Computational Linguistics in the Netherlands Journal 14 (2025) 67-84 Submitted 12/2024; Published 07/2025

Bag of Lies: Robustness in Continuous Pre-training BERT

Ine Gevers* INE.GEVERS@UANTWERPEN.BE
Walter Daelemans* WALTER.DAELEMANS@UANTWERPEN.BE

* *University of Antwerp, CLiPS, Lange Winkelstraat 40, Belgium*

Abstract

This study aims to acquire more insights into the continuous pre-training phase of BERT regarding entity knowledge, using the COVID-19 pandemic as a case study. Specifically, we focus on to what extent entity knowledge can be acquired through continuous pre-training, and how robust this process is. Since the pandemic emerged after the last update of BERT's pre-training data, the model has little to no prior entity knowledge about COVID-19. Using continuous pre-training, we control what entity knowledge is available to the model. We use a fact-checking benchmark about the entity, namely Check-COVID, as an evaluative framework, comparing a baseline BERT model with continuous pre-trained variants on this task. To test the robustness of continuous pre-training, we experiment with several adversarial methods to manipulate the input data, such as using misinformation and shuffling the word order until the input becomes nonsensical. Our findings reveal that these methods do not degrade, and sometimes even improve, the model's downstream performance. This suggests that continuous pre-training of BERT is robust against these attacks, but that BERT obtaining entity-specific knowledge is susceptible to writing style changes in the data. Furthermore, we are releasing a new dataset, consisting of original texts from academic publications in the LitCovid repository and their AI-generated (false) counterparts.

1. Introduction

While pre-trained Large Language Models achieve remarkable results on a variety of downstream tasks, it is also known that their performance decreases when they are applied to tasks relying on information outside of the scope of their original pre-training distribution (Oren et al. 2019). Standard practice to alleviate this issue is to continue pre-training the models (i.e., Masked Language Modeling on large unlabeled text data sets) before fine-tuning (i.e., using smaller labeled task-specific data after pre-training). This technique aims at bridging the gap between a model's original knowledge and the specialized information required for its current application. For instance, continuous pre-training (CPT) has shown its merits for specialised in-domain applications (e.g., Lee et al. (2020), Chalkidis et al. (2020)).

We aim to explore to what extent BERT can learn entity knowledge about topics diverging from its original pre-training data through CPT, and how robust this process is. Focusing on a case study enables us to isolate and examine the specific impact of CPT on entity knowledge, sidestepping the confounding factors that often complicate such analyses.
In this case study, we focus on entity knowledge regarding the COVID-19 pandemic, a topic that emerged after the end of BERT's initial pre-training phase. Although BERT's original dataset may include abstract knowledge about viruses and diseases, the specifics of the COVID-19 pandemic present a novel challenge. By leveraging a COVID-19 fact-checking benchmark, Check-COVID (Wang et al. 2023), as our evaluative framework, we aim to shed light on questions surrounding the stability and robustness of knowledge acquisition during CPT.

Our methodology examines various factors that could influence the efficacy of CPT, including the size of the data set (cf. Rietzler et al. (2020)), the veracity of information, the source of the data, the degree to which the training data is aligned with the task data (cf. Gururangan et al. (2020)), the word order within the data, and model size (regarding data memorization, cf. Kharitonov et al. (2021)). The original data sources we examine are academic publications (from the LitCovid repository (Chen et al. 2022)), task-adaptive data (from the fact-checking benchmark Check-COVID (Wang et al. 2023)), and social media data (from Reddit).

We employ two adversarial techniques to manipulate the original input data: (1) misinformation; and (2) shuffling the word order. We continue pre-training BERT on the diverse forms of input data related to COVID-19, consequently fine-tuning and evaluating the models' performance on the Check-COVID benchmark. Figure 1 illustrates the experimental setup described in this study. Among the key findings of our study are the positive effects of CPT on downstream performance, and the surprising robustness of the models against adversarial techniques, with certain adversarial inputs even enhancing the model's performance. We observe that learning entity knowledge through CPT is susceptible to writing style changes, and using data with a simpler writing style during CPT yields the best downstream results. We release the dataset[1] we created for this purpose, which contains the texts extracted from academic publications in LitCovid paired with their AI-generated misinformation and AI-generated paraphrasing.

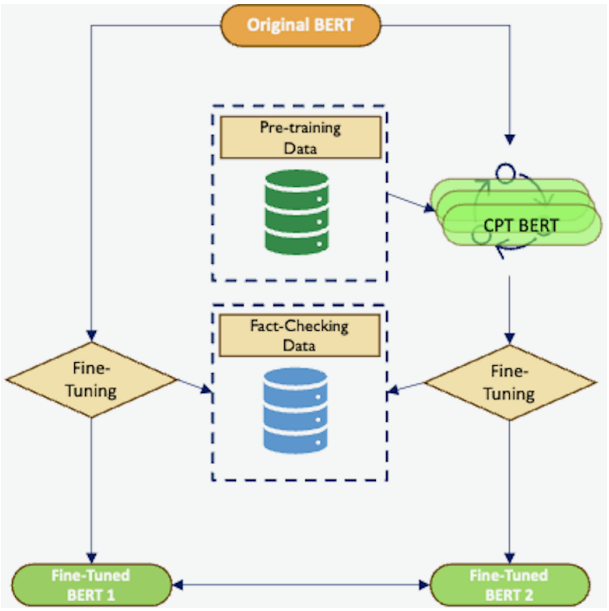

Figure 1: Illustration of the model architecture. We compare a baseline BERT model, fine-tuned on Check-COVID, to BERT models continuously pre-trained on different datasets. After fine-tuning the resulting CPT BERT models on the same Check-COVID benchmark, we compare the performance of the models on the fact-checking task.

The rest of the paper is structured as follows. In Section 2, we start with an overview of the related work concerning continuous pre-training and model evaluations. In Section 3, we discuss our research questions and hypotheses, and in Section 4 we dive into the methodology concerning the data curation and experiments. Further on, in Section 5, we present the results of our analyses. The final Section 6 concludes our research, giving an overview of the findings and suggestions for further research.

2. Related Work

Knowing what information is used by models at inference time is crucial to ascertain the model's trustworthiness and ability to generalize. However, uncovering this is a major challenge: a whole field of study, explainable AI, is dedicated entirely to studying the inner working of transformers (e.g., BERTology for BERT (Rogers et al. 2021)). While this field is concerned with a variety of knowledge, such as syntactic, semantic, or world knowledge, others are mainly interested in uncovering what factual information is used by BERT.

1. Zenodo url: `https://zenodo.org/records/15055492`, DOI: 10.5281/zenodo.15055492

For instance, using Masked Language Modeling (MLM) probing, Penha and Hauff (2020) observe that BERT stores knowledge about books, movies, and music in its parameters. Further, Podkorytov et al. (2021) analyse the internal components of BERT that are responsible for the output, to measure factual knowledge present in the transformer model and its ability for generalization in downstream tasks. They find that BERT's knowledge is fragile, and based on token co-occurrence in the pre-training data. Conversely, Petroni et al. (2019) convert factual triplets (subject-relation-object) into prompts to probe factual knowledge and find that BERT has a strong ability to recall factual information. However, Guimarães et al. (2024) find that this approach is not fool-proof, since adding negation to the prompt distorts the results.
As can be seen, there is some discrepancy in the reported results about how and what factual knowledge is stored in BERT. This could be an effect of the variety of evaluation metrics used to measure knowledge in BERT. Previous research has used both intrinsic and extrinsic evaluations. Intrinsic variants include fill-in the gap probes in Masked Language Modeling (MLM); using self-attention weights; and probing classifiers using different BERT representations as input (Rogers et al. 2021). Extrinsic evaluations on downstream NLP tasks have been carried out on benchmarks such as CREAK (Onoe et al. 2021) which tests for entity knowledge. We propose fact-checking tasks as an additional extrinsic evaluation.

While results do not always agree about the extent to which knowledge is reliably incorporated in the parameters of BERT, it is generally agreed that the textual data used to (pre-)train the transformer model has a significant role in the acquisition of that knowledge. Using large language models' perplexity on masked spans in texts about entities that are excluded from the original pre-training data, Onoe et al. (2022) demonstrate that models struggle with making inferences about unseen entities, from which can be derived that the knowledge about these entities is limited.
Updating the available knowledge, then, is fundamental to improving the models' performance. However, pre-training the model from scratch is computationally expensive and time-consuming (Lamproudis et al. 2021), but relying on extensive fine-tuning can lead to catastrophic forgetting (Chen et al. 2020). Therefore, intermediate techniques such as continuous pre-training (CPT) could alleviate this issue (Cossu et al. 2022).
CPT is especially important for events that took place after the last update of the model, or for specific topics that are not well represented in the original training data (e.g., biomedical data: BioBERT (Lee et al. 2020), legal: LEGAL-BERT (Chalkidis et al. 2020)). Lemmens et al. (2022) continue pre-training the Dutch RobBERT model on COVID-19 related Tweets, and show that the resulting model outperforms the original model on vaccine hesitancy detection. Additionally, Gururangan et al. (2020) show that task-adaptive pre-training, in which the downstream task's unlabeled data is used for CPT, is a promising method compared to using large amounts of in-domain data. However, the amount of data needed for CPT is domain-dependent (Rietzler et al. 2020). Most research uses human-generated data as input to pre-train transformer models, but given the increasing rise of generative language models such as the GPT-family, some research has experimented with using AI-generated data, showcasing its effectiveness (Eldan and Li 2023).
Although CPT is now standard practice, questions persist about its effectiveness and optimal configurations (Bacco et al. 2023). Also, the stability of the process has been questioned since even one sentence can alter the model's downstream performance (Bacco et al. 2023), and fine-tuning the model on a large dataset can obliterate the effects of CPT (Zhu et al. 2021).
As such, efforts have been made to test the boundaries by exploring adversarial techniques. Literature shows that using nonsensical input texts (i.e., randomly selected n-grams, non-human language, or different word order) in CPT does not lead to worse results (Chiang and Lee 2020, Krishna et al. 2021, Sinha et al. 2021). It is hypothesized that pre-training mainly teaches the model hierarchical structures, long-distance dependencies, and higher-order word co-occurrences, for which the distributional information of the input text is enough. Some research notes that adding noise to the input data, which is argued to encourage the diversity of the embedding vectors, even helps downstream performance (Wang et al. 2019).

However, to the best of our knowledge, no one experimented with using factually incorrect data as a confounding factor, which is especially relevant for fact-checking tasks: models' pre-training data could include unverified information, so if misinformation influences the models' output, this needs to be addressed.

3. Research questions and hypotheses

This study is guided by the following research questions and hypotheses:

1. **Does BERT utilize entity knowledge for fact verification?** We investigate whether BERT is capable of leveraging specific entity knowledge to verify facts within given statements during fine-tuning. To measure this, we focus on entity knowledge that was not present in the original pre-training data, but introduced during the CPT phase. We hypothesize that BERT can benefit from new entity knowledge seen during CPT to perform fact verification tasks. Additionally, task-adaptive pre-training will enhance performance: using CPT data that is more aligned with the specific language use of the downstream task is beneficial (Gururangan et al. 2020).

2. **Is the veracity of that entity knowledge important for the accuracy of fact verification by BERT?** This question aims to understand the impact of the truthfulness of the provided entity knowledge on the model's performance. We expect the presence of erroneous information to negatively affect the model's ability to accurately verify facts during fine-tuning. In like manner, using questionable sources such as Reddit as input data will also decrease performance.

3. **How robust is the CPT phase?** We examine whether CPT still helps performance on fact verification when the input data is manipulated to confuse the model (i.e., misinformation, shuffled word order). We hypothesize that the CPT phase is not robust when it comes to misinformation (as mentioned above), but in accordance with prior work, we assume that the use of nonsensical data (i.e., shuffled word order) should not decrease the results on downstream tasks (Sinha et al. 2021, Krishna et al. 2021, Chiang and Lee 2020). Following previous literature, we assume that small amounts of CPT data will already show differences in downstream tasks (see Bacco et al. (2023)), but we hypothesize that larger pre-training datasets will make these effects more robust, with more correct information leading to better performance and more incorrect information resulting in worse outcomes.

It is important to note that the primary goal of this study is not to surpass the current state-of-the-art (SOTA) models in fact-checking: we do not expect that by CPT alone we could match the current SOTA. Rather, we focus on the effect of adding new entity knowledge in the CPT phase, and a fact-checking setup gives us a controlled environment to evaluate the importance of this entity knowledge in a downstream task that revolves around entity knowledge. However, insights gained from this research could assist other techniques focused on improving fact-checking performances. This includes understanding the potential benefits of using limited pre-training data, evaluating the significance of the data source for pre-training (e.g., the use of AI-generated data and data from social media platforms), the impact of using misinformation during CPT, the possible gains from task-adaptive pre-training, and evaluating the process' robustness by manipulating the word order of the input data.

4. Methods

4.1 Data

We use various sources to create the input data to continue pre-training BERT. In this section, we describe these sources, and the subsequent transformations we applied to test the robustness of the CPT process.

Figure 2 demonstrates the implemented procedures through examples. Additionally, we describe the benchmark we use to fine-tune and evaluate the resulting CPT models on.

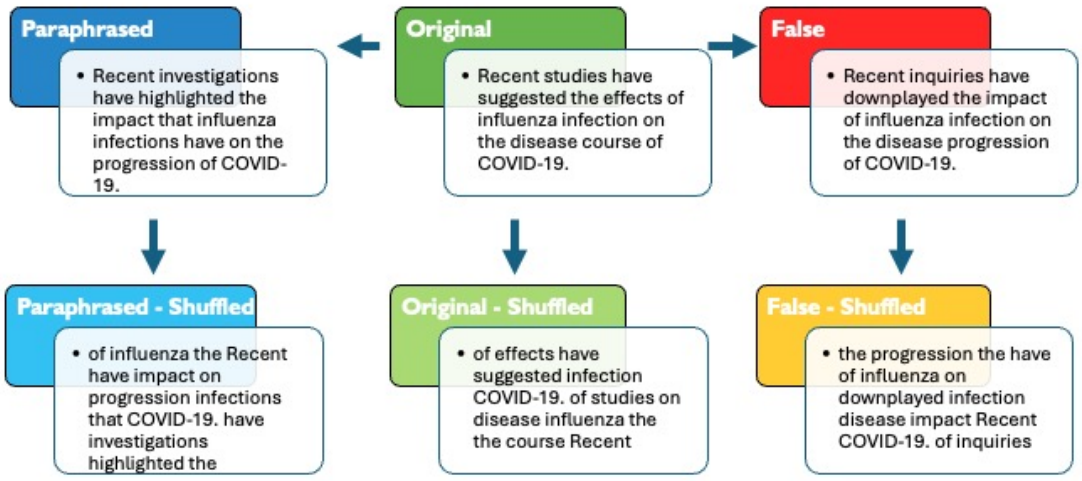

Figure 2: Illustration of adversarial transformations of the input text. Using GPT-4, we paraphrase and falsify the original text from LitCovid. Then, we shuffle the word order in each text.

As a starting point, we extract texts from the LitCovid repository, which contains academic publications about the COVID-19 pandemic (Chen et al. 2022). These texts, following the structure of the dataset, can be titles, abstracts, or entire paragraphs from original academic publications. For more details on the curation of this repository, we refer to the original publication: Chen et al. (2022). To verify the impact of the data size used for CPT, we compare a model that is continuously pre-trained on 200 text units with one on 10,000 texts.[2] On average, each text has around 120 words.
Following Gururangan et al. (2020), we also implement task-adaptive pre-training. For this purpose, we use the unlabeled text data from the fact-checking benchmark as input during CPT, prior to fine-tuning the model on the labeled data. We use the entire dataset to CPT BERT, because the dataset is rather small (1,000 instances). Since the labels are not seen during CPT, half of the information is factually correct, the other half incorrect; if factuality in CPT data plays a role, the model performance should be around random chance. To verify that topical vocabulary plays a role, we add a baseline model with CPT on similar task data, but other topics. To this end, we gather data from the fact-checking benchmarks Liar (containing statements from PolitiFact.com) (Wang 2017) and VitaminC (based on Wikipedia revisions) (Schuster et al. 2021), and we filter out all data points that mention 'covid', 'pandemic', 'corona', 'covid-19', or 'coronavirus'.

2. We experimented with larger sizes up to 1 million texts from the original LitCovid data, but we focus our discussion on the 200 and 10,000 variants for two main reasons. First, the model's performance plateaued early: incremental increases of 500 texts showed no significant improvement beyond 10,000 texts. Second, because of resource limitations, we could not generate this many texts using GPT-4.

The last original data source we consider in this study is user-generated data on the social media platform Reddit[3]. The dataset contains posts and comments mentioning COVID-19, from which we randomly sample 10,000 comments to be used as input data.
There are two adversarial techniques that we apply to modify the original input data. For the academic texts (derived from LitCovid), we use the GPT-4Turbo API to revert the truthfulness of the text, artificially generating misinformation. To ensure that the text's veracity is important besides the specific AI-generated language use, we also use GPT-4Turbo to paraphrase the original data. For all text sources (i.e., LitCovid, Reddit, task-adaptive, and AI-generated), we shuffle the word order to distort the text. The shuffling is done both inside one line of text as well as for the entire collection of texts.
We calculate various metrics to evaluate the main datasets of this study (i.e., LitCovid, paraphrased, and misinformation for both 200 (small) and 10,000 (large) samples, all of them in original and shuffled version). For the detailed results of each metric, we refer to Appendix A.
First, the smaller datasets and the larger datasets are similar in all aspects, only the perplexity (measured by exponentiating the average negative log-likelihood per token in the text using GPT2) is slightly higher for the larger datasets. Second, comparing the original texts to their shuffled variants, we note that the readability[4] is higher, but this is likely because disrupting the word order alters local syllable distributions and punctuation patterns that make the sentence structures seemingly simpler; the formula does not measure coherence or meaning. This hypothesis is corroborated by a higher perplexity score for the shuffled texts. Third, comparing the original LitCovid texts to their AI-generated counterparts (misinformation and paraphrased), we observe that AI-generated texts are generally longer, have a lower readability score, and higher perplexity. A manual analysis of the LitCovid 10K datasets comparing human to AI-generated texts reveals that in 1% of the cases, the original language was not English, and is translated by AI. We note that in cases where the original text consists of a collection of numbers or results, this is presented in more naturally flowing text in the AI-generated texts.

We evaluate the resulting CPT models on the same benchmark, Check-COVID (Wang et al. 2023). The benchmark combines claims from newspaper articles[5], annotated by experts, and evidence from the CORD-19 repository (Wang et al. 2020) to test LLM's fact-verification abilities concerning the COVID-19 pandemic. In total, there are 1,500 claims (evenly distributed over the labels 'support', 'refute', and 'not-enough-information'). For more statistics about the benchmark, please see (Wang et al. 2023). While the approach presented in their research focuses on rationale selection from texts for the task, our own study, in contrast, aims to examine the role of entity knowledge in BERT's CPT data, for which the fact-checking setup offers a good case study. For our purposes, we concentrate on the 'support' and 'refute' labels and exclude the 'not-enough-information' label. To overcome variability issues observed when evaluating on the original test set only, we apply 5-fold cross-validation. As mentioned earlier, we do not compare our results to those reported on the fact-verification task; we do not expect our results to be competitive with this task. Rather, we focus on the relative performance difference between a base model without entity knowledge, and models that have access to entity knowledge through CPT.

We focus on a case study about COVID-19, an entity that is not present in the original pre-training data, for various reasons. Mainly, investigating an entity that is present in the original data would require us to delete parts of BERT's original pre-training data, which could introduce unexpected variance into the model's performance. Alternatively, pre-training a model from scratch excluding the relevant entity

3. `https://www.kaggle.com/datasets/pavellexyr/the-reddit-covid-dataset?select=the-reddit-covid-dataset-comments.csv`
4. Flesch Reading Ease:

$$\text{FRES} = 206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

5. either "extracted": copied verbatim from the newspaper article, or "composed": rewritten claim by an annotator based on the content of the newspaper article

knowledge is both time-intensive and resource-heavy, making it an impractical approach for our research objectives.

4.2 Experiments

Similar to previous work (e.g., Bacco et al. (2023)), we focus on the BERT-base model (Devlin et al. 2019). We deliberately choose a smaller, older model, because it has been used in previous research concerning CPT, making it easier to situate and compare our work. Additionally, the smaller size allows us to run our experiments locally in a manageable time frame.
We run three baseline models to measure the influence and robustness of entity knowledge added during CPT. Specifically, we use the BERT-base model[6], an in-domain pre-trained model BioBERT (Lee et al. 2020), and a BERT-base model CPT on task-adaptive but non-COVID topics (following Gururangan et al. (2020)). We use a fact-checking setup because it is a good case to measure the model's ability to learn from data about an entity inserted in CPT. The binary classification ("support" and "refute") offers a straightforward way to evaluate models' performance. By keeping the rest of the setup identical, and by only changing the input data for CPT, we can pinpoint the relative impact of this entity knowledge on the model's performance. We highlight that since we are interested in the effect of entity knowledge during CPT, we focus on the relative performance difference between a baseline BERT model to BERT CPT on different dataset variants. We make no claims about differences between BERT models CPT on different dataset variants, since this would introduce additional confounding factors that are out of the scope of this research.

4.2.1 Continue pre-training BERT

As mentioned before, we do our main experiments with BERT-base, but given that model size is an important variable for data memorization (cf. Kharitonov et al. (2021)), we add additional results with BERT-large[7],[8]. For model specifications, we refer to Appendix D.1. After further pre-training, we fine-tune and evaluate the resulting models on the down-stream task. We report results across 5 random seeds in 5-fold cross-validation, indicating the average performance as well as the standard deviation across the seeds.

4.2.2 Prompt GPT API to revert truthfulness and paraphrase

In order to create the adversarial data, for which we artificially generate misinformation, we leverage the OpenAI GPT-4Turbo API.[9] We use the original texts extracted from the LitCovid repository as input. We experimented with several variations of the prompt. For transparency, we add the final version of the prompt in Appendix B. Since simple negations such as "not" are known to confuse BERT-base models (Truong et al. 2023), we take care to instruct the generative model to go beyond simple negations: we include this specifically in the prompt, and we found that by increasing the temperature setting, the model complies better with this demand. We use the gpt-4-turbo-preview[10] model, setting the model temperature to 0.5. We keep the output length approximately as long as the original input text[11] and we manually verify the quality of a sample of the generated output. We use the same model and technique to generate paraphrased data from the original texts, the prompt can be found in Appendix C. We release

6. `https://huggingface.co/google-bert/bert-base-uncased`
7. `https://huggingface.co/google-bert/bert-large-uncased`
8. We cannot compare with BioBERT, since this model only has BERT-base as underlying model.
9. The cost of generating the misinformation dataset was $148.30, the paraphrased dataset $111.96.
10. `https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo`, model date: 25/01/2024
11. We use the tiktoken package (`https://pypi.org/project/tiktoken/0.1.1/`) to estimate the number of input tokens.

the resulting paired dataset (i.e., the original texts sampled from the academic publications in LitCovid (Chen et al. 2022), and the generated counterparts).[12]

5. Results and discussion

Model	BERT-base			BERT-large		
	Macro F1	**Precision**	**Recall**	**Macro F1**	**Precision**	**Recall**
BERT-baseline	61.76 (±3.52)	63.80	62.56	65.56 (±2.88)	68.24	66.40
BioBERT	66.48 (±3.43)*	67.36	66.72	/	/	/
Task-adaptive other topic	62.64 (±2.16)	64.64	63.60	63.16 (±2.95)	65.84	64.32
Task-adaptive Check-COVID	64.96 (±2.73)*	66.96	65.76	65.04 (±3.45)	66.48	65.88
Task-adaptive Check-COVID shuffled	62.60 (±4.67)	65.60	64.12	65.00 (±1.61)	66.44	65.64
LitCovid200 (True)	58.76 (±3.48)	62.04	60.12	65.12 (±2.49)	66.72	65.96
LitCovid200 (True) shuffled	64.60 (±3.03)*	66.56	65.20	64.88 (±2.99)	66.52	65.56
LitCovid200 (Paraphrased)	64.92 (±2.99)*	67.04	65.64	64.88 (±2.36)	66.56	65.52
LitCovid200 (Paraphrased) shuffled	64.52 (±2.58)*	66.48	65.20	65.08 (±2.74)	66.80	65.76
LitCovid200 (False)	64.80 (±2.76)*	66.40	65.32	63.56 (±2.57)	65.56	64.56
LitCovid200 (False) shuffled	64.16 (±2.61)*	66.00	65.08	64.48 (±2.57)	67.00	65.40
LitCovid10K (True)	60.72 (±2.40)	62.28	61.40	66.28 (±1.42)	67.76	66.76
LitCovid10K (True) shuffled	63.24 (±5.99)*	65.12	64.28	64.08 (±7.27)	65.72	65.44
LitCovid10K (Paraphrased)	64.32 (±2.99)*	66.56	65.36	65.64 (±1.86)	67.00	66.08
LitCovid10K (Paraphrased) shuffled	65.08 (±1.12)*	67.12	65.76	66.56 (±2.92)	68.60	67.28
LitCovid10K (False)	63.88 (±3.04)*	65.64	64.76	66.00 (±1.41)	67.36	66.44
LitCovid10K (False) shuffled	64.00 (±0.98)*	66.40	65.12	64.28 (±6.41)	66.88	65.44
Reddit	64.56 (±2.71)*	67.16	65.52	65.48 (±1.91)	66.36	65.92
Reddit shuffled	64.60 (±4.57)*	66.92	65.56	65.72 (±4.02)	67.88	66.68

Table 1: Model performance on Check-COVID. We compare three baseline models (BERT, BioBERT, task-adaptive model on other topics) with the CPT models. We report the average result of the models in 5-fold cross-validation across 5 random seeds. We indicate the relative standard deviation for the Macro F1-score across the seeds. '*' denotes whether the difference in macro F1 performance from the baseline BERT model is statistically significant. Following prior work, we use the McNemar test for this purpose. If $\alpha < 0.05$, we can assume that the model is significantly different from the baseline. The effect sizes (calculated with Cohen's g) are small to medium. For the exact measures, we refer to Appendix E.

We report the performance of the models on Check-COVID in Table 1. We focus first on BERT-base, since this is also the model discussed in prior literature. Our analysis shows that CPT generally helps performance on this downstream task, which confirms prior literature. Specifically, CPT on in-domain (i.e., biomedical) data (Lee et al. 2020) and task-adaptive pre-training (Gururangan et al. 2020) significantly improve downstream performance. However, improvements with task-adaptive pre-training are contingent on the data being topically aligned. Simply having similar data from tasks, but about different topics, is insufficient to gain significant improvements.

12. The full dataset is available on Zenodo, url: `https://zenodo.org/records/15055492`, DOI: 10.5281/zenodo.15055492.

Surprisingly, using accurate information extracted from academic publications (LitCovid True) does not improve performance, and this holds true even when larger datasets are employed. However, the BERT-base models CPT on the generated misinformation (LitCovid False, both on the smaller and larger datasets) significantly improved compared to the baseline model and the models CPT on the original academic texts (LitCovid True). Also the models CPT on AI-generated paraphrased data (LitCovid Paraphrased) improve. So, using AI-generated text helps in this context, but the veracity of the text plays no role: there is no significant difference between BERT-base models CPT on correct or incorrect AI-generated text. We performed a manual qualitative error analysis on a sample of the test data, comparing the models' output to each other and the gold standard, but this did not reveal any distinct patterns.

We hypothesize that this can be explained by the language use of AI-generated texts being more diverse, which helps BERT to learn the relevant patterns (Eldan and Li 2023). Also, the data analysis in Section 4.1 showed that AI-generated texts have a higher perplexity compared to the original human generated text, which could be a reason why the BERT-base model performs better with this data: this more complex data could teach BERT richer, more robust language representations that generalize better. An alternative explanation could be that there is not yet enough data provided during CPT in our experiments, and that with more data a breaking point will be observed (i.e., with larger data size, incorrect data will eventually lead to worse performance, and correct data to better performance).[13] We leave this for future research to explore.

To verify that not only AI-generated language is responsible for this remarkable result, we use Reddit comments about COVID-19 -of which the veracity of the content can be questioned- as input data. We observe that also in this setup, the resulting model significantly outperforms the baseline model. This unexpected finding could lead to further research in domains with restricted data availability: user-generated data from social media platforms are generally omitted in these contexts exactly because of their questionable veracity and quality, but this result could indicate that including social media data is a viable option. We hypothesize that the writing style of the Reddit comments is more central in the model's trainingdata representation, and thus being more familiar to the model, compared to the academic writing in LitCovid.

Consistent with earlier studies, the shuffling of word order does not significantly affect downstream performance in most scenarios. However, we observe a notable exception: in instances where CPT does not yield improvements over the baseline model (i.e., when correct information from academic texts is used, LitCovid True), the shuffling of this data leads to improved performance. Additionally, when the two adversarial attacks used in this study (i.e., misinformation and shuffling word order) are combined, we observe that the resulting models still outperform the baseline. This could suggest that CPT is rather robust, and primarily focuses on learning associations on document level. In the cases where the original data does not outperform the baseline model without CPT, which could occur because the language or writing style of the CPT-data deviates significantly from the language used in the task (as described in Gururangan et al. (2020)), reformatting the texts using generative AI techniques and/or shuffling the word order can potentially aid the model to generalize.

We repeated our experiments using BERT-large to measure the effect of model size. First, there is no significant difference in performance between the BERT-large based CPT models: CPT does not bring improvements, but adversarial attacks also do not degrade model performance. This is unexpected, since larger models are generally associated with more data memorization (Kharitonov et al. 2021), which could have lead to models CPT on incorrect information performing worse. Second, there are no significant differences when comparing the paired BERT-base and BERT-large models (e.g., BERT-base CPT on Reddit and BERT-large CPT on Reddit). Thus, while model size mitigates the effects of CPT, the model is still robust against adversarial attacks.

13. As explained in Section 4.1, we experimented with 1 million original texts, but did not generate as many falsified texts: this remains to be explored in further research.

In summary, we can answer the research questions from Section 3 as follows:

1. **Does BERT utilize entity knowledge for fact verification?** Relevant entity knowledge generally helps BERT's downstream performance. However, we do observe that the language use of the input data should be aligned to the task data: as expected, task-adaptive pre-training yields improvements, but contrary to our initial expectations, using academic texts from LitCovid does not improve results. However, a larger model does not significantly benefit from CPT.

2. **Is the veracity of that entity knowledge important for the accuracy of fact verification by BERT?** No, using misinformation or questionable data sources as input does not degrade the model's performance. On the contrary: despite their questionable nature in terms of veracity, there is a significant improvement compared to the baseline performance for BERT-base. We hypothesize that this is a result of their writing style being more diverse from the model's training data representation (as indicated by a higher perplexity), helping the model to generalize.

3. **How robust is CPT in enhancing the model's ability for fact verification?** We find that CPT is robust against the two adversarial techniques we present in this work (i.e., using misinformation, as described above; and shuffling the word order of the texts), also when combined. Using larger data sizes shows the same conclusions as smaller data sizes.

6. Conclusion

Continuous pre-training has become a standard practice for addressing the limitations of language models for niche or not well-represented areas, or to update a model's information after the initial pre-training. Nevertheless, the stability of this process has been questioned, highlighting the need for further investigation into its reliability and impact on model performance (Bacco et al. 2023). In this study, we examine a specific aspect of CPT by focusing on entity knowledge. While considerable research efforts have investigated in-domain pre-training (e.g., Lee et al. (2020), Chalkidis et al. (2020)), few have looked at entity knowledge: to the best of our knowledge, the benchmark CREAK is one of its kind investigating common sense reasoning over entity knowledge (Onoe et al. 2021). However, we propose using fact-checking benchmarks as a means to assess a model's grasp on entity knowledge.

In this case study, we focus on the COVID-19 pandemic. Since the pandemic emerged after the last update of BERT's pre-training data, the model has little to no entity knowledge about COVID-19. Using CPT, we control what entity knowledge is available to the model. We compare the baseline BERT model with the CPT variants on the fact-checking benchmark Check-COVID (Wang et al. 2023). We compare three baseline models (i.e., a vanilla BERT (Devlin et al. 2019), an in-domain pre-trained BioBERT (Lee et al. 2020), and a task-adaptive model on other topics (based on (Gururangan et al. 2020)) with BERT models CPT on relevant entity knowledge. For this, we use three data sources: academic publications (the LitCovid repository (Chen et al. 2022)), task data (the unlabeled texts from CheckCovid (Wang et al. 2023)), and social media (Reddit). Further, we compare performance of two model sizes: BERT-base and BERT-large.

Since the robustness of the CPT process is sometimes questioned, we explore two adversarial attacks that manipulate this input data: deliberately using misinformation (which we apply to LitCovid, generating misinformation with the OpenAI GPT-4 API), and shuffling the word order (which we apply to all three data sources).

Consequently, we compare the baseline models with the CPT models by fine-tuning and evaluating them on the same fact-checking benchmark Check-COVID. We apply McNemar tests on the models' predictions to confirm significance and cohen's g to report effect size. A manual error analysis on a sample of the models' output did not reveal any distinct patterns.

Surprisingly, our findings indicate that the veracity of the text is not an important factor. BERT-base

models CPT on AI-generated data perform better than BERT-base models CPT on original correct information, but there is no significant difference whether that AI-generated data is correct (paraphrased) or incorrect (misinformation). Additionally, using a source of questionable content quality (i.e., Reddit) also improves BERT-base performance. Consistent with prior results, shuffling word order has no effect (Chiang and Lee 2020, Krishna et al. 2021, Sinha et al. 2021). However, we note that in the cases where CPT does not lead to improvements on the baseline performance (i.e., when correct information from academic texts is used), shuffling the word order of that data results in significantly better performance. We observe that even when the two adversarial attacks are combined, this does not have a negative effect on the downstream performance. We observe that a larger model size (i.e., BERT-large) is less impacted by CPT, but is still robust against the adversarial attacks.

Looking ahead, we suggest several avenues for further research, including an examination of the internal representations within the models similar to the methods proposed by Bacco et al. (2023) and the utilization of larger CPT datasets. In our approach, we leverage the GPT-4 API to generate paraphrases and misinformation, but there is a potential variability and superficiality in the outputs. Future research might benefit from creating texts manually to compare with the automated outputs of GPT-4. Further, this case study demonstrates that user-generated data from social media platforms, despite their questionable veracity and quality, can be used as input data for CPT. This could inspire future work on domains with restricted data availability to use social media data. We also like to point out that this research is conducted on a small controlled case-study, looking at the COVID-19 pandemic, and that generalizations to other topics or domains should be investigated in further research.

7. Limitations

While our work is, to the best of our knowledge, the first to tackle misinformation in CPT, there are certain limitations that were not addressed in this case-study. First, since it is a case-study, this exploratory study is limited to one entity (COVID-19), and one downstream dataset (Check-COVID). This is partly due to a lack of qualitative labeled COVID-19 fact-checking data. After considering the possibilities, we decided to do an in-depth study on one topic, and include other topics in future research. However, it is thus possible that results could vary for other topics. Second, the data size used for CPT is relatively limited (up to 10,000 texts): using more data during CPT could affect the trends observed in this study. However, in our experiments we noted no difference between models CPT on 10,000 or 1 million texts from the original LitCOVID repository.[14] Additionally, we compared model performances using incrementally more texts during CPT (each time, we added 500 texts), but the performance of the models plateaued quickly. Third, in the task-adaptive set-up we use the same dataset during CPT as is used for the cross-validation. Since the labels are not seen during CPT, half of the texts are factually correct and the other half incorrect without the model knowing which ones are correct. While there is a potential concern regarding overfitting, the unsupervised nature of CPT mitigates this risk by focusing on language patterns rather than label information. However, additional experiments using different subsets could also investigate the data size for task-adaptive pre-training for fact-checking tasks. Fourth, while a fact-checking set-up introduces an indirectness to measure a model's entity knowledge, we argue that this framework is suitable for this case study because of its clear set-up and evaluation metrics. Besides, we focus on the effect of various settings of CPT (e.g., using correct or incorrect information) before we fine-tune on the fact-checking benchmark. While fine-tuning on the fact-checking task could mitigate some of the effects of CPT, we are interested in the observed difference in performance when the data used for CPT is the only changing variable: this set-up therefore gives us a window to explore entity knowledge learnt during CPT. However, to measure

14. We did not go beyond 10,000 texts for the AI-generated counterparts because of resource limitations.

the level of information that is retained by the models from the CPT phase, we suggest that internal analyses, such as probing, would give more insights.

Acknowledgments

This research was made possible with a grant from the Fonds Wetenschappelijk Onderzoek (FWO) project 11P3824N.

References

Akiba, Takuya, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama (2019), Optuna: A next-generation hyperparameter optimization framework, *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Bacco, Luca, Gosse Minnema, Tommaso Caselli, Felice Dell'Orletta, Mario Merone, and Malvina Nissim (2023), On the instability of further pre-training: Does a single sentence matter to BERT?, *Natural Language Processing Journal* **5**, pp. 100037, Elsevier.

Chalkidis, Ilias, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos (2020), LEGAL-BERT: The muppets straight out of law school, *in* Cohn, Trevor, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, Online, pp. 2898–2904. https://aclanthology.org/2020.findings-emnlp.261.

Chen, Qingyu, Alexis Allot, Robert Leaman, Chih-Hsuan Wei, Elaheh Aghaarabi, John J Guerrerio, Lilly Xu, and Zhiyong Lu (2022), LitCovid in 2022: an information resource for the COVID-19 literature, *Nucleic Acids Research* **51** (D1), pp. D1512–D1518. https://doi.org/10.1093/nar/gkac1005.

Chen, Sanyuan, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu (2020), Recall and learn: Fine-tuning deep pretrained language models with less forgetting, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Association for Computational Linguistics, Online, pp. 7870–7881. https://www.aclweb.org/anthology/2020.emnlp-main.634.

Chiang, Cheng-Han and Hung-yi Lee (2020), Pre-training a language model without human language, *arXiv preprint arXiv:2012.11995*.

Cossu, Andrea, Antonio Carta, Lucia Passaro, Vincenzo Lomonaco, Tinne Tuytelaars, and Davide Bacciu (2022), Continual pre-training mitigates forgetting in language and vision, *Available at SSRN 4495233*.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019), BERT: Pre-training of deep bidirectional transformers for language understanding, *in* Burstein, Jill, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, pp. 4171–4186. https://aclanthology.org/N19-1423.

Eldan, Ronen and Yuanzhi Li (2023), Tinystories: How small can language models be and still speak coherent English?, *arXiv preprint arXiv:2305.07759*.

Guimarães, Nuno, Ricardo Campos, and Alípio Jorge (2024), Pre-trained language models: What do they know?, *WIREs Data Mining and Knowledge Discovery* **14** (1), pp. e1518. https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1518.

Gururangan, Suchin, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith (2020), Don't stop pretraining: Adapt language models to domains and tasks, *in* Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online, pp. 8342–8360. https://aclanthology.org/2020.acl-main.740.

Kharitonov, Eugene, Marco Baroni, and Dieuwke Hupkes (2021), How BPE affects memorization in transformers, *arXiv e-prints* pp. arXiv–2110.

Krishna, Kundan, Jeffrey Bigham, and Zachary C. Lipton (2021), Does pretraining for summarization require knowledge transfer?, *in* Moens, Marie-Francine, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, Association for Computational Linguistics, Punta Cana, Dominican Republic, pp. 3178–3189. https://aclanthology.org/2021.findings-emnlp.273.

Lamproudis, Anastasios, Aron Henriksson, and Hercules Dalianis (2021), Developing a clinical language model for swedish: continued pretraining of generic BERT with in-domain data, *International Conference Recent Advances in Natural Language Processing (RANLP'21), online, September 1-3, 2021*, INCOMA Ltd., pp. 790–797.

Lee, Jinhyuk, 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), pp. 1234–1240, Oxford University Press.

Lemmens, Jens, Jens Van Nooten, Tim Kreutz, and Walter Daelemans (2022), CoNTACT: A Dutch COVID-19 adapted BERT for vaccine hesitancy and argumentation detection, *in* Calzolari, Nicoletta, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, pp. 6837–6845. https://aclanthology.org/2022.coling-1.595.

Onoe, Yasumasa, Michael JQ Zhang, Eunsol Choi, and Greg Durrett (2021), CREAK: A dataset for commonsense reasoning over entity knowledge, *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Onoe, Yasumasa, Michael Zhang, Eunsol Choi, and Greg Durrett (2022), Entity cloze by date: What LMs know about unseen entities, *in* Carpuat, Marine, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Findings of the Association for Computational Linguistics: NAACL 2022*, Association for Computational Linguistics, Seattle, United States, pp. 693–702. https://aclanthology.org/2022.findings-naacl.52.

Oren, Yonatan, Shiori Sagawa, Tatsunori B. Hashimoto, and Percy Liang (2019), Distributionally robust language modeling, *in* Inui, Kentaro, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, pp. 4227–4237. https://aclanthology.org/D19-1432.

Penha, Gustavo and Claudia Hauff (2020), What does BERT know about books, movies and music? probing BERT for conversational recommendation, *Proceedings of the 14th ACM conference on recommender systems*, pp. 388–397.

Petroni, Fabio, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander H Miller, and Sebastian Riedel (2019), Language models as knowledge bases?, *arXiv preprint arXiv:1909.01066*.

Podkorytov, Maksim, Daniel Biś, and Xiuwen Liu (2021), How can the [MASK] know? the sources and limitations of knowledge in BERT, *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp. 1–8.

Rietzler, Alexander, Sebastian Stabinger, Paul Opitz, and Stefan Engl (2020), Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification, *in* Calzolari, Nicoletta, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, pp. 4933–4941. https://aclanthology.org/2020.lrec-1.607.

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky (2021), A primer in BERTology: What we know about how BERT works, *Transactions of the Association for Computational Linguistics* **8**, pp. 842–866, MIT Press One Rogers Street, Cambridge, MA 02142-1209, USA journals-info ....

Schuster, Tal, Adam Fisch, and Regina Barzilay (2021), Get your vitamin C! robust fact verification with contrastive evidence, *in* Toutanova, Kristina, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Online, pp. 624–643. https://aclanthology.org/2021.naacl-main.52.

Sinha, Koustuv, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela (2021), Masked language modeling and the distributional hypothesis: Order word matters pre-training for little, *in* Moens, Marie-Francine, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp. 2888–2913. https://aclanthology.org/2021.emnlp-main.230.

Truong, Thinh Hung, Timothy Baldwin, Karin Verspoor, and Trevor Cohn (2023), Language models are not naysayers: an analysis of language models on negation benchmarks, *in* Palmer, Alexis and Jose Camacho-collados, editors, *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, Association for Computational Linguistics, Toronto, Canada, pp. 101–114. https://aclanthology.org/2023.starsem-1.10.

Wang, Dilin, Chengyue Gong, and Qiang Liu (2019), Improving neural language modeling via adversarial training, *International Conference on Machine Learning*, PMLR, pp. 6555–6565.

Wang, Gengyu, Kate Harwood, Lawrence Chillrud, Amith Ananthram, Melanie Subbiah, and Kathleen McKeown (2023), Check-COVID: Fact-checking COVID-19 news claims with scientific evidence, *in* Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, Association for Computational Linguistics, Toronto, Canada, pp. 14114–14127. https://aclanthology.org/2023.findings-acl.888.

Wang, Lucy Lu, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier (2020), CORD-19: The COVID-19 open

research dataset, *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Association for Computational Linguistics, Online. https://www.aclweb.org/anthology/2020.nlpcovid19-acl.1.

Wang, William Yang (2017), "liar, liar pants on fire": A new benchmark dataset for fake news detection, *in* Barzilay, Regina and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, Vancouver, Canada, pp. 422–426. https://aclanthology.org/P17-2067.

Zhu, Qi, Yuxian Gu, Lingxiao Luo, Bing Li, Cheng Li, Wei Peng, Minlie Huang, and Xiaoyan Zhu (2021), When does further pre-training MLM help? an empirical study on task-oriented dialog pre-training, *Proceedings of the Second Workshop on Insights from Negative Results in NLP*, pp. 54–61.

Appendix A. Evaluation of datasets

In Table 2 we summarize the results evaluating the different variants of the LitCovid 200 and 10K datasets respectively. Table 3, we include the results for the shuffled datasets.

	LitCovid 200			**LitCovid 10K**		
	True	False	Paraphrased	True	False	Paraphrased
type-token-ratio	0.70	0.70	0.70	0.69	0.70	0.70
mean length of utterance	126	134	143	120	127	135
Flesch reading ease	35.16	26.28	27.55	35.89	27.18	27.79
average sentence length	30	30	31	29	31	31
average word length	4.6	4.9	4.9	4.79	5.02	5.03
unique word count	5,993	5,915	5,936	73,285	68,890	60,263
Shannon entropy	5.61	5.70	5.78	5.59	5.69	5.79
perplexity (GPT2)	25	29	26	27	30	27

Table 2: Metrics for the LitCovid 200 and LitCovid 10K datasets, in their True, False, and Paraphrased setting.

	LitCovid 200 shuffled			**LitCovid 10K shuffled**		
	True	False	Paraphrased	True	False	Paraphrased
type_token_ratio	0.80	0.85	0.76	0.80	0.79	0.78
mean_length_of_utterance	126	85	142	120	127	135
flesch_reading_ease	50.50	40.88	38.32	45.78	37.35	34.13
average_sentence_length	32	32	33	30	31	32
average_word_length	4.6	4.9	4.9	4.8	5	5
unique_word_count	6,025	5,944	5,969	73,447	69,051	60,581
shannon_entropy	5.94	5.09	6.02	5.92	5.97	5.99
perplexity GPT2	899	1,034	1,031	949	1,023	1,114

Table 3: Metrics for the shuffled LitCovid 200 and LitCovid 10K datasets, in their True, False, and Paraphrased setting.

Appendix B. Prompt misinformation

Prompt to generate misinformation from original texts in academic publications present in the LitCovid repository:

You are an AI language model tasked with processing academic texts related to COVID-19. Your objective is to rewrite an Input Text A but sustain the opposite position.
In simpler terms, let's imagine there are two statements (A and B) related to COVID-19, and they cannot both be true at the same time. If the Input Text supports fact A, the task is to rewrite the text in a way that now supports statement B.
Let's first understand the problem by reading the instructions, then extract relevant variables, and make a plan. Then, let's carry out the plan, calculate intermediate variables (pay attention to commonsense), solve the problem step by step, and show the answer. Look at the example cases below to understand the task. Each example consists of an example input (Input Text), what your output should look like (Expected Output), and an example of undesirable output (Wrong Output). After having read and understood the examples, transform the Input Text.

Instructions:

1. **Read the input text carefully.** It will be an abstract or a paragraph from an academic publication about COVID-19.

2. **Transform the text.** Your goal is to reverse the truthfulness of the information presented in the text. If the Input Text supports a statement, make sure the Output Text supports the opposite.

3. **Maintain academic tone and style.** Despite the transformation, the resulting text should preserve the formal and structured nature of academic writing.

4. **Make sure the output is coherent.** While making an Output Text contradictory to the Input Text, make sure the Output Text is coherent and logical.

5. **Use various linguistic techniques.** Avoid relying solely on negations. Employ a range of linguistic strategies such as rephrasing with antonyms, altering contexts, introducing contrary facts, or any creative method that inverts the factual basis of the content.

6. **Output format:** Return the transformed text as a string. If multiple texts are provided in one session, separate each transformed text with a newline.

7. **Length of response:** The Output Text should be approximately the same length as the Input Text to ensure that the essence and detail of the original content are mirrored in the transformation.

Example Case 1:
- **Input Text:** "Recent studies indicate that COVID-19 primarily spreads through respiratory droplets." - **Expected Output:** "In-depth analyses suggest that COVID-19's transmission is unrelated to respiratory droplets." - **Wrong Output:** "Old studies indicate that COVID-19 does not spread through respiratory droplets."
Example Case 2:
- **Input Text:** "In a groundbreaking discovery, researchers have identified a specific protein that plays a crucial role in the severity of COVID-19 symptoms. Understanding the interaction of this protein with the virus could lead to targeted therapeutic interventions and improved outcomes for patients." - **Expected Output:** "In a groundbreaking discovery, researchers have refuted the existence of a specific protein that plays a crucial role in the severity of COVID-19 symptoms. Since no specific protein interacts with the virus dismisses, there is no possibility of targeted therapeutic interventions, challenging the potential for improved outcomes for patients." - **Wrong Output:** "In an unoriginal discovery, researchers have not identified a specific protein that plays an uncrucial role in the severity of COVID-19 symptoms. Disregarding the interaction of this protein with the virus could not lead to targeted therapeutic interventions and improved outcomes for patients."

Read the instructions and example cases carefully. Only once you fully comprehend your task, proceed with transforming the provided COVID-19 academic text(s) according to these instructions.

Appendix C. Prompt paraphrasing

Prompt to generate paraphrases from original texts in academic publications present in the LitCovid repository:

Context: You are an AI language model tasked with processing academic texts related to COVID-19. Your role involves creatively paraphrasing these texts. The original texts may include abstracts or paragraphs from academic publications. Your objective is to paraphrase its content using a variety of linguistic techniques. This exercise aims to explore the flexibility of language and understand how the same information can be presented in various ways while maintaining logical coherence and readability.

Instructions:

1. **Read the input text carefully.** It will be an abstract or a paragraph from an academic publication about COVID-19.

2. **Transform the text.** Your goal is to paraphrase the information presented in the text. Avoid using only synonyms. Employ a range of linguistic strategies such as changing word classes, using a different grammatical structure or voice (active vs. passive), elaborating on the original text, or any creative method that paraphrases the content.

3. **Maintain academic tone and style.** Despite the transformation, the resulting text should preserve the formal and structured nature of academic writing.

4. **Make sure the output is coherent.** While paraphrasing, make sure the output is coherent and logical.

5. **Do not only use synonyms.** Rely on other paraphrasing techniques besides using synonyms of terms used in the original sentence.

6. **Output format:** Return the transformed text as a string. If multiple texts are provided in one session, separate each transformed text with a newline.

7. **Length of response:** The transformed text should be approximately the same length as the input text to ensure that the essence and detail of the original content are paraphrased in the transformation.

Example:

- **Input Text:** "Recent studies indicate that COVID-19 primarily spreads through respiratory droplets." - **Expected Output:** "Recent research shows that COVID-19 mainly transmits via respiratory droplets."

Read the context and instructions carefully. Only once you fully comprehend your task, proceed with transforming the provided COVID-19 academic text(s) according to these instructions.

Appendix D. Model specifications

D.1 Model specifications continual pre-training BERT

To continue pre-training BERT, we follow this procedure. We pre-train for one step using the MLM objective, for which we use the baseline script on HuggingFace (which was also used for the research in (Gururangan et al. 2020). As is standard practice, we mask 15% of the tokens. The learning rate was set at 5e-05, consistent with the usual rate for domain adaptation (Bacco et al. 2023, Gururangan et al. 2020). We train for one epoch, using mixed precision (fp16) to accelerate the process. Since we use a collection of text units as input data, we opt for line-by-line, which directs the model to use the text inputs as separate sequences. When the CPT is completed, we save and upload the model to HuggingFace, making

it accessible for further use.

D.2 Model specifications fine-tuning BERT on fact verification

To fine-tune BERT on the downstream task of fact verification, we follow this procedure. We maintain stable hyperparameter settings to ensure consistency in the experimental conditions (similar to Bacco et al. (2023)). Hyperparameter tuning was conducted on the baseline BERT model using the Optuna library (Akiba et al. 2019). We set the learning rate to 3.5e-05, adjust the batch size to 32 to accommodate hardware limitations, use 5 training epochs, and implement early stopping after two epochs to prevent overfitting.

Appendix E. Detailed results comparing BERT-base to CPT BERT-base

In Table 4, we give the p-values and effect sizes when comparing the BERT-base baseline to BERT-base CPT models.

models	p-value	Cohen's g
bert - biobert	0.010	0.20
bert - litcov200 True shuffled	0.020	0.20
bert - litcov200 False	0.001	0.18
bert - litcov200 False shuffled	0.010	0.21
bert - litcov200 paraphrased	0.010	0.17
bert - litcov200 paraphrased shuffled	0.007	0.19
bert - checkcovid	0.007	0.17
bert - reddit	0.010	0.20
bert - reddit shuffled	0.040	0.17
bert - litcov10K False	0.030	0.14
bert - litcov10K paraphrased	0.030	0.14
bert - litcov10K paraphrased shuffled	0.002	0.18

Table 4: P-values (McNemar) and effect sizes (Cohen's g) comparing BERT-base baseline and BERT-base CPT models.