Rewire-then-Probe: A Contrastive Recipe for Probing Biomedical Knowledge of Pre-trained Language Models

Anonymous ACL submission

Abstract

Knowledge probing is crucial for understanding the knowledge transfer mechanism behind the pre-trained language models (PLMs). Despite the growing progress of probing knowledge for PLMs in the general domain, specialised areas such as biomedical domain are vastly under-explored. To facilitate this, we release a well-curated biomedical knowledge probing benchmark, `MedLAMA`, constructed based on the Unified Medical Language System (UMLS) Metathesaurus. We test a wide spectrum of state-of-the-art PLMs and probing approaches on our benchmark, reaching at most 3% of acc@10. While highlighting various sources of domain-specific challenges that amount to this underwhelming performance, we illustrate that the underlying PLMs have a higher potential for probing tasks. To achieve this, we propose Contrastive-Probe, a novel self-supervised contrastive probing approach, that adjusts the underlying PLMs without using any probing data. While Contrastive-Probe pushes the acc@10 to 28%, the performance gap still remains notable. Our human expert evaluation suggests that the probing performance of our Contrastive-Probe is still under-estimated as UMLS still does not include the full spectrum of factual knowledge. We hope `MedLAMA` and Contrastive-Probe facilitate further developments of more suited probing techniques for this domain.[1]

1 Introduction

Pre-trained language models (PLMs; Devlin et al. 2019; Liu et al. 2020) have orchestrated incredible progress on myriads of few- or zero-shot language understanding tasks, by pre-training model parameters in a task-agnostic way and transferring knowledge to specific downstream tasks via fine-tuning (Brown et al., 2020; Petroni et al., 2021).

To better understand the underlying knowledge transfer mechanism behind these achievements, many knowledge probing approaches and benchmark datasets have been proposed (Petroni et al., 2019; Jiang et al., 2020a; Kassner et al., 2021; Zhong et al., 2021). This is typically done by formulating knowledge triples as cloze-style queries with the objects being masked (see Table 1) and using the PLM to fill the single (Petroni et al., 2019) or multiple (Ghazvininejad et al., 2019) [Mask] token(s) without further fine-tuning.

	Query	Answer(s)
Hard Queries	*Riociguat* **has physiologic effect** [Mask].	*Vasodilation*
	Entecavir **may prevent** [Mask].	*Hepatitis B*
	Invasive Papillary Breast Carcinoma **disease mapped to gene** [Mask].	*[ERBB2 Gene, CCND1 Gene]*
Easy Queries	*Posttraumatic arteriovenous fistula* **is associated morphology of** [Mask].	*Traumatic arteriovenous fistula*
	Acute Myeloid Leukemia with Mutated RUNX1 **disease mapped to gene** [Mask].	*RUNX1 Gene*
	Magnesium Chloride **may prevent** [Mask].	*Magnesium Deficiency*

Table 1: Example probing queries from `MedLAMA`. **Bold** font denotes UMLS relation.

In parallel, it has been shown that specialised PLMs (e.g., BioBERT; Lee et al. 2020, BlueBERT; Peng et al. 2019 and PubMedBERT; Gu et al. 2020) substantially improve the performance in several biomedical tasks (Gu et al., 2020). The biomedical domain is an interesting testbed for investigating knowledge probing for its unique challenges (including vocabulary size, multi-token entities), and the practical benefit of potentially disposing the expensive knowledge base construction process. However, research on knowledge probing in this domain is largely under-explored.

To facilitate research in this direction, we present a well-curated biomedical knowledge probing benchmark, `MedLAMA`, that consists of 19 thoroughly selected relations. Each relation contains $1k$ queries ($19k$ queries in total with at most 10 answers each), which are extracted from the large UMLS (Bodenreider, 2004) biomedical knowledge graph and verified by domain experts.

[1] Code and data are attached in the submission.

We use automatic metrics to identify the hard examples based on the hardness of exposing answers from their query tokens. See Table 1 for a sample of easy and hard examples from MedLAMA.

A considerable challenge in probing in biomedical domain is handling multi-token encoding of the answers (e.g. in MedLAMA only 2.6% of the answers are single-token, while in the English set of mLAMA; Kassner et al. 2021, 98% are single-token), where all existing approaches (i.e., mask predict; Petroni et al. 2019, retrieval-based; Dufter et al. 2021, and generation-based; Gao et al. 2020) struggle to be effective.[2] For example, the mask predict approach (Jiang et al., 2020a) which performs well in probing multilingual knowledge achieves less than 1% accuracy on MedLAMA.

To address the aforementioned challenge, we propose a new method, **Contrastive-Probe**, that first adjusts the representation space of the underlying PLMs by using a retrieval-based contrastive learning objective (like '*rewiring*' the switchboard to the target appliances Liu et al. 2021c) then retrieves answers based on their representation similarities to the queries. Notably, our Contrastive-Probe does not require using the MLM heads during probing, which avoids the vocabulary bias across different models. Additionally, retrieval-based probe is effective for addressing the multi-token challenge, as it avoids the need to generate multiple tokens from the MLM vocabulary. We show that Contrastive-Probe facilitates absolute improvements of up-to ~6% and ~25% on the acc@1 and acc@10 probing performance compared with the existing approaches.

We further highlight that the elicited knowledge by Contrastive-Probe is not gained from the additional random sentences, but from the original pre-trained parameters, which echos the previous finding of Liu et al. (2021b); Glavaš and Vulić (2021). Additionally, we demonstrate that different state-of-the-art PLMs and transformer layers are suited for different types of relational knowledge, and different relations requires different depth of tuning, suggesting that both the layers and tuning depth should be considered when infusing knowledge over different relations. Furthermore, expert evaluation of PLM responses on a subset of MedLAMA highlights that expert-crafted resources such as UMLS still do not include the full spectrum of factual knowledge, indicating that the factual information encoded in PLMs is richer than what is reflected by the automatic evaluation.

The findings of our work, along with the proposed MedLAMA and Contrastive-Probe, highlight both the unique challenges of the biomedical domain and the unexploited potential of PLMs. We hope our research to shed light on what domain-specialised PLMs capture and how it could be better resurfaced, with minimum cost, for probing.

2 MedLAMA

To facilitate research of knowledge probing in the biomedical domain, we create the MedLAMA benchmark based on the largest biomedical knowledge graph UMLS (Bodenreider, 2004). UMLS[3] is a comprehensive metathesaurus containing 3.6 million entities and more than 35.2 million knowledge triples over 818 relation types which are integrated from various ontologies, including SNOMED CT, MeSH and the NCBI taxonomy.

Creating a LAMA-style (Petroni et al., 2019) probing benchmark from such a knowledge graph poses its own challenges: (1) UMLS is a collection of knowledge graphs with more than 150 ontologies constructed by different organisations with very different schemata and emphasis; (2) a significant amount of entity names (from certain vocabularies) are unnatural language (e.g., *t(8;21)(q22;q22)* denoting an observed karyotypic abnormality) which can hardly be understood by the existing PLMs, with tokenisation tailored for natural language; (3) some queries (constructed from knowledge triples) can have up to hundreds of answers (i.e., 1-to-N relations), complicating the interpretation of probing performance; and (4) some queries may expose answers in themselves (e.g., answer within queries), making it challenging to interpret relative accuracy scores.

Selection of Relationship Types. In order to obtain high-quality knowledge queries, we conducted multiple rounds of manual filtering on the relation level to exclude uninformative relations or relations that are only important in the ontological context but do not contain interesting semantics as a natural language (e.g, taxonomy and measurement relations). We also excluded relations

[2] Prompt-based probing approaches such as AutoPrompt (Shin et al., 2020a), SoftPrompt (Qin and Eisner, 2021), and OptiPrompt (Zhong et al., 2021) need additional labelled data for fine-tuning prompts, but we restrict the scope of our investigation to methods that do not require task data.

[3] Release version 2021AA: https://download.nlm.nih.gov/umls/kss/2021AA/umls-2021AA-full.zip

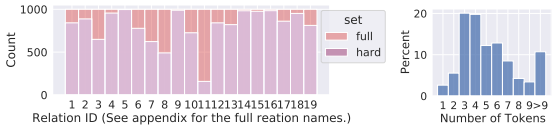

Figure 1: Left: Count over full and hard sets. Right: Percentage of answers over number of tokens.

with insufficient triples/entities. Then, we manually checked the knowledge triples for each relation to filter out those that contain unnatural language entities and ensure that their queries are semantically meaningful. Additionally, in the cases of 1-to-N relations where there are multiple gold answers for the same query, we constrained all the queries to contain at most 10 gold answers. These steps resulted in 19 relations with each containing 1k randomly sampled knowledge queries. See *Appendix* for the detailed relation names and their corresponding prompts.

Easy vs. Hard Queries. Recent works (Poerner et al., 2020; Shwartz et al., 2020) have discovered that PLMs are overly reliant on the surface form of entities to guess the correct answer of a knowledge query. The PLMs "cheat" by detecting lexical overlaps between the query and answer surface forms instead of exercising their abilities of predicting factual knowledge. For instance, PLMs can easily deal with the triple <*Dengue virus live antigen CYD serotype 1*, `may-prevent`, *Dengue*> since the answer is part of the query. To mitigate such bias, we also create a hard query set for each relation by selecting a subset of their corresponding 1k queries using token and matching metrics (i.e., exact matching and ROUGE-L (Lin and Och, 2004)). For more details see the *Appendix*. We refer to the final filtered and original queries as the **hard sets** and **full sets**, respectively. Figure 1 (left) shows the count of hard vs. full sets.

The Multi-token Issue. One of the key challenges for probing `MedLAMA` is the multi-token decoding of its entity names. In `MedLAMA` there are only 2.6% of the entity names that are single-token[4] while in the English set of mLAMA (Kassner et al., 2021) and LAMA (Petroni et al., 2019) the percentage of single-token answers are 98% and 100%, respectively. Figure 1 (right) shows the percentage of answers by different token numbers.

[4]Tokenized by Bert-base-uncased.

Approach	Type	Answer space	MLM
Fill-mask (Petroni et al., 2019)	MP	PLM Vocab	✓
X-FACTR (Jiang et al., 2020a)	MP	PLM Vocab	✓
Generative PLMs (Lewis et al., 2020)	GB	PLM Vocab	✗
Mask average (Kassner et al., 2021)	RB	KG Entities	✓
Contrastive-Probe (Ours)	RB	KG Entities	✗

Table 2: Comparison of different approaches. Types of probing approaches: Mask predict (MP), Retrieval-based (RB) and Generation-based (GB).

3 Existing Multi-token Knowledge Probing Approaches

While the pioneer works in PLM knowledge probing mainly focused on the single-token entities, many recent works have started exploring the solutions for the multi-token scenario (Kassner et al., 2021; Jiang et al., 2020a; De Cao et al., 2021). These knowledge probing approaches can be categorised, based on answer search space and reliance on MLM head, into three categories: *mask predict*, *generation-based*, and *retrieval-based*. Table 2 summarises their key differences.

Mask Predict. Mask predict (Petroni et al., 2019; Jiang et al., 2020a) is one of the most commonly used approaches to probe knowledge for masked PLMs (e.g. BERT). The mask predict approach uses the MLM head to fill a single mask token for a cloze-style query, and the output token is subjected to the PLM vocabulary (Petroni et al., 2019). Since many real-world entity names are encoded with multiple tokens, the mask predict approach has also been extended to predict multi-token answers using the conditional masked language model (Jiang et al., 2020a; Ghazvininejad et al., 2019). Figure 2(a) shows the prediction process. Specifically, given a query, the probing task is formulated as: 1) filling masks in parallel independently (*Independent*); 2) filling masks from left to right autoregressively (*Order*); 3) filling tokens sorted by the maximum confidence greedily (*Confidence*). After all mask tokens are replaced with the initial predictions, the predictions can be further refined by iteratively modifying one token at a time until convergence or until the maximum number of iterations is reached (Jiang et al., 2020a). For example, *Order+Order* represents that the answers are initially predicted by *Order* and then refined by *Order*. In this paper we examined two of these approaches, i.e. *Independent* and *Order+Order*, based on our initial exploration.

Generation-based. Recently, many generation based PLMs have been presented for text generation tasks, such as BART (Lewis et al., 2020) and

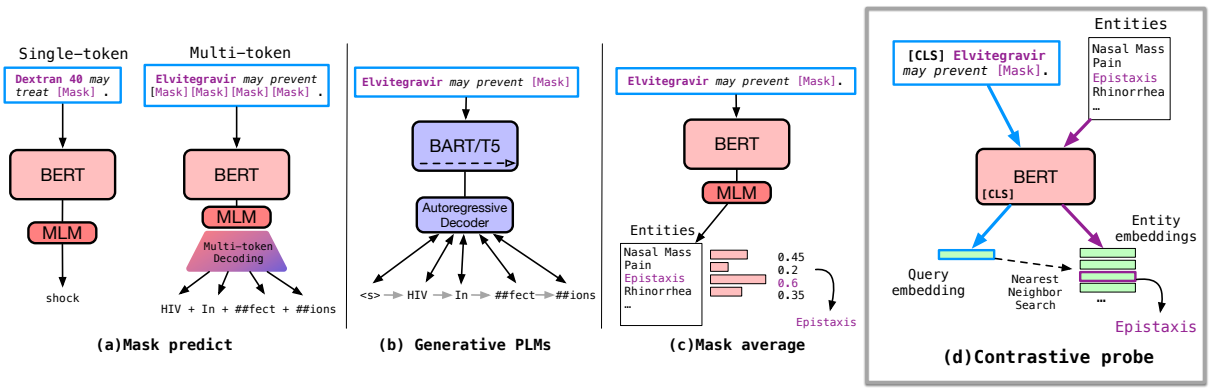

Figure 2: Comparison of different probing approaches. (d) is our proposed CONTRASTIVE-PROBE.

T5 (Raffel et al., 2020). These generative PLMs are trained with a de-noising objective to restore its original form autoregressively (Lewis et al., 2020; Raffel et al., 2020). Such an autoregressive generation process is analogous to the *Order* probing approach, thus the generative PLMs can be directly used to generate answers for each query. Specifically, we utilize the cloze-style query with a single [MASK] token as the model input. The model then predicts the answer entities that correspond to the [MASK] token in an autoregressive manner. An illustration is provided in Figure 2(b).

Retrieval-based. Mask predict and Generation-based approaches need to use the PLM vocabulary as their search spaces for answer tokens, which may generate answers that are not in the answer set. In particular, when probing the masked PLMs using their MLM heads, the predicted result might not be a good indicator for measuring the amount of knowledge captured by these PLMs. This is mainly because the MLM head will be eventually dropped during the downstream task fine-tuning while the MLM head normally accounts for more than 20% of the total PLM parameters. Alternatively, the retrieval-based probing (Dufter et al., 2021; Kassner et al., 2021) are applied to address this issue. Instead of generating answers based on the PLM vocabulary, the retrieval-based approach finds answers by ranking the knowledge graph candidate entities based on the query and entity representations, or the entity generating scores. To probe PLMs on `MedLAMA`, we use *mask average* (Kassner et al., 2021), an approach that takes the average log probabilities of entity's individual tokens to rank the candidates. The retrieval-based approaches address the multi-token issue by restricting the output space to the valid answer set and can be used to probe knowledge in different types of PLMs (e.g. BERT vs. fastText; Dufter et al. 2021). However, previous works (Kassner et al., 2021; Dufter et al., 2021) only report results based on the type-restricted candidate set (e.g. relation) which we observed to decay drastically under the full entity set.

4 CONTRASTIVE-PROBE: Cloze-style Task as a Self-retrieving Game

To better transform the PLM encoders for the cloze-style probing task, we propose CONTRASTIVE-PROBE which pre-trains on a small number of sentences sampled from the PLM's *original pre-training corpora* with a contrastive self-supervising objective, inspired by the Mirror-BERT (Liu et al., 2021b). Our contrastive pre-training does not require the MLM head or any additional external knowledge, and can be completed in less than one minute on $2 \times$ 2080Ti GPUs.

Self-supervised Contrastive Rewiring. We randomly sample a small set of sentences (e.g. 10k, see §5.2 for stability analysis of CONTRASTIVE-PROBE on several randomly sampled sets), and replace their tail tokens (e.g. the last 50% excluding the full stop) with a [MASK] token. Then these transformed sentences are taken as the queries of the *cloze-style self-retrieving game*. In the following we show an example of transforming a sentence into a cloze-style query:

> **Sentence:** Social-distancing largely reduces coronavirus infections.
> **Query:** Social-distancing largely [MASK].

where "reduces coronavirus infections" is marked as a positive answer of this query.

Given a batch, the *cloze-style self-retrieving game* is to ask the PLMs to retrieve the positive answer from all the queries and answers in the same batch. Our CONTRASTIVE-PROBE tackles this by optimising an InfoNCE objective (Oord et al., 2018),

$$\mathcal{L} = -\sum_{i=1}^{N} \log \frac{\exp(\cos(f(x_i), f(x_p))/\tau)}{\sum_{x_j \in \mathcal{N}_i} \exp(\cos(f(x_i), f(x_j))/\tau)}, \quad (1)$$

where $f(\cdot)$ is the PLM encoder (with the MLM head chopped-off and `[CLS]` as the contextual representation), N is batch size, x_i and x_p are from a query-answer pair (i.e., x_i and x_p are from the same sentence), $\mathcal{N}_i$ contains queries and answers in the batch, and τ is the temperature. This objective function encourages f to create similar representations for any query-answer pairs from the same sentence and dissimilar representations for queries/answers belonging to different sentences.

Retrieval-based Probing. For probing step, the query is created based on the prompt-based template for each knowledge triple (see *Appendix* for our manual prompts), as shown in the following:

> **Triple:** <*Elvitegravir*, `may-prevent`, *Epistaxis*>
> **Query:** Elvitegravir may prevent [MASK].

and we search for nearest neighbours from all the entity representations encoded by the same model.

5 Experiments

In this section we conduct extensive experiments to verify whether CONTRASTIVE-PROBE is effective for probing biomedical PLMs. First, we experiment with CONTRASTIVE-PROBE and existing probing approaches on `MedLAMA` benchmark (§5.1). Then, we conduct in-depth analysis of the stability and applicability of CONTRASTIVE-PROBE in probing biomedical PLMs (§5.2). Finally, we report an evaluation of a biomedical expert on the probing predictions and highlight our findings (§5.3).

CONTRASTIVE-PROBE Rewiring. We train our CONTRASTIVE-PROBE based on 10k sentences which are randomly sampled from the PubMed texts[5] using a mask ratio of 0.5. The best hyperparameters and their tuning options are provided in *Appendix*.

Probing Baselines. For the mask predict approach, we use the original implementation of X-FACTR (Jiang et al., 2020a), and set the beam size and the number of masks to 5. Both mask predict and retrieval-based approaches are tested under both the general domain and biomedical domain BERT models, i.e. Bert-based-uncased (Devlin et al., 2019), BlueBERT (Peng et al., 2019), BioBERT (Lee et al., 2020), PubMedBERT (Gu et al., 2020).[6] For generation-based baselines, we test five PLMs, namely BART-base (Lewis et al., 2020), T5-small and T5-base (Raffel et al., 2020) that are general domain generation PLMs, and SciFive-base & SciFive-large (Phan et al., 2021) that are pre-trained on large biomedical corpora.

Approach	PLM	Full Set	
		acc@1	acc@10
Generative PLMs	BART-base	0.16	1.39
	SciFive-base	0.53	2.02
	SciFive-large	0.55	2.03
	T5-small	0.70	1.72
	T5-base	0.06	0.19
X-FACTR (*Confidence*)	BERT	0.05	-
	BlueBERT	0.74	-
	BioBERT	0.17	-
X-FACTR (*Order+Order*)	BERT	0.06	-
	BlueBERT	0.50	-
	BioBERT	0.11	-
Mask average	BERT	0.06	0.73
	BlueBERT	0.05	1.39
	BioBERT	0.28	3.03
CONTRASTIVE-PROBE (Ours)	BERT	0.94	4.96
	BlueBERT	<u>5.20</u>	<u>20.06</u>
	BioBERT	4.54	19.59
	PubMedBERT	**7.32**	**27.70**

Table 3: Performance of different probing approaches on the full set of `MedLAMA`. Since the MLM head of PubMedBERT is not available, the mask predict and mask average approaches cannot be applied. Best results are in **bold** and the second bests are <u>underlined</u>.

5.1 Benchmarking on `MedLAMA`

Comparing Various Probing Approaches. Table 3 shows the overall results of various probing baselines on `MedLAMA`. It can be seen that the performances of all the existing probing approaches (i.e. *generative PLMs*, *X-FACTR* and *mask predict*) are very low (<1% for acc@1 and <4% for acc@10) regardless of the underlying PLM, which are not effective indicators for measuring knowledge captured. In contrast, our CONTRASTIVE-PROBE obtains absolute improvements by up-to ~ 6% and ~ 25% on acc@1 and acc10 respectively comparing with the three existing approaches, which validates its effectiveness on measuring the knowledge probing performance. In particular, PubMedBERT model obtains the best probing performance (7.32% in accuracy) for these biomedical queries, validating its effectiveness of capturing biomedical knowledge comparing with other PLMs (i.e. BERT, BlueBERT and BioBERT).

Benchmarking with CONTRASTIVE-PROBE. To further examine the effectiveness of PLMs in captur-

[5] We use the text from PubMed used for pre-training of BlueBERT model (Peng et al., 2019).

[6] The MLM head of PubMedBERT is not publicly available and cannot be evaluated by X-FACTR and *mask average*.

Model	acc@1/acc@10	
	Full Set	Hard Set
BERT (Devlin et al., 2019)	0.94/4.96	0.74/2.31
BlueBERT (Peng et al., 2019)	5.20/20.06	4.34/18.91
BioBERT (Lee et al., 2020)	4.54/19.59	3.84/15.01
ClinicalBERT (Huang et al., 2019)	1.03/5.61	0.66/4.38
SciBERT (Beltagy et al., 2019)	4.17/16.90	2.83/14.78
PubMedBERT (Gu et al., 2020)	7.32/**27.70**	4.86/**23.51**
SapBERT (Liu et al., 2021a)	4.65/24.53	2.02/21.78
UmlsBERT (Michalopoulos et al., 2021)	1.68/7.19	0.94/5.68
CoderBERT (Yuan et al., 2020)	**8.38**/24.05	**6.40**/21.79

Table 4: Benchmarking biomedical PLMs on MedLAMA (Full and Hard) via Contrastive-Probe. The bottom panel are knowledge-enhanced PLMs.

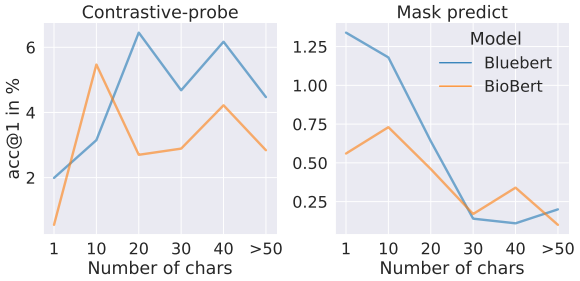

Figure 3: Performance over answer lengths.

ing biomedical knowledge, we benchmarked several state-of-the-art biomedical PLMs (including pure pre-trained and knowledge-enhanced models) on MedLAMA through our Contrastive-Probe. Table 4 shows the probing results over the full and hard sets (detailed macro and micro accuracies are provided in *Appendix*). In general, we can observe that these biomedical PLMs always perform better than general-domain PLMs (i.e., BERT). Also, we observe the decay of performance of all these models on the more challenging hard set queries. While PubMedBERT performs the best under all metrics, CoderBERT (Yuan et al., 2020) (which is the knowledge infused PubMedBERT) achieves better performance on micro acc@1, highlighting the benefits of knowledge infusion pre-training.

Comparison per Answer Length. Since different PLMs use different tokenizers, we use char length of the query answers to split MedLAMA into different bins and test the probing performance over various answer lengths. Figure 3 shows the result. We can see that the performance of retrieval-based probing in Contrastive-Probe increases as the answer length increase while the performance of mask predict dropped significantly. This result validates that our Contrastive-Probe (retrieval-based) are more reliable at predicting longer answers than the mask predict approach since the latter heavily relies on the MLM head.[7]

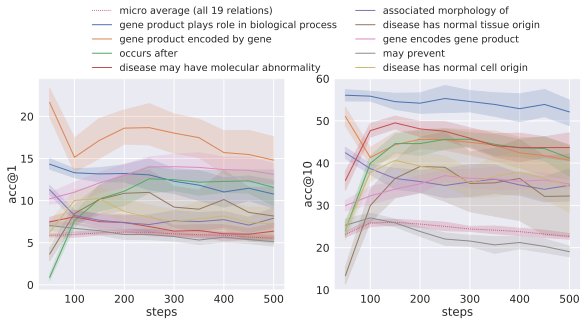

Figure 4: Performance over training steps on full set. The shaded regions are the standard deviations.

5.2 In-depth Analysis of Contrastive-Probe

Since our Contrastive-Probe involves many hyperparameters and stochastic factors during self-retrieving pre-training, it is critical to verify if it behaves consistently under (1) different randomly sampled sentence sets; (2) different types of relations; and (3) different pre-training steps.

Stability of Contrastive-Probe. To conduct this verification, we sampled 10 different sets of 10k sentences from the PubMed corpus[8] and probed the PubMedBERT model using our Contrastive-Probe on the full set. Figure 4 shows the acc@1 performance over top 9 relations and the micro average performance of all the 19 relations. We can see that the standard deviations are small and the performance over different sets of samples shows the similar trend. This further highlights that the probing success of Contrastive-Probe is not due the selected pre-training sentences. Intuitively, the contrastive self-retrieving game (§4) is equivalent to the formulation of the cloze-style filling task, hence tuning the underlying PLMs makes them better suited for knowledge elicitation needed during probing (like 'rewiring' the switchboards). Additionally, from Figure 4 we can also observe that different relations exhibit very different trends during pre-training steps of Contrastive-Probe and peak under different steps, suggesting that we need to treat different types of relational knowledge with different tuning depths when infusing knowledge. We leave further exploration of this to future work.

[7]For the single-token answer probing scenario, Contrastive-Probe does not outperform the mask predict approach, particularly in the general domain. This is expected since most of the masked PLMs are pre-trained by a single-token-filling objective.

[8]The tuning corpus itself is unimportant, since we can obtain the similar results even using Wikipedia.

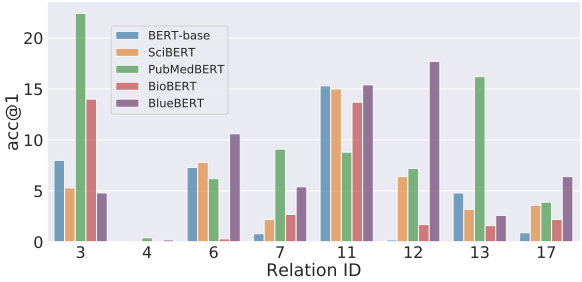

Figure 5: Performance of PLMs over partial relations. See appendix for the full relation ID and their names.

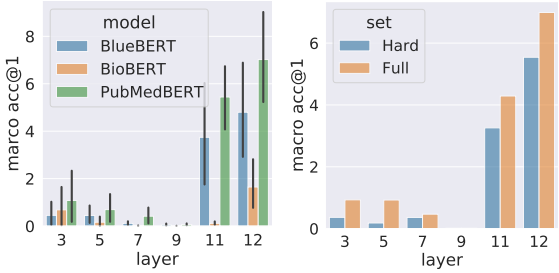

Figure 6: Performance over different layers.

Probing by Relations. To further analyse the probing variance over different relations, we also plot the probing performance of various PLMs over a subset of relations of `MedLAMA` in Figure 5 (for full plot see the *Appendix*). We can observe that different PLMs exhibit different performance rankings over different types of relational knowledge (e.g. BlueBERT peaks at *relation 12* while PubMedBERT peaks at *relation 3*). This result demonstrates that different PLMs are suited for different types of relational knowledge. We speculate this to be reflective of their training corpora.

Probing by Layer. To investigate how much knowledge is stored in each Transformer layer, we chopped the last layers of PLMs and applied Contrastive-Probe to evaluate the probing performance based on the first $L \in \{3, 5, 7, 9, 11, 12\}$ layers on `MedLAMA`. In general, we can see in Figure 6 that the model performance drops significantly after chopping the last 3 layers, while its accuracy is still high when dropping only last one layer. In Figure 7, we further plot the layer-wise probing performance of PubMedBERT over different relations. Surprisingly, we find that different relations do not show the same probing performance trends over layers. For example, with only the first 3 layers, PubMedBERT achieves the best accuracy (>15%) on *relation 11*[9] queries. This result demonstrates that both relation types and PLM layers are confounding variables in capturing factual knowledge, which helps to explain the

[9] *relation 11*: [X] is associated morphology of [Y].

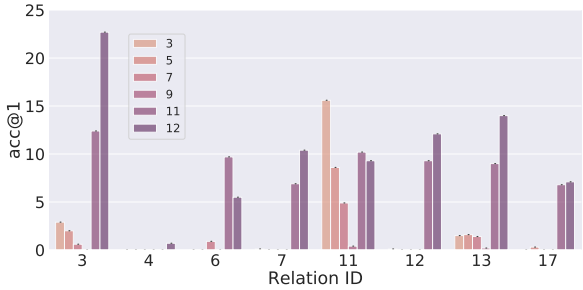

Figure 7: Performance of PubMedBERT over layers. See appendix for the full relation ID and their names.

difference of training steps over relations in Figure 4. This result also suggests that layer-wise and relation-wise training could be the key to effectively infuse factual knowledge for PLMs.

5.3 Expert Evaluation on Predictions

To assess whether the actual probing performance could be possibly higher than what is reflected by the commonly used automatic evaluation, we conducted a human evaluation on the prediction result. Specifically, we sample 15 queries and predict their top-10 answers using Contrastive-Probe based on PubMedBERT and ask the assessor[10] to rate the predictions on a scale of [1,5]. Figure 8 shows the confusion matrices.[11] We observe the followings: (1) There are 3 UMLS answers that are annotated with score level 1-4 (precisely, level 3), which indicates UMLS answers might not always be the perfect answers. (2) There are 20 annotated perfect answers (score 5) in the top 10 predictions that are not marked as the gold answers in the UMLS, which suggests the UMLS does not include all the expected gold knowledge. (3) In general, PubMedBERT achieves an 8.67% (13/150) acc@10 under gold answers, but under the expert annotation the acc@10 is 22%(33/150), which means the probing performance is higher than what evaluated using the automatically extracted answers.

6 Related Work and Discussion

Knowledge Probing Benchmarks for PLMs. LAMA (Petroni et al., 2019), which starts this line of work, is a collection of single-token knowledge triples extracted from sources including Wikidata and ConceptNet (Speer et al., 2017). To mitigate the problem of information leakage from the

[10] A senior Ph.D. graduate in Cell Biology.

[11] In the *Appendix*, we provide examples with their UMLS gold answers, human annotated answers and probing predictions of different probing approaches.

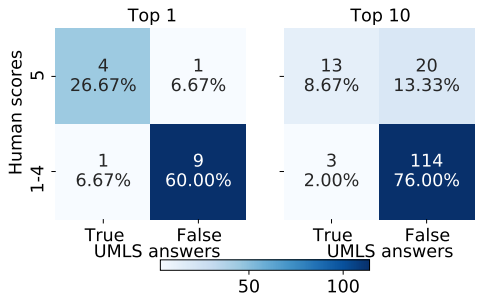

Figure 8: Confusion matrices of expert annotated scores versus the extracted UMLS answers. Five annotation score levels: 5-*Perfectly answer the query*; 4-*Similar to the gold answer, could somehow be the answer*; 3-*Related to the query but not correct*; 2-*Same domain or slight relation*; 1-*Completely unrelated*.

head entity, Poerner et al. (2019) propose LAMA-UHN, which is a hard subset of LAMA that has less token overlaps in head and tail entities. X-FACTR (Jiang et al., 2020a) and mLAMA (Kassner et al., 2021) extend knowledge probing to the multilingual scenario and introduce multi-token answers. They each propose decoding methods that generate multi-token answers, which we have shown to work poorly on `MedLAMA`. BioLAMA (Sung et al., 2021) is a concurrent work that also releases a benchmark for biomedical knowledge probing. We provide a comparison between LAMA, BioLAMA and `MedLAMA` in terms of (# relations, # queries, avg # answers per query, avg # characters per answer) in the *Appendix*.[12]

Probing via Prompt Engineering. Knowledge probing is sensitive to what prompt is used (Jiang et al., 2020b). To bootstrap the probing performance, Jiang et al. (2020b) mine more prompts and ensemble them during inference. Later works parameterised the prompts and made them trainable (Shin et al., 2020b; Fichtel et al., 2021; Qin and Eisner, 2021). We have opted out prompt-engineering methods that require training data in this work, as tuning the prompts are essentially tuning an additional (parameterised) model on top of PLMs. As pointed out by Fichtel et al. (2021), prompt tuning requires large amounts of training data from the task. Since task training data is used, the additional model parameters are exposed to the target data distribution and can solve the set set by overfitting to such biases (Cao et al., 2021). In our work, by adaptively finetuning the model with a small set of raw sentences, we elicit the knowledge out from PLMs but do not expose the data biases from the benchmark (`MedLAMA`).

Biomedical Knowledge Probing. Nadkarni et al. (2021) train PLMs as KB completion models and test on the same task to understand how much knowledge is in biomedical PLMs. BioLAMA focuses on the continuous prompt learning method OptiPrompt (Zhong et al., 2021), which also requires ground-truth training data from the task. Overall, compared to BioLAMA, we have provided a more comprehensive set of probing experiments and analysis, including proposing a novel probing technique and providing human evaluations of model predictions.

7 Conclusion

In this work, we created a carefully curated biomedical probing benchmark, `MedLAMA`, from the UMLS knowledge graph. We illustrated that state-of-the-art probing techniques and biomedical pre-trained languages models (PLMs) struggle to cope with the challenging nature (e.g. multi-token answers) of this specialised domain, reaching only an underwhelming 3% of acc@10. To reduce the gap, we further proposed a novel contrastive recipe which rewires the underlying PLMs without using any probing-specific data and illustrated that with a lightweight pre-training their accuracies could be pushed to 28%.

Our experiments also revealed that different layers of transformers encode different types of information, reflected by their individual success at handling certain types of prompts. Additionally, using a human expert, we showed that the existing evaluation criteria could overpenalise the models as many valid responses that PLMs produce are not in the ground truth UMLS knowledge graph. This further highlights the importance of having a human in the loop to better understand the potentials and limitations of PLMs in encoding domain specific factual knowledge.

Our findings indicate that the real lower bound on the amount of factual knowledge encoded by PLMs is higher than we estimated, since such bound can be continuously improved by optimising both the encoding space (e.g. using our self-supervised contrastive learning technique) and the input space (e.g. using the prompt optimising techniques (Shin et al., 2020a; Qin and Eisner, 2021)). We leave further exploration of integrating the two possibilities to future work.

[12]Our comparison indicated that `MedLAMA` and BioLAMA have no overlaps in queries, allowing both resources to complement each other.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *EMNLP*, pages 3606–3611.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *EACL*.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *ACL*, pages 1860–1874.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *ICLR*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186.

Philipp Dufter, Nora Kassner, and Hinrich Schütze. 2021. Static embeddings as efficient knowledge bases? In *NAACL*, pages 2353–2363.

Leandra Fichtel, Jan-Christoph Kalo, and Wolf-Tilo Balke. 2021. Prompt tuning or fine-tuning-investigating relational knowledge in pre-trained language models. In *AKBC*.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. In *ACL*.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer. 2019. Mask-predict: Parallel decoding of conditional masked language models. In *EMNLP*, pages 6112–6121.

Goran Glavaš and Ivan Vulić. 2021. Is supervised syntactic parsing beneficial for language understanding tasks? an empirical investigation. In *ACL*, pages 3090–3104.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020a. X-factr: Multilingual factual knowledge retrieval from pretrained language models. In *EMNLP*, pages 5943–5959.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020b. How can we know what language models know? *TACL*, 8:423–438.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual lama: Investigating knowledge in multilingual pretrained language models. In *ACL*, pages 3250–3258.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *ACL*, pages 7871–7880.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *ACL*, pages 605–612.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021a. Self-alignment pre-training for biomedical entity representations. *NAACL*.

Fangyu Liu, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2021b. Fast, effective, and self-supervised: Transforming masked language models into universal lexical and sentence encoders. *EMNLP*.

Qianchu Liu, Fangyu Liu, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2021c. Mirrorwic: On eliciting word-in-context representations from pretrained language models. In *CoNLL*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Roberta: A robustly optimized bert pretraining approach. In *ICLR*.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. In *NAACL*, pages 1744–1753.

Rahul Nadkarni, David Wadden, Iz Beltagy, Noah A Smith, Hannaneh Hajishirzi, and Tom Hope. 2021. Scientific language models for biomedical knowledge base completion: An empirical study. *arXiv preprint arXiv:2106.09700*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *BioNLP Workshop*, pages 58–65.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2021. KILT: a benchmark for knowledge intensive language tasks. *NAACL*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *EMNLP*, pages 2463–2473.

Long N Phan, James T Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598*.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2019. E-bert: Efficient-yet-effective entity embeddings for bert. *arXiv preprint arXiv:1911.03681*.

Nina Poerner, Ulli Waltinger, and Hinrich Schütze. 2020. E-bert: Efficient-yet-effective entity embeddings for bert. In *EMNLP: Findings*, pages 803–818.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying lms with mixtures of soft prompts. In *NAACL*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020a. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *EMNLP*.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. 2020b. Eliciting knowledge from language models using automatically generated prompts. In *EMNLP*, pages 4222–4235.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. "you are grounded!": Latent name artifacts in pre-trained language models. In *EMNLP*, pages 6850–6861.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*.

Mujeen Sung, Jinhyuk Lee, Sean Yi, Minji Jeon, Sungdong Kim, and Jaewoo Kang. 2021. Can language models be biomedical knowledge bases? In *EMNLP*.

Zheng Yuan, Zhengyun Zhao, and Sheng Yu. 2020. Coder: Knowledge infused cross-lingual medical term embedding for term normalization. *arXiv preprint arXiv:2011.02947*.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [mask]: Learning vs. learning to recall. In *NAACL*.

A Appendix

A.1 Details Relation Names and Prompts

Table 5 shows the detailed relation names and their manual prompts of our `MedLAMA`.

A.2 Details of the Hardness Metrics

In this paper, we use two automatic metrics to distinguish hard and easy queries. In particular, we first filter out easy queries by an exact matching metric (i.e. the exactly matching all the words of answer from queries). Since our `MedLAMA` contains multiple answers for queries, we use a threshold on the average exact matching score, i.e. **avg-match**>0.1, to filter out easy examples, where **avg-match** is calculated by:

$$\textbf{avg-match} = \frac{\textbf{Count}(\text{matched answers})}{\textbf{Count}(\text{total answers})}.$$

This metric can remove all the queries that match the whole string of answers. However, some common sub-strings between queries and answers also prone to reveal answers, particularly benefiting those retrieval-based probing approaches. E.g. <*Magnesium Chloride*, `may-prevent`, *Magnesium Deficiency*>. Therefore, we further calculate the **ROUGE-L** score (Lin and Och, 2004) for all the queries by regarding <query, answers> pairs as the <hypothesis, reference> pairs, and further filter out the **ROUGE-L**>0.1 queries.

A.3 Comparing with BioLAMA

During the writing of this work, we notice that Sung et al. (2021) also released a biomedical knowledge probing benchmark, called BioLAMA, which is a work concurrent to ours. In Table 6, we compare our `MedLAMA` with LAMA (Petroni et al., 2019) and BioLAMA (Sung et al., 2021) in terms of their statistics. We found that there is only 1 overlapped relation (i.e. may treat) between BioLAMA and our `MedLAMA`, and no same query can be found. Moreover, Sung et al. (2021) only use two existing probing approach on their proposed BioLAMA, while in this paper we further proposed a new probing approach Contrastive-Probe. We also evaluate our Contrastive-Probe in BioLAMA, and the result is shown in Table 7. We can see that, without additional training data from the biomedical knowledge facts, our `MedLAMA` shows very competitive performance comparing with OptiPrompt approach, which needs further training data. Additionally, since Mask Predict and OptiPrompt require using the MLM head, it is impossible to compare a model without MLM head being released (e.g. PubMedBERT). In contrast, our Contrastive-Probe not only provides a good indicator of comparing these models in terms of their captured knowledge, but also makes layer-wise knowledge probing possible.

A.4 Hyperparameters Tuning.

We train our Contrastive-Probe based on 10k sentences which are randomly sampled from the original pre-training corpora of the corresponding PLMs. Since most of the biomedical BERTs use PubMed texts as their pre-training corpora, for all biomedical PLMs we sampled random sentences from a version of PubMed corpus used by BlueBERT model (Peng et al., 2019), while for BERT we sampled sentences from its original wikitext corpora. For the hyperparamters of our Contrastive-Probe, Table 8 lists our search options and the best parameters used in our paper.

A.5 The Impact of Mask Ratios

To further investigate the impact of the mask ratio to the probing performance, we also test our Contrastive-Probe based on PubMedBERT over different mask ratios ({0.1, 0.2, 0.3, 0.4, 0.5}) under the 10 random sentence sets, the result of which is shown in Figure 9. We can see that over different mask ratios the Contrastive-Probe always reaches their best performance under certain pre-training steps. And the performance curves of mask ratios are different over the full and hard sets, but they all achieves a generally good performance when the mask ratio is 0.5, which validates that different mask ratios favour different types queries.

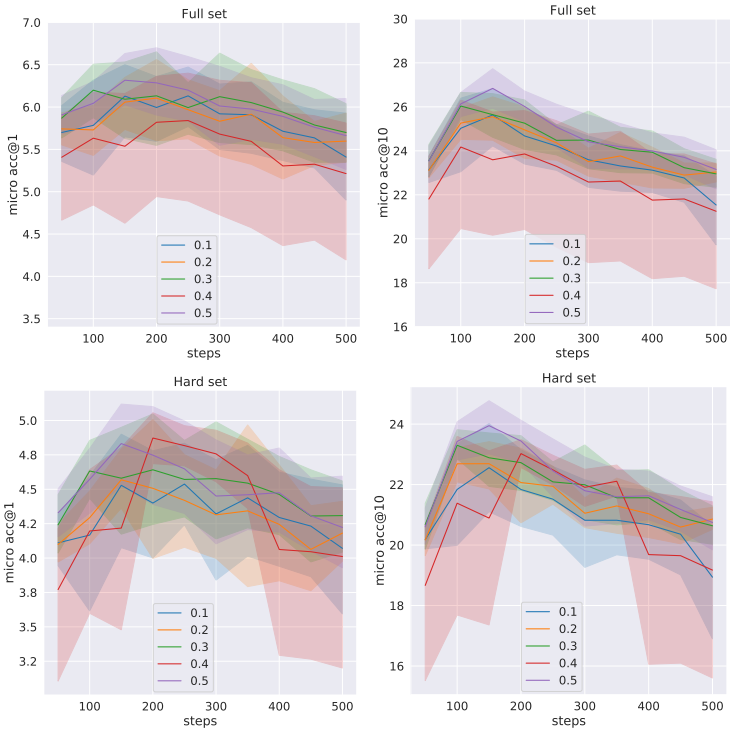

Figure 9: Performance of CONTRASTIVE-PROBE based on PubMedBERT over different mask ratios. The shaded regions are the standard deviations under 10 different random sentence sets sampled from the PubMed corpus.

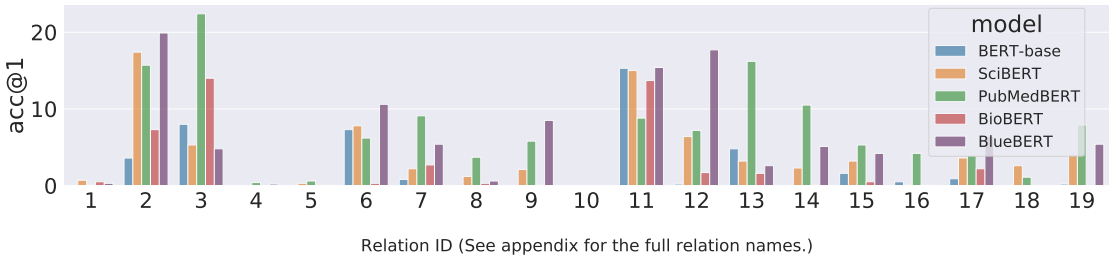

Figure 10: Performance of various PLMs over relations.

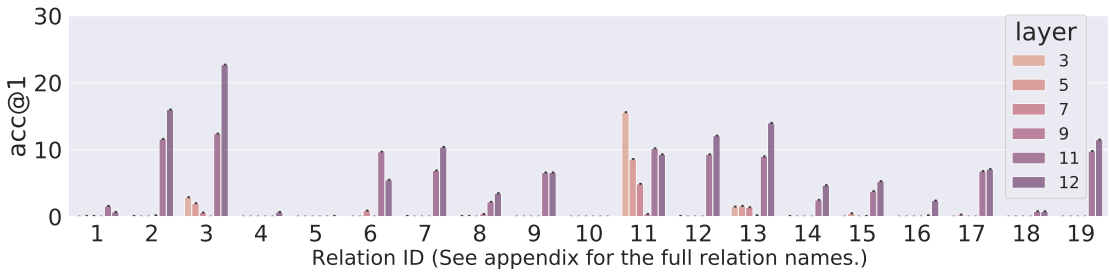

Figure 11: Performance of PubMedBERT under different layers over relations.

ID	Relation	Manual Prompt
1	disease may have associated disease	The disease [X] might have the associated disease [Y] .
2	gene product plays role in biological process	The gene product [X] plays role in biological process [Y] .
3	gene product encoded by gene	The gene product [X] is encoded by gene [Y] .
4	gene product has associated anatomy	The gene product [X] has the associated anatomy [Y] .
5	gene associated with disease	The gene [X] is associatied with disease [Y] .
6	disease has abnormal cell	[X] has the abnormal cell [Y] .
7	occurs after	[X] occurs after [Y] .
8	gene product has biochemical function	[X] has biochemical function [Y] .
9	disease may have molecular abnormality	The disease [X] may have molecular abnormality [Y] .
10	disease has associated anatomic site	The disease [X] can stem from the associated anatomic site [Y] .
11	associated morphology of	[X] is associated morphology of [Y] .
12	disease has normal tissue origin	The disease [X] stems from the normal tissue [Y] .
13	gene encodes gene product	The gene [X] encodes gene product [Y] .
14	has physiologic effect	[X] has physiologic effect of [Y] .
15	may treat	[X] might treat [Y] .
16	disease mapped to gene	The disease [X] is mapped to gene [Y] .
17	may prevent	[X] may be able to prevent [Y] .
18	disease may have finding	[X] may have [Y] .
19	disease has normal cell origin	The disease [X] stems from the normal cell [Y] .

Table 5: The 19 relations and their corresponding manual prompts in MedLAMA.

Benchmark	# Relations	# Queries	Avg. # answer	Avg. # Char	% Single-Tokens
LAMA	41	41k	1	7.11	100%
BioLAMA	36	49k	1	18.40	2.2%
MedLAMA	19	19k	2.3	20.88	2.6%

Table 6: Comparison of LAMA, BioLAMA and our MedLAMA. Note that there is only 1 overlapped relation (i.e. may treat) between BioLAMA and our MedLAMA, and no same query between them.

Probe	Model	CTD		wikidata		UMLS	
		acc@1	acc@5	acc@1	acc@5	acc@1	acc@5
Mask Predict	BERT	0.06	1.20	1.16	6.04	0.82	1.99
	BioBERT	0.42	3.25	3.67	11.20	1.16	3.82
	Bio-LM	1.17	7.30	**11.97**	25.92	3.44	8.88
OptiPrompt	BERT	3.56	6.97	3.29	8.13	1.44	3.65
	BioBERT	**4.82**	9.74	4.21	12.91	5.08	13.28
	Bio-LM	2.99	**10.19**	10.60	25.15	**8.25**	**20.19**
Contrastive-Probe	BlueBERT	1.62	5.84	6.64	25.97	2.63	11.46
	BioBERT	0.20	0.99	1.04	4.51	0.89	3.89
	Bio-LM	1.70	4.26	4.32	18.74	1.27	5.01
	PubMedBERT	2.60	8.87	10.20	**35.14**	4.93	18.33

Table 7: Performance on BioLAMA benchmark. Note that both the mask predict and opti-prompt require using the MLM head and opti-prompt needs further training data, so it is impossible to compare a model without MLM head being released (e.g. PubMedBERT). In contrast, our Contrastive-Probe provides a good indicator of comparing these models in terms of their captured knowledge.

Hyperparameters	Search space
rewire training learning rate	{1e-5, 2e-5*, 5e-5}
rewire training steps	500
rewire training mask ratio	{0.1, 0.2, 0.3, 0.4*, 0.5*}
τ in InfoNCE of rewire training	{0.02,0.03*,0.04,0.05}
rewire training data size	{1k, 10k*, 20k,100k}
step of checkpoint for probing	{50, 150*, 200, 250}
max_seq_length of tokeniser for queries	50
max_seq_length of tokeniser for answers	25

Table 8: Hyperparameters along with their search grid. $*$ marks the values used to obtain the reported results.

Query 1: The gene product *HLA Class II Histocompatibility Antigen, DP(W4) Beta Chain* is encoded by gene [Y] . **UMLS Answers:** *MHC Class II Gene*, *HLA-DPB1 Gene*, *Immunoprotein Gene* **Human Answers:** *MHC Class II Gene*, *HLA-DPB1 Gene*			
Model	**CONTRASTIVE-PROBE (PubMedBERT)**	**X-FACTR (BlueBERT)**	**Generative PLMs (SciFive-large)**
Top-5	*MHC Class II Gene*	*b*	*HLA-DRB1*
	MHC Class I Gene	*hla*	*encoding HLA*
	HLA-A Gene	*dqb1*	*DP(W)*
	HLA-DPB1 Gene	*locus dqb1*	*HLA-B*
	HLA-F Gene	*2 , dq beta 2*	*HLA-DQ*
Query 2: The gene product *Tuberin* is encoded by gene [Y] . **UMLS Answers:** *TSC2 Gene*, *Signaling Pathway Gene* **Human Answers:** *TSC2 Gene*, *Tuberin*			
Model	**CONTRASTIVE-PROBE (PubMedBERT)**	**X-FACTR (BlueBERT)**	**Generative PLMs (SciFive-large)**
Top-5	*TSC2 Gene*	*family of tuberins*	*“”*
	SKA2 Gene	*##t1*	*TUB*
	TSPY1 Gene	*symbol tuber*	*Tuberin*
	Tuberin	*(tuber)*	*TUBE*
	TSC1 Gene	*a*	*TUBB*
Query 3: *Refractory Monomorphic Post-Transplant Lymphoproliferative Disorder* may have [Y] . **UMLS Answers:** *Lymphadenopathy*, *Aggressive Clinical Course*, *Extranodal Disease* **Human Answers:** *Early post-transplant lymphoproliferative disorder*, *Lymphoproliferative disorder following transplantation* , *Refractory Polymorphic Post-Transplant Lymphoproliferative Disorder*, *Aggressive Clinical Course*, *Post transplant lymphoproliferative disorder* *Neoplastic Post-Transplant Lymphoproliferative Disorder*, *Refractory Monomorphic Post-Transplant Lymphoproliferative Disorder*			
Model	**CONTRASTIVE-PROBE (PubMedBERT)**	**X-FACTR (BlueBERT)**	**Generative PLMs (SciFive-large)**
Top-5	*Early post-transplant lymphoproliferative disorder*	*manifestations*	*similar to this*
	Lymphoproliferative disorder following transplantation	*relapses*	*in this study*
	Refractory Polymorphic Post-Transplant Lymphoproliferative Disorder	*phenotype*	*similar to our case*
	Aggressive Clinical Course	*- specific phenotype*	*similar to ours*
	Post transplant lymphoproliferative disorder	*features*	*similar to this case*
Query 4: *moexipril* might treat [Y] . **UMLS Answers:** *Diabetic Nephropathies*, *Heart Failure*, *Hypertension*, *Ventricular Dysfunction, Left* **Human Answers:** *Essential Hypertension*, *Hypertension*			
Model	**CONTRASTIVE-PROBE (PubMedBERT)**	**X-FACTR (BlueBERT)**	**Generative PLMs (SciFive-large)**
Top-5	*Essential Hypertension*	*hypertension*	*“”*
	Posttransplant hyperlipidemia	*diabetes mellitus*	*this*
	Hypertension	*essential hypertension*	*them*
	Atherosclerotic Cardiovascular Disease	*diabetes*	*migraine*
	Type 1 Diabetes Mellitus	*in patients with hypertension*	*patients*

Table 9: Example predictions of different probing approaches. The human answers are annotated based on the CONTRASTIVE-PROBE predictions.

Type	Model	Full set		Hard set	
		macro acc@1/@10	micro acc@1/@10	macro acc@1/@10	micro acc@1/@10
Pure pre-trained	BERT (Devlin et al., 2019)	0.94/4.96	0.94/4.96	0.74/2.31	0.68/2.13
	BlueBERT (Peng et al., 2019)	5.20/20.06	5.20/20.06	4.34/18.91	4.49/18.51
	BioBERT (Lee et al., 2020)	4.54/19.59	4.54/19.59	3.84/15.01	3.90/14.91
	ClinicalBERT (Huang et al., 2019)	1.03/5.61	1.03/5.61	0.66/4.38	0.61/3.74
	SciBERT (Beltagy et al., 2019)	4.17/16.90	4.17/16.90	2.83/14.78	2.81/14.08
	PubMedBERT (Gu et al., 2020)	7.32/**27.70**	7.32/**27.70**	4.86/**23.51**	4.71/**22.88**
Knowledge-enhanced	SapBERT (Liu et al., 2021a)	4.65/24.53	4.65/24.53	2.02/21.78	1.96/18.98
	CoderBERT (Yuan et al., 2020)	**8.38**/24.05	**8.38**/24.05	**6.40**/21.79	**6.24**/20.32
	UmlsBERT (Michalopoulos et al., 2021)	1.68/7.19	1.68/7.19	0.94/5.68	0.85/5.00

Table 10: Benchmarking biomedical PLMs on MedLAMA via our CONTRASTIVE-PROBE.