augment the training set and select 7 qualified rephrase sentences for each training sample. We also include M$_{\text{hypo}}$ into the training set. In this way, 762 explanations were used for training. Adam algorithm (Kingma and Ba, 2015) is used as the optimizer. We tune the learning rate in $[10^{-5}, 10^{-3}]$ with the maximum epoch 50. The batch size is tuned in [10, 100], the dropout rate of LoRA is 0.1, and the attention dimension of LoRA is tuned in [32, 128]. For KG prompt template, we sample 3 functional nodes from a KG subgraph to fill the template. During the inference stage, the probability of sampling a set of predicted tokens is tuned in [0.6, 0.92]. We run each experiment five times on an NVIDIA A40 GPU and compute the average results.

4.3. Prompt-based results

Zero-shot inference We first investigated the performance of the four pre-trained generative models in the zero-shot inference scenario. Each model adopts two types of prompts respectively. The results are shown in Table 3. Models named “*-D1" means only the instruction prompts are used for input gene pairs, while “*-D2" means that the instruction prompts and KG prompts are combined. We can see that the performance of most models is improved with increased model parameters. For each model, combining instruction with KG prompts yields better results than using only the instruction prompts. Such comparison is more obvious in BioGPT, the model with the least parameters. As the model parameters increased, the combination of instruction and the KG prompts mainly improved the text quality of the generated explanations, especially ROUGE-2-based scores.

LoRA fine-tuning Next, we evaluated pre-trained models in a parameter-efficient fine-tuning scenario. Considering the size of our dataset, we fine-tuned BioGPT and BioMedLM. (PMC-LLaMA cannot be fine-tuned due to the computational resources) The experimental results are displayed in Table 3. Compared to zero-shot inference results, the performance of fine-tuned BioGPT and BioMedLM is noticeably improved, demonstrating that the fine-tuning strategy is beneficial for generating explanations. Furthermore, BioGPT-D2 outperforms BioGPT-D1 across all metrics, consistent with the observation of zero-shot inference, meaning that KG prompts can boost the quality of explanations generated by BioGPT. For BioMedLM-based models, BioMedLM-D2 performs better than BioMedLM-D1 on FMR score, meaning that KG prompts can enhance the SL-related features for specific explanations. Interestingly, BioMedLM-D2 also surpasses BioMedLM-D1 on BLEU-4 and ROUGE2-based scores, indicating that combining instruction with KG prompts can aid the model in capturing more semantic information.

Ablation study To validate the efficacy of data augmentation method, we removed the 175 augmented training samples and then fine-tuned BioMedLM again. Moreover, to further test whether our designed prompts are beneficial for fine-tuning, we removed prompts and took only the names of target gene pairs as the input words for fine-tuning BioMedLM.

Table 4 shows the results evaluated by seven metrics. We can see that compared to BioMedLM using all training samples (i.e., BioMedLM-D2), excluding augmented training samples decreases the model performance. This means that the rephrased explanations effectively assist in fine-tuning the model. Additionally, compared to BioMedLM model taking prompts, removing prompts decreases performance across all metrics. When inspecting the predictions of BioMedLM without prompt, we found that many of these sentences explained the wrong target gene pairs, although these sentences include the ground truth features. This gene-pair-mismatch issue was substantially mitigated when we integrated prompts into the input sequences (see section 4.4 for specific cases). This is probably because the personalized prompts can provide specific SL knowledge for gene pairs, thereby guiding the model for more reasonable predictions.

4.4. Case study

We provide specific examples to demonstrate the effectiveness of fine-tuning and our prompts for generating SL explanations, as shown in Fig. 3. In the first case, given the same gene pair and prompts (both instruction and KG prompts), the zero-shot inference from BioGPT-D2 is too short and not reasonable, while the prediction of fine-tuned BioGPT-D2 is more reasonable, although gene functions are not included. For BioMedLM, the zero-shot explanation mentions that ERH is important to KRAS-mutant cancer cells. However, the response does not illustrate the specific functions of ERH that make it so essential. After fine-tuning, the generated sentence points out why ERH is important for KRAS mutation, i.e., ERH plays a vital role in regulating RNA-related transcriptions. This reason reveals a more profound SL mechanism and is closer to the key features in the ground truth, indicating that the model can better capture the knowledge in the KG prompts after fine-tuning, thus enhancing explainability.

In the second case, the fine-tuned model without prompt conducts an explanation for a wrong gene pair. However, when given a prompt encapsulating specific knowledge for a gene pair, the fine-tuned model does not lose the target gene pairs for the explanation. Moreover, since the prompt highlights DNA-related functions, the generated explanation also emphasizes the key role of DNA repair in the SL interaction between BRCA2 and RAD52.

5. Conclusion

In this study, we propose a prompt-based pipeline for generating natural language explanations. We first present NexLeth, a dataset for elucidating synthetic lethality mechanisms for existing gene pairs which is built using New Bing. NexLeth offers valuable insights for comprehensive research. Factual explanations in NexLeth can support biologists in analyzing the regularities of SL mechanisms, while the hypothetical explanations provide direction for wet-lab experiments and the identification of potential drug targets. Additionally, NexLeth acts as a benchmark for evaluating SL explanation models. We then designed personalized prompts by merging KG subgraph structures with instructions. Based on the prompts, we fine-tuned BioGPT and BioMedLM, two pre-trained biomedical language models, via an efficient tuning method. The experimental results demonstrate the effectiveness of our proposed prompt and the fine-tuning model in generating reliable SL explanations, positioning the fine-tuned model as a valuable tool for explaining new SL gene pairs.

In the future, we plan to continue updating NexLeth using our data collection pipeline and fine-tuning the model iteratively. The fine-tuning process can be refined by Reinforcement Learning from Human Feedback, incorporating expert feedback and using their selected explanations as training samples. We will also adopt knowledge graphs with more comprehensive SL knowledge and enhance the reasoning ability of language models, so as to further improve the explainability of generated natural language explanations. Our current model can only interpret potential SL gene pairs, without predicting the relationships of gene pairs. A more realistic scenario is to first predict the relationship between any gene pairs and then interpret the predicted results. Therefore, we will consider enhancing the model with predictive abilities before interpretation.

Cases	ground truth
Case 1 ***Input:*** KRAS and ERH may share common functions, including mRNA processing, mRNA splicing, metal ion binding, explain the synthetic lethality mechanism between KRAS and ERH: ***Zero-shot inference by BioGPT-D2:*** KRAS and ERH can cause or impair ERH functions. ***Zero-shot inference by BioMedLM-D2:*** KRAS-driven cancer cells are sensitive to ERH depletion, and that this sensitivity can be restored by restoring ERH expression. These results indicate that ERH plays a critical role in the KRAS-driven tumorigenesis. In this study, we identified ERH as a critical regulator of cell proliferation. Depletion of ERH causes growth inhibition in a variety of human cancer cell lines, and overexpression. ***Fine-tuned BioGPT-D2 output:*** when KRAS is upregulated, inhibition of ERH and KRAS induces cell death in cells with KRAS mutations and activating ERH, while overexpression of KRAS or ERH leads to cell death in cells with KRAS mutation but not in cells with KRAS loss of function. its membrane as in KRAS-mutant cancer cells but its membrane as in KRAS-normal cells involves its signaling as in KRAS-null cells. ***Fine-tuned BioMedLM-D2 output:*** KRAS and ERH have a synthetic lethality relationship, because ERH is a transcription factor that regulates RNA metabolism and RNA biogenesis, which is a process that is essential for KRAS-mutant cancer cells.	KRAS and ERH have a synthetic lethality relationship, because ERH controls the mRNA splicing of CENP-E, a key regulator of mitotic progression, and KRAS mutant cells are more sensitive to CENP-E depletion than wild-type cells.
Case 2 ***Input without prompt:*** BRCA2 RAD52 ***Fine-tuned BioMedLM output:*** KRAS and RB1 have a synthetic lethality relationship, because KRAS is a key oncogene that regulates cell proliferation and survival, and RB1 is a tumor suppressor gene that regulates cell cycle progression and DNA damage response. ***Input with prompt:*** BRCA2 and RAD52 may share common functions, including DNA Damage Response, DNA repair, explain the synthetic lethality mechanism between BRCA2 and RAD52: ***Fine-tuned BioMedLM-D2 output:*** BRCA2 and RAD52 have a synthetic lethality relationship, because BRCA2 is a key gene for homologous recombination, while RAD52 is a key protein that regulates homologous recombination. When homologous recombination is an essential pathway for cell survival under DNA damage, inhibiting both of RAD52 and BRCA2 can trigger synthetic lethality in BRCA2-mutant cancer cells under PARP inhibitors.	BRCA2 and RAD52 have a synthetic lethality relationship, because RAD52 plays a role in DNA repair and recombination. In BRCA2-deficient cells, which are involved in homologous recombination repair of DNA double-strand breaks (DSBs), RAD52 can facilitate an alternative mechanism of DNA repair known as single-strand annealing (SSA). SSA is an error-prone mechanism that can result in deletions and genomic instability.

Figure 3: Two cases of explaining SL mechanisms. We compared the predictions under different settings, i.e., zero-shot inference vs. fine-tuning and without prompt vs. with prompts. Key features are highlighted.

A. Appendix

A.1. Human annotation pipeline

Algorithm 2: Human annotation

```
Input: GenepairsCollection $Q_{nb}$,
  AnswerCollection $M_{nb}$, CitationCollection $R_{nb}$
FactGenePairsCollection $Q_{fact} \leftarrow \{\}$
FactAnswerCollection $M_{fact} \leftarrow \{\}$
FactCitationCollection $R_{fact} \leftarrow \{\}$
FeatureCollection $F_{fact} \leftarrow \{\}$
HypotheticalAnswerCollection $M_{hypo} \leftarrow \{\}$
for $i \leftarrow 1$ to $len(M_{nb})$ do
    gene pair $(u, v) = Q_{nb}[j]$
    answer $\leftarrow M_{nb}[i]$
    citations $\leftarrow R_{nb}[i]$
    if AnnotatorReadandCheck(answer, citations)
     then
        features $\leftarrow$ FeatureAnnotation(answer)
        $Q_{fact} \leftarrow$ (u, v)
        $M_{fact} \leftarrow$ answer
        $R_{fact} \leftarrow$ citations
        $F_{fact} \leftarrow$ features
    else
        add answers to $M_{hypo}$
    end
    $pairs_{new} \leftarrow$
     AnnotatorMiningNewpairs(citations)
    if $pairs_{new} \neq \varnothing$ then
        $answers_{new} \leftarrow$
         AnnotatorSummarization(citations)
        features $\leftarrow$ FeatureAnnotation(answers)
        $Q_{fact} \leftarrow pairs_{new}$
        $M_{fact}$ $answers_{new}$
        $R_{fact} \leftarrow$ citations
        $F_{fact} \leftarrow$ features
    end
end
```

A.2. A case of data augmentation prompt

You are a helpful assistant that rephrases text and makes sentences smooth. I will give you a sample, please rephrase the partial sentence after the word "because" of the sample, then give me 10 rephrased answers. Each answer should include the exact noun phrases which I will give you, and each answer must start with "because". The complete sample is: *TP53 and CDK2 have a synthetic lethality relationship, because TP53 is a tumor suppressor that regulates cell cycle arrest, apoptosis and DNA repair, and CDK2 is a cyclin-dependent kinase that controls cell cycle progression and DNA replication. Therefore, inhibition of CDK2 in TP53-mutant cells results in synergistic cell death due to impaired DNA repair and increased DNA damage.* The phrases are *"DNA repair"*, *"DNA damage"*, *"cell cycle progression"*, *"cell death"*.

A.3. From a KG subgraph to a personalized KG prompt

According to Fig. A.3, RAD52 shares two functions with BRAC2's SL partner genes, so we assume that RAD52 and BRAC2 also share the functions. The KG prompt for BRAC2 and RAD52 is: *BRAC2* and *RAD52* may share common functions, including *DNA Damage Response*, *DNA repair*.

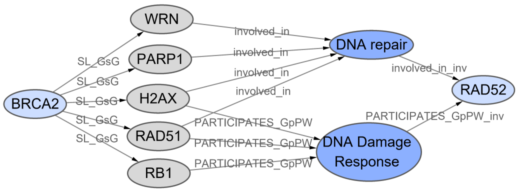


Figure 4: A KG subgraph for SL pair (BRCA2, RAD52). Light blue nodes are target genes, grey nodes are other genes having SL relationships with BRAC2, and dark blue nodes are gene functions.

B. References

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Haixing Dai, Zhengliang Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, et al. 2023. Chataug: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Alan Huang, Levi A Garraway, Alan Ashworth, and Barbara Weber. 2020. Synthetic lethality as an engine for cancer drug target discovery. *Nature reviews Drug discovery*, 19(1):23–38.

Heena Jariyal, Frank Weinberg, Abhinav Achreja, Deepak Nagarath, and Akshay Srivastava. 2020. Synthetic lethality: a step forward for personalized medicine in cancer. *Drug Discovery Today*, 25(2):305–320.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. 2023. GeneGPT: Augmenting large language models with domain tools for improved access to biomedical information. *arXiv preprint arXiv:2304.09667*.

Jinhao Ju, Deqing Yang, and Jingping Liu. 2022. Commonsense knowledge base completion with relational graph attention network and pre-trained language model. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 4104–4108.

William G Kaelin. 2005. The concept of synthetic lethality in the context of anticancer therapy. *Nature reviews cancer*, 5(9):689–698.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, San Diego, CA, USA*.

Zeljko Kraljevic, Anthony Shek, Daniel Bean, Rebecca Bendayan, James Teo, and Richard Dobson. 2021. MedGPT: Medical concept prediction from clinical narratives. *arXiv preprint arXiv:2107.03134*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Joo Sang Lee, Avinash Das, Livnat Jerby-Arnon, Rand Arafeh, Noam Auslander, Matthew Davidson, Lynn McGarry, Daniel James, Arnaud Amzallag, Seung Gu Park, et al. 2018. Harnessing synthetic lethality to predict the response to cancer treatment. *Nature communications*, 9(1):2546.

Joo Sang Lee, Nishanth Ulhas Nair, Gal Dinstag, Lesley Chapman, Youngmin Chung, Kun Wang, Sanju Sinha, Hongui Cha, Dasol Kim, Alexander V Schperberg, et al. 2021. Synthetic lethality-mediated precision oncology via the tumor transcriptome. *Cell*, 184(9):2487–2502.

Lei Li, Yongfeng Zhang, and Li Chen. 2021. Personalized transformer for explainable recommendation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 4947–4957. Association for Computational Linguistics.

Lei Li, Yongfeng Zhang, and Li Chen. 2023. Personalized prompt learning for explainable recommendation. *ACM Transactions on Information Systems*, 41(4):1–26.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for*

Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4582–4597, Online. Association for Computational Linguistics.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 1073–1094. Association for Computational Linguistics.

Peng Liu, Lemei Zhang, and Jon Atle Gulla. 2023a. Pre-train, prompt and recommendation: A comprehensive survey of language modelling paradigm adaptations in recommender systems. *arXiv preprint arXiv:2302.03735*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Xin Liu, Jiale Yu, Siyu Tao, Beiyuan Yang, Shike Wang, Lin Wang, Fang Bai, and Jie Zheng. 2022. PiLSL: pairwise interaction learning-based graph neural network for synthetic lethality prediction in human cancers. *Bioinformatics*, 38(Supplement_2):ii106–ii112.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6).

Assaf Magen, Avinash Das Sahu, Joo Sang Lee, Mahfuza Sharmin, Alexander Lugo, J Silvio Gutkind, Alejandro A Schäffer, Eytan Ruppin, and Sridhar Hannenhalli. 2019. Beyond synthetic lethality: charting the landscape of pairwise gene expression states associated with survival in cancer. *Cell reports*, 28(4):938–948.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. 2022. PEFT: State-of-the-art Parameter-Efficient Fine-Tuning methods. `https://github.com/huggingface/peft`.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. 2023. Unifying large language models and knowledge graphs: A roadmap. *arXiv preprint arXiv:2306.08302*.

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

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

A Venigalla, J Frankle, and M Carbin. 2022. Biomedlm: a domain-specific large language model for biomedical text. *MosaicML. Accessed: Dec*, 23(3):2.

Benyou Wang, Qianqian Xie, Jiahuan Pei, Prayag Tiwari, Zhao Li, et al. 2021a. Pre-trained language models in biomedical domain: A systematic survey. *arXiv preprint arXiv:2110.05006*.

Jie Wang, Min Wu, Xuhui Huang, Li Wang, Sophia Zhang, Hui Liu, and Jie Zheng. 2022a. SynLethDB 2.0: a web-based knowledge graph database on synthetic lethality for novel anticancer drug discovery. *Database*, 2022.

Jing Wang, Qinglong Zhang, Junshan Han, Yanpeng Zhao, Caiyun Zhao, Bowei Yan, Chong Dai, Lianlian Wu, Yuqi Wen, Yixin Zhang, et al. 2022b. Computational methods, databases and tools for synthetic lethality prediction. *Briefings in Bioinformatics*, 23(3):bbac106.

Shike Wang, Fan Xu, Yunyang Li, Jie Wang, Ke Zhang, Yong Liu, Min Wu, and Jie Zheng. 2021b. KG4SL: knowledge graph neural network for synthetic lethality prediction in human cancers. *Bioinformatics*, 37(Supplement_1):i418–i425.

Chengwei Wei, Yun-Cheng Wang, Bin Wang, and C-C Jay Kuo. 2023. An overview on language models: Recent developments and outlook. *arXiv preprint arXiv:2303.05759*.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*.

Xuansheng Wu, Kaixiong Zhou, Mingchen Sun, Xin Wang, and Ninghao Liu. 2023b. A survey of graph prompting methods: Techniques, applications, and challenges. *arXiv preprint arXiv:2303.07275*.

An Yan, Zhankui He, Jiacheng Li, Tianyang Zhang, and Julian McAuley. 2023. Personalized showcases: Generating multi-modal explanations for recommendations.

Ke Zhang, Min Wu, Yong Liu, Yimiao Feng, and Jie Zheng. 2023. KR4SL: knowledge graph reasoning for explainable prediction of synthetic lethality. *Bioinformatics (Oxford, England)*, 39(Supplement_1):i158–i167.