

# A Taxonomy for Data Contamination in Large Language Models

Medha Palavalli and Amanda Bertsch and Matthew R. Gormley

School of Computer Science

Carnegie Mellon University

[mpalaval, abertsch, mgormley]@cs.cmu.edu

## Abstract

Large language models pretrained on extensive web corpora demonstrate remarkable performance across a wide range of downstream tasks. However, a growing concern is data contamination, where evaluation datasets may be contained in the pretraining corpus, inflating model performance. Decontamination, the process of detecting and removing such data, is a potential solution; yet these contaminants may originate from altered versions of the test set, evading detection during decontamination. How different types of contamination impact the performance of language models on downstream tasks is not fully understood. We present a taxonomy that categorizes the various types of contamination encountered by LLMs during the pretraining phase and identify which types pose the highest risk. We analyze the impact of contamination on two key NLP tasks—summarization and question answering—revealing how different types of contamination influence task performance during evaluation.

## 1 Introduction

Advancements in machine learning have traditionally relied on benchmark datasets to evaluate and compare model performance (Raji et al., 2021; Gururaja et al., 2023). With the surge of large language models (LLMs) in recent years, these benchmarks are now leveraged to showcase remarkable abilities across diverse tasks.

However, the shelf life of benchmarks is incredibly low, with Roberts et al. (2023) demonstrating that newer models with updated training cutoff dates are iteratively rendering existing benchmarks stale. The presence of internet-sourced data in both pretraining and evaluation datasets increases the risk of data contamination (Brown et al., 2020; Magar and Schwartz, 2022) and challenges the notion of fair evaluation for models pretrained on massive corpora. Both GPT-3 and C4 training corpora were found to contain test data for several benchmarks

(Dodge et al., 2021; Raffel et al., 2020; Brown et al., 2020), raising serious concerns about the validity of evaluation scores for many pretrained models (Lee et al., 2022; Chang et al., 2023b).

The research community lacks consensus on best practices for data contamination, and different works define contamination in subtly different ways. Without standardization of terminology, it is difficult to develop best practices for contamination—or even to characterize the problem at all. To address this gap, we suggest a formal definition of contamination and taxonomize subtypes of contamination (§ 2). We map prior work on both the detection and impact of contamination into this taxonomy, revealing several understudied forms of contamination (§ 2.3). We also measure the impact of different types of contamination on downstream summarization (§ 4) and QA (§ 5) performance through continued pretraining experiments assessing indirect/approximate test set contamination effects.

Our findings reveal that for GPT-2 Large models, it is often the case that having in-domain data present during training is as beneficial as having the test data present during training. Moreover, we observe that certain contamination types exhibit task-dependent effects on evaluation performance, further complicating decontamination best practices. Our findings enable recommendations for identifying and mitigating problematic contamination during LLM development to ensure reliable evaluations (§ 7).

## 2 Taxonomy

Consider a model  $M : \mathcal{X} \rightarrow \mathcal{Y}$  which, given an input of some type  $x \in \mathcal{X}$ , outputs text  $\hat{y} \in \mathcal{Y}$ . While  $x$  can be of any format, we will restrict ourselves to cases where  $\hat{y}$  is in the space of the *natural language* ( $\mathcal{Y} \subseteq \Sigma^*$  for some alphabet in  $\Sigma$ ). Let  $D$  be the *test set*, consisting of  $|D|$  examples  $\langle x_i, \hat{y}_i \rangle$ .

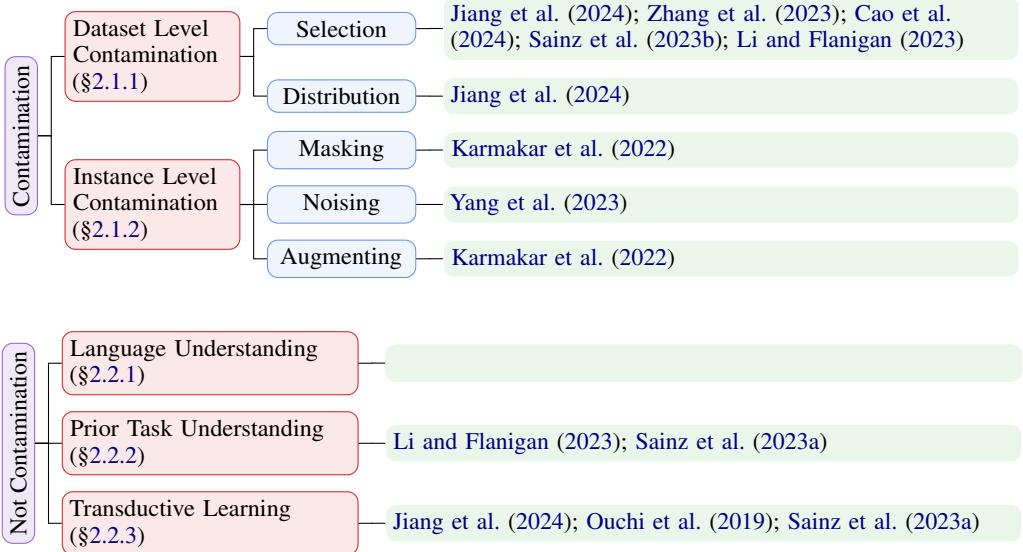


Figure 1: Taxonomy of Contamination, with some representative works in the literature that address each category.

## 2.1 Contamination

We define **contamination** as any leakage of information that provides a signal for the correct label for at least one example in the test set  $D$ . When contamination occurs, some subset of the pretraining data can be characterized as the result of a function  $f(D)$ , which may be a composition of multiple contamination functions  $f = f^{(1)} \circ f^{(2)} \circ \dots \circ f^{(n)}$ . We characterize types of contamination by their *dataset-level* (§ 2.1.1) and *example-level* (§ 2.1.2) properties. Figure 1 provides an overview of our taxonomy.

### 2.1.1 Dataset-level Properties

For dataset-level contamination, consider a function  $g$  that leaves the individual examples  $\langle x_i, \hat{y}_i \rangle$  intact. In the simplest case,  $g$  is the identity function; this is the leakage of a full test set, e.g. from scraping a file containing the test set instances and labels. The following are types of functions  $g(D)$  can take on.

- **Selection:** A function that selects some group of examples  $D' \subset D$ , such that only a subset of the test set is leaked. This is likely when the test data is drawn from several sources, only some of which appear in the pretraining data; when some of the test data is more recent than other data and the pretraining data contains an older snapshot of the contamination source; or when the data is contained in several documents and the cleaning of the pretraining data only removes some of these documents. Ver-

*batim contamination* refers to when  $g$  is the identity function.

- **Distribution:** A function which combines the contaminated data  $D$  with some additional, non-contaminating documents, such that the examples from  $D$  are not all sequential in the pretraining data. This can occur during data shuffling, or if the contamination comes from multiple documents. Practically, this means that the contaminated region of the pretraining data  $g(D)$  spans more tokens.

### 2.1.2 Instance-level Properties

In instance-level contamination, the function  $f$  applies some function  $h$  to each individual leaked example  $f(D) = \{h(\langle x_i, \hat{y}_i \rangle)\}_{i=1}^{|D|}$ .<sup>1</sup> A few representative examples in this class are enumerated below:

- **Masking:** A function that removes some or all of the input (can be done in combination with the output), e.g.  $h(\langle x_i, \hat{y}_i \rangle) = \hat{y}_i$  or removing all incorrect answer choices in a multiple choice question. *This primarily qualifies as contamination for generation tasks;* for a classification task, leaking the label-space in advance may not be a concern if the labels don't have inherent contextual value without the input, such as binary labels like 0s and 1s or positives and negatives. However, if the

<sup>1</sup>Note that this is a strict subset of all functions applied to the leaked dataset,  $f(D)$ ; however, we distinguish this set of functions that operate on individual examples.

labels carry meaningful information on their own, their premature disclosure would indeed constitute contamination. Note that masking *all of the output*, leaving only the inputs from the test set, is generally considered to be a type of transductive learning, *not* contamination; see § 2.2.3 for more discussion.

- **Noising:** A function that modifies the surface form of the example, e.g. by paraphrasing the inputs or outputs, by presenting the output before the input, or by using silver rather than gold labels for each example. Note that this can also take the form of *alternate correct answers* being present in the pretraining data: for instance, in book summarization, a different summary of the book being present in the pretraining data is still contamination.
- **Augmenting:** A function that adds additional context, which may or may not be relevant to the example. For instance, for a task where the model must answer an open-ended question at test time, an *augmented* contaminated example in pretraining would be a multiple-choice test with the same questions. While this provides the correct answer, it also introduces new (distractor) information that is not present at test time. Another example would be including additional context paragraphs for QA in addition to the necessary context and answer. Note the difference between example-level *augmenting* and dataset-level *distribution*.

## 2.2 Phenomena that aren’t Contamination

For clarity, we describe several phenomena that lead to improved performance on test sets downstream but are *not* considered contamination under our taxonomy.

### 2.2.1 Language Understanding

Pretraining enables models to produce (generally) fluent text and encodes some representation of meaning for words commonly used in task definitions; for instance, the model has some representation of meaning for the labels “positive” and “negative” in sentiment analysis. While this representation is likely helpful for performing downstream tasks (Min et al., 2022), this is not inherently contamination.

### 2.2.2 Prior Task Understanding

We define prior task understanding as an ability to perform a task learned from non-contamination sources, and such prior knowledge has been demonstrated to boost model performance when evaluated on unseen instances of said task (Li and Flanigan, 2023). For instance, fine-tuning a model on a training dataset for the task is clearly not contamination of the test set, although it generally improves performance on that test set; likewise, pretraining on other related datasets is not contamination for a given test set. For closed-book QA and tasks requiring world knowledge, prior task understanding from training data is essential. Closed-book QA demands answering without external resources, relying solely on the model’s training on similar question-answer pairs or related datasets.

In general, scrutinizing the training data’s sources and nature is crucial to maintain model integrity and generalizability. Prior task understanding may violate the assumption of “zero-shot” performance: that the model has not seen training data for that task.

### 2.2.3 Transductive Learning

Transductive learning (Vapnik, 1998) incorporates an unlabeled test set into training. During training, the raw text inputs of the test set can be used, but the labels are not seen. The model, once trained, is then evaluated on the same test set during the test phase. Transductive LM fine-tuning has shown to consistently improve neural models in both in-domain and out-of-domain settings (Ouchi et al., 2019), although concerns have been raised about blurring the line between training and evaluation (Jiang et al., 2024).

We generally do not consider pretraining on the *inputs* of the test set to be contamination,<sup>2</sup> although we note that this will likely improve performance, in the same manner than pretraining on training set text improves downstream performance by providing some domain adaptation to the testing domain (Gururangan et al., 2020; Krishna et al., 2023). Some prior work refers to the presence of inputs-only in the pretraining data as contamination for classification tasks (Jiang et al., 2024; Ouchi et al., 2019); however, under our taxonomy, we consider this a type of transductive learning.

<sup>2</sup>A key exception is tasks where the input/output distinction does not apply, such as perplexity evaluation on a dataset  $D = \{x_1, \dots, x_{|D|}\}$  of sentences  $x_i$ .

### 2.3 Mapping prior work exploring contamination into this taxonomy

The effects of *selection* have been explored by experiments that compare LLM performance over time (Li and Flanigan, 2023; Cao et al., 2024), prompting the model to generate samples from specific dataset splits (Sainz et al., 2023b), and training LLMs that *select* some subset of an evaluation dataset (Zhang et al., 2023; Jiang et al., 2024).

Jiang et al. (2024) also explores the effects of the frequency in which contaminated data appears *distributed* throughout the pretraining data.

Through zero-shot experimentation on the Codex model (Chen et al., 2021), Karmakar et al. (2022) investigates the effects of prompts *masking* out input specifications and prompts with *augmented* objectives. Additionally, Yang et al. (2023) showcases memorization of evaluation samples by prompting LLMs with *noisy* samples.

A prior position paper (Sainz et al., 2023a) defined three categories of data contamination: their *guideline contamination* falls under our definition of prior task understanding; their *raw text contamination* is transductive learning; and their *annotation contamination* equates to our definition of data contamination in § 2.1. Our work further categorizes and explores types of annotation contamination.

### 2.4 Detecting Data Contamination

**Methods with access to pretraining data** Early research on LLM data contamination primarily employed methods akin to high-order n-gram overlap detection between pretraining and evaluation data (Radford et al., 2019a; Brown et al., 2020; Wei et al., 2021; Touvron et al., 2023). Tools for qualitative analysis on large-scale corpora (such as Data Portraits (Marone and Durme, 2023) and the ROOTS Search Tool (Piktus et al., 2023)) have further increased the practicality of this type of contamination detection. However, these approaches have several limitations: they remain fairly computationally expensive, assume access to pretraining data, and generally can only detect contamination when a cluster of several test set examples co-occur (as most methods leverage data sketching (Broder, 1997) tools that are only effective for sequences above a certain length).

Yang et al. (2023) proposes an LLM-based de-contamination method, which leverages embedding similarity search followed by evaluation with a strong language model (e.g. GPT-4), to identify

and mitigate contamination. This is computationally costly but can identify *noisy* contamination

#### Methods without access to pretraining data

Some approaches are capable of detecting contamination without direct access to pretraining data, but assume that the test data has not been modified or distributed across the pretraining corpus. These methods leverage *metadata* from the dataset to detect contamination, e.g. by leveraging dataset ordering (Sainz et al., 2023b) or the assignment of examples to specific data splits (Golchin and Surdeanu, 2023). Golchin and Surdeanu (2024) introduce the Data Contamination Quiz, a streamlined method that efficiently detects and estimates verbatim contamination in LLMs by crafting multiple choice questions that prompt a model to correctly dataset-specific content among similar but noisy alternatives.

Chang et al. (2023a) detect contamination of books (which serve as inputs for many long-context evaluation datasets) using domain specific features—a name cloze test and a publication-year evaluation. This is powerful for detecting the presence of the exact text of the book, but its efficacy on detecting related artifacts (e.g. summaries of the book, which may serve as test set outputs) is unknown.

Shi et al. (2023) introduces a new detection method MIN-K% PROB, which is capable of detecting whether a piece of text was in the pretraining corpora by leveraging the variability of the tokens’ probabilities according to the model. This has the potential to detect distributed or masked contamination, but is not robust to noising operations, which change the token sequence.

Most contemporary data-contamination detection techniques are designed to identify contamination of full, non-distributed test datasets, resulting in a significant gap in detecting noisy or partial contamination. The methods most well-adapted to detect noisy contamination, while powerful, require access to pretraining data and expensive operations; more work is necessary to lower the barrier to detection.

## 3 Methodology

In all our experiments, we employ GPT-2 Large (Radford et al., 2019b).<sup>3</sup> This will be referred to as the *initial model*. Since the pretraining corpus for GPT-2 is not publicly accessible, there is a chance

<sup>3</sup>Our implementation uses nanoGPT (Karpathy, 2023) initialized with OpenAI’s gpt2-large weights.

that these *learned* weights of GPT-2 might be contaminated. Consequently, the outcomes of our experiments serve as a conservative estimate or lower bound on the effects of data contamination.

For each of our datasets, we create train/in-domain/test splits of equal size, aiming to establish a fair and comparable evaluation environment. To disentangle the effects of exposure to test data during pretraining from those of prior task understanding, we constructed an in-domain data split, allowing us to train models on task-relevant but uncontaminated data for comparison against the various contaminated settings. To partially mitigate the potential recency bias from continued pretraining, we incorporate an additional 10,000 samples of Open AI’s WebText (Radford et al., 2019b) into the continued pretraining data.

During continued pretraining, we use a blocksize of 1024 tokens with a batchsize of 1. For finetuning, the training data is seen sample by sample. To obtain deterministic results during our experiments, we set the temperature to zero and capped the maximum completion length at 200 tokens.

### 3.1 Training Settings

We consider several settings for incorporating data:

- ZERO-SHOT (not contamination): prompt the initial model with the test sample and a simple instruction for the task.
- BASELINE (not contamination): finetune initial model with train split
- CHEATING (contamination at fine-tuning time, rather than pre-training): finetune initial model with test split
- *Contamination Setting(s)* (standard contamination during pretraining): continued pretraining with  $f(\text{test split})$  and finetune with train split; the details of each contamination setting are specific to the task (§ 4 and § 5)
- *In-Domain Setting(s)* (not contamination): continued pretraining with  $f(\text{in-domain split})$  and finetune with train split—for each contamination setting in § 4 and § 5, there is an associated in-domain model.

For each setting+dataset, we average results over models trained on 3 random shuffles of the data. Standard deviations are computed over these 3 runs and error bars indicate  $\pm$  one standard deviation.

## 4 Case Study: Summarization

For this case study, we use the following summarization datasets: XSum (Narayan et al., 2018), SAMSUM (Gliwa et al., 2019), and CNN/Daily Mail (Nallapati et al., 2016). We explore 5 contamination settings:

1. VERBATIM (dataset level, selection):  $f = \text{identity function on test split}$
2. DISTRIBUTION (dataset level, distribution):  $f = \text{shuffle test data with WebText}$
3. MASKED (instance level, masking):  $h = \text{mask out input documents in test split}$
4. NOISED (instance level, noising):  $h = \text{swap in GPT-3.5}^4 \text{ generated summaries on test split}$
5. REFORMATTED (instance level, noising):  $h = \text{swap format from document-summary to summary-document for test data}$

Table 5 provides examples of each setting.

### 4.1 Results

In this section, we consider the overall performance of each contamination method across summarization datasets. Figure 2 shows an example of the results from one task and one metric (SAMSUM, ROUGE-L). See Appendix A for full results on all tasks and metrics, specifically Figures 4, 5, 6 or Table 3.

Consistently, the CHEATING setting outperforms all others; this is expected, given that deliberately finetuning on the test data is an extreme form of contamination.

Overall, continued pretraining with the approximate contamination methods improves performance above the BASELINE setting, often substantially. This suggests that exposure to these forms of contamination during pretraining can impact the reliability of evaluations on this data downstream.

While VERBATIM setting performs slightly better than the other contamination settings, this improvement isn’t significant for most settings. Note that most contaminated settings outperform the baseline, and exist within a standard deviation of each other. This suggests that the performance boost may simply be attributed to the increase in in-domain data seen during the training stage rather than encountering the test split during continued pretraining.

Note that for the most part, the VERBATIM and INDOMAIN-VERBATIM settings perform on par

---

<sup>4</sup>gpt-3.5-turbo-0125 with temperature=0.5

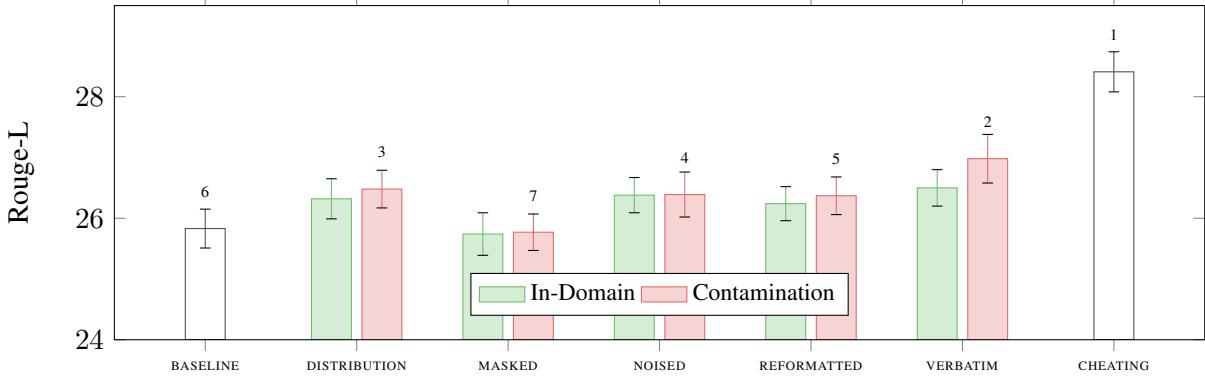


Figure 2: Bar Chart of all SAMSum models compared for Rouge-L.

with each other. This trend seems to hold true for the other *contamination* and *in-domain* model pairs. The comparable performance further suggests that exposure to contaminated data may not be the primary factor boosting model performance in the *contamination* settings studied.

Dataset	R-1	R-2	R-L	R-Lsum
CNN	38.70	14.14	24.90	32.11
SAMSum	37.92	13.78	28.73	28.75
XSum	24.21	4.89	16.60	16.60

Table 1: Rouge scores (R-) for summaries generated by GPT-3.5. These summaries are used as silver labels for our NOISED contamination setting.

While the majority of these settings have metrics that fall within one standard deviation of each other, there are exceptions. For instance, in the case of the XSum dataset, the NOISED setting fails to surpass the BASELINE. This discrepancy can be attributed to the idiosyncrasies of the XSum dataset, where ground truth summaries may deviate significantly from typical summaries, thus posing a challenge for the model in generating accurate outputs. Table 1 shows that the summaries generated by GPT-3.5 (Brown et al., 2020) for the XSum dataset have lower rouge scores than the other two datasets.

Additionally, underperformance of the MASKED contamination setting compared to the BASELINE across all datasets is noteworthy, suggesting that exposure only to summaries during pretraining may fail to achieve the benefits of seeing *in-domain* data.

## 5 Case Study: Question Answering

For this case study, we consider open-ended QA with SQuAD (Rajpurkar et al., 2016) and multiple-

choice QA with the Children’s Book Test (CBT) (Hill et al., 2016). We explore 6 contamination settings:

1. VERBATIM (dataset level, selection):  $f =$  identity function on test split
2. DISTRIBUTION (dataset level, distribution):  $f =$  shuffle test data with WebText
3. MASKED (instance level, masking):  $h =$  mask out context passage in test split
4. NOISED (instance level, noising):  $h =$  encounter GPT-3.5 generated answers to test split questions
5. REFORMATTED (instance level, augmenting/-masking)<sup>5</sup>:  $h_{\text{SQuAD}} =$  introduce 3 distractor multiple choice answer options;  $h_{\text{CBT}} =$  mask out incorrect answer options
6. AUGMENTED (instance level, augmenting):  $h =$  prompt GPT-3.5 to add additional content to the context passages in the test split

Table 6 provides examples of each setting.

### 5.1 Results

In this section, we consider the overall performance of each contamination method across Question Answering datasets. Figure 3 shows an example of the results from one task and one metric (SQuAD, Exact Match). See Appendix B for full evaluation results on all tasks and metrics, specifically Figures 7, 8 or Table 4.

Once again, the CHEATING setting outperforms all others by a noticeable margin. With the exception of the MASKED setting for the SQuAD

<sup>5</sup>For SQuAD, this is a form of augmented contamination, as additional (distractor) information is introduced. For CBT, this is a form of masked contamination, as information is removed.

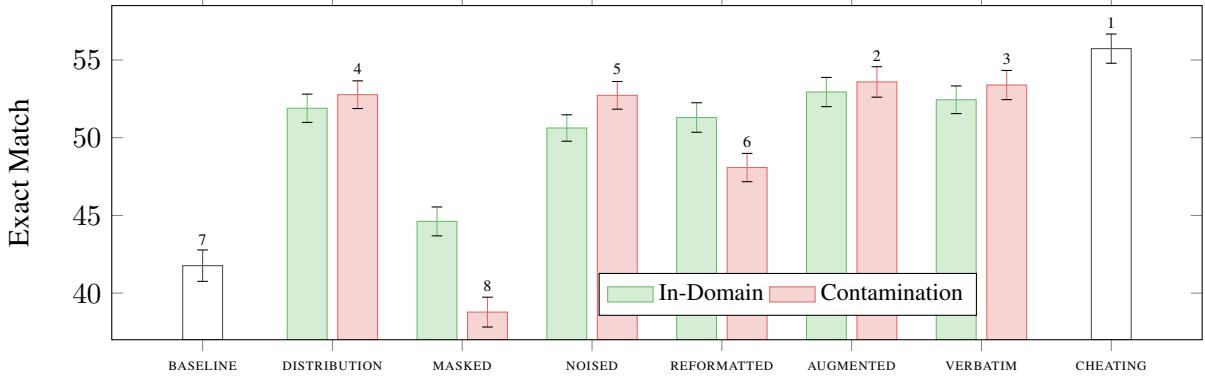


Figure 3: Bar Chart of all SQuAD models compared for Exact Match.

dataset, all contaminated settings exhibit better performance compared to the BASELINE setting by a considerable margin. This indicates that the increased data diversity experienced by both the *in-domain* and *contaminated* models during training improved their performance during evaluation.

Dataset	Exact Match	F1 Score
SQuAD	74.56	88.15
CBT	77.21	79.78

Table 2: Exact match and F1 scores for answers generated by GPT-3.5.

Note that the NOISED setting performs almost as well as the VERBATIM contamination setting. We attribute this to the fairly high quality of silver labels generated by GPT-3.5 (see Table 2).

Exposure to *in-domain* data during pretraining appears to improve model performance. However, our results show that contaminated settings such as NOISED, VERBATIM, and DISTRIBUTION tend to outperform the corresponding *in-domain* settings during evaluation. This suggests that seeing data from the test set positively impacts model performance for question answering tasks. Note that for these three model setups, the format of context, question, and answer is almost consistent with the format and content seen during evaluation time.

Reformatting (augmenting) free-form questions from SQuAD into multiple-choice answers during pretraining appears to have a negative effect on model performance, though it still outperforms the BASELINE setting. Conversely, converting multiple-choice questions from CBT into free-form questions (masking) during pretraining yields positive results, with the REFORMATTED setting outperforming most other contaminated settings.

Furthermore, we observe variations in the performance of AUGMENTED setting across the two datasets. While this setting perform well for SQuAD, its performance is not as impressive for CBT. This discrepancy may be attributed to the nature of data augmentation, where the additional information provided for SQuAD is more relevant and beneficial to the wikipedia paragraphs compared to the irrelevant introductions, such as ‘once upon a time’ style introductions generated by GPT-3.5 for these book excerpts, added to CBT stories. It is important to note that since this information doesn’t significantly contribute to the task, this form of augmentation falls in a blurry space between distribution and augmentation branches of the taxonomy. It could also be viewed as unrelated information being added between samples during pretraining, complicating its categorization.

## 6 Analysis

Unsurprisingly, the CHEATING and VERBATIM contamination settings consistently outperform the BASELINE across both tasks. The *in-domain* settings’ consistent outperformance of the BASELINE underscores the advantages of exposure to related samples during pretraining (Krishna et al., 2023).

Far more concerning is that several approximate contamination settings outperform both the BASELINE and their respective *in-domain* settings, suggesting that the model in these settings benefits not only from seeing *in-domain* text but from unfairly leveraging prior knowledge of the test examples. In particular, the NOISED setting, which is generally not detectable with existing decontamination methods, produces scores inflated over BASELINE in all datasets, and scores more than one standard deviation above its corresponding *in-domain* setting in

several datasets.

The MASKED setting generally performs around or worse than the BASELINE, possibly due to the more extreme formatting mismatch between this data and the test data. We expect that the MASKED setting may be encountered in the wild if a file of outputs for the dataset is in the pretraining data; the limited impact of this contamination on downstream performance is thus good news, though more investigation would be necessary to conclusively say MASKED contamination is not a concern.

For many of the *contaminated* settings and their corresponding *in-domain* settings, the effect of approximate contamination is not greater than effect of in-domain data seen during pretaining. However, research has shown that memorization in LLMs significantly grows as the size of the model increases (Carlini et al., 2023). The number of times a sample has been duplicated in the pretraining corpora has also been shown to increase a model’s memorization capabilities (Carlini et al., 2023; Golchin and Surdeanu, 2023).

Some behavior is task- or dataset-specific, emphasizing that there is no one-size-fits-all approach to data curation: the importance of removing each type of contamination from the pretraining corpus is at least partially linked to the specific task’s formatting. However, some types of approximate contamination do lead to inflated scores, emphasizing that considering a more broad definition of contamination when de-contaminating pretraining corpora is a worthwhile endeavor.

## 7 Conclusion

Our analysis highlights the importance of data format, with models performing better when pretraining data matches the evaluation format. We also observe task-specific effects, with certain contamination methods benefiting particular tasks more than others. Additionally, we find that some late-stage pretraining contamination can actually be *unhelpful* to downstream performance, if it occurs in a substantially different format from the downstream task. Our findings underscore gaps in current de-contamination practices, which primarily focus on full-dataset-level contamination and are often unable to detect approximate or noisy contamination.

We demonstrate that different types of contamination can have variable effects on model performance, highlighting the need for careful consideration during training and evaluation. With the

creation of our taxonomy, we hope to promote standardization regarding the definition and categories of contamination within the research community, facilitating clear communication and collaboration, while also enabling precise detection and mitigation of contamination in pretraining data. We recommend researchers decontaminating pretraining corpora for LLMs prioritize developing techniques that address noisy evaluation data, while also ensuring rigorous scrutiny to prevent any shuffled or interleaved evaluation data from inadvertently persisting in the pretraining data. It is not enough to merely remove instances of the full test dataset in the pretraining corpus; fragments or noised versions of the test set can also inflate performance. We hope our work inspires future work on detecting and mitigating specific types of contamination.

## 8 Limitations

Due to resource constraints, we only investigate the impact of encountering contaminated data towards the end of pretraining (i.e. with continued pretraining), rather than randomly throughout pre-training. This may introduce recency bias, influencing our findings. Additionally, our focus on a single language model limits the generalizability of our results. GPT-2 pretraining data is not publicly accessible so our results may only offer an approximation of contamination effects. Different model architectures, training procedures, and datasets may yield varying impacts of contamination. Conducting experiments on larger LLMs could potentially reveal more pronounced effects of contamination, as larger models have been shown to exhibit greater tendencies of memorization (Carlini et al., 2023). Further research involving multiple models and comprehensive evaluations is needed to establish more robust conclusions across diverse settings.

## Acknowledgements

The authors would like to thank Lori Levin for her early guidance and the anonymous reviewers for their thoughtful comments. This work was supported in part by grants from 3M, the Pittsburgh Supercomputing Center, and the National Science Foundation Graduate Research Fellowship under Grant No. DGE2140739.

## References

- A.Z. Broder. 1997. **On the resemblance and containment of documents.** In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No.97TB100171)*, pages 21–29.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. **Language models are few-shot learners.**
- Jialun Cao, Wuqi Zhang, and Shing-Chi Cheung. 2024. **Concerned with data contamination? assessing countermeasures in code language model.**
- Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. **Quantifying memorization across neural language models.**
- Kent K. Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023a. **Speak, memory: An archaeology of books known to chatgpt/gpt-4.**
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023b. **A survey on evaluation of large language models.**
- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. **Evaluating large language models trained on code.**
- Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. **Documenting large webtext corpora: A case study on the colossal clean crawled corpus.** In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Bogdan Gliwa, Iwona Mochol, Maciej Bieseck, and Aleksander Wawer. 2019. **SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization.** In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.
- Shahriar Golchin and Mihai Surdeanu. 2023. **Time travel in llms: Tracing data contamination in large language models.**
- Shahriar Golchin and Mihai Surdeanu. 2024. **Data contamination quiz: A tool to detect and estimate contamination in large language models.**
- Sireesh Gururaja, Amanda Bertsch, Clara Na, David Widder, and Emma Strubell. 2023. **To build our future, we must know our past: Contextualizing paradigm shifts in natural language processing.** In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13310–13325, Singapore. Association for Computational Linguistics.
- Suchin Gururangan, 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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.
- Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. **The goldilocks principle: Reading children’s books with explicit memory representations.**
- Minhao Jiang, Ken Ziyu Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. 2024. **Investigating data contamination for pre-training language models.**
- Anjan Karmakar, Julian Aron Prenner, Marco D’Ambros, and Romain Robbes. 2022. **Codex hacks hackerrank: Memorization issues and a framework for code synthesis evaluation.**
- Andrej Karpathy. 2023. **nanogpt.**
- Kundan Krishna, Saurabh Garg, Jeffrey P. Bigham, and Zachary C. Lipton. 2023. **Downstream datasets make surprisingly good pretraining corpora.**
- Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. **Deduplicating training data makes language models better.** In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,

- pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.
- Changmao Li and Jeffrey Flanigan. 2023. Task contamination: Language models may not be few-shot anymore.
- Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation.
- Marc Marone and Benjamin Van Durme. 2023. Data portraits: Recording foundation model training data.
- Sewon Min, Xinxin Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gülcöhre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.
- Hiroki Ouchi, Jun Suzuki, and Kentaro Inui. 2019. Transductive learning of neural language models for syntactic and semantic analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3665–3671, Hong Kong, China. Association for Computational Linguistics.
- Aleksandra Piktus, Christopher Akiki, Paulo Villegas, Hugo Laurençon, Gérard Dupont, Alexandra Sasha Luccioni, Yacine Jernite, and Anna Rogers. 2023. The roots search tool: Data transparency for llms.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019a. Language models are unsupervised multitask learners.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019b. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- 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. *CoRR*, abs/1910.10683.
- Inioluwa Deborah Raji, Emily M. Bender, Amanda-lynne Paullada, Emily Denton, and Alex Hanna. 2021. Ai and the everything in the whole wide world benchmark.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Manley Roberts, Himanshu Thakur, Christine Herlihy, Colin White, and Samuel Dooley. 2023. Data contamination through the lens of time.
- Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023a. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.
- Oscar Sainz, Jon Ander Campos, Iker Garcia-Ferrero, Julen Etxaniz, , and Eneko Agirre. 2023b. Did chatgpt cheat on your test?
- Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2023. Detecting pretraining data from large language models.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esibou, Jude Fernandes, Jeremy Fu, Wenjin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Miaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Bin Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.
- Vladimir Vapnik. 1998. Statistical learning theory. *John Wiley & Sons*.
- Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2021. Finetuned language models are zero-shot learners. *CoRR*, abs/2109.01652.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. 2023. [Rethinking benchmark and contamination for language models with rephrased samples](#).

Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2023. [Counterfactual memorization in neural language models](#).

## A Full results for Summarization Case Study

We present the full results of the summarization case study. For each setting and dataset, we have included a table of the Rouge metrics along with their standard deviations. The data is also presented through a series of bar charts for easier interpretability of the results for the reader. Standard deviations are measured over the results of the 3 models trained on random shuffles of the data.

Dataset	Model	Contaminated Pretraining Data	Contaminated Fine-tuning Data	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSUM
CNN	ZERO-SHOT	-	-	21.98 ± 0.26	5.076 ± 0.01	13.63 ± 0.10	18.51 ± 0.10
	BASELINE	-	✗	27.22 ± 0.53	7.436 ± 0.13	18.15 ± 0.43	24.90 ± 0.36
	CHEATING	-	✓	<b>33.60 ± 0.58</b>	<b>10.198 ± 0.16</b>	<b>20.52 ± 0.33</b>	<b>29.61 ± 0.32</b>
	VERBATIM	✓	✗	29.84 ± 0.48	9.488 ± 0.14	19.50 ± 0.38	26.98 ± 0.40
	DISTRIBUTION	✓	✗	29.73 ± 0.33	9.557 ± 0.13	<i>19.50 ± 0.22</i>	27.12 ± 0.26
	MASKED	✓	✗	28.34 ± 0.22	8.326 ± 0.13	18.01 ± 0.29	25.96 ± 0.20
	NOISED	✓	✗	31.31 ± 0.52	8.821 ± 0.15	19.19 ± 0.32	28.85 ± 0.30
	REFORMATTED	✓	✗	29.21 ± 0.28	8.887 ± 0.13	18.88 ± 0.29	26.27 ± 0.30
	INDOMAIN-VERBATIM	✗	✗	29.81 ± 0.48	9.277 ± 0.13	18.93 ± 0.25	26.88 ± 0.31
	INDOMAIN-DIST.	✗	✗	28.86 ± 0.30	8.910 ± 0.13	18.41 ± 0.27	26.10 ± 0.30
SAMSum	INDOMAIN-MASK	✗	✗	28.87 ± 0.39	8.493 ± 0.15	18.24 ± 0.30	26.40 ± 0.29
	INDOMAIN-NOISE	✗	✗	31.16 ± 0.42	8.596 ± 0.10	18.85 ± 0.26	26.53 ± 0.35
	INDOMAIN-REFORM.	✗	✗	28.80 ± 0.31	8.681 ± 0.12	18.75 ± 0.24	26.07 ± 0.32
	ZERO-SHOT	-	-	11.73 ± 0.14	1.357 ± 0.01	8.377 ± 0.19	9.331 ± 0.16
	BASELINE	-	✗	32.95 ± 0.57	10.22 ± 0.15	25.83 ± 0.32	25.59 ± 0.29
	CHEATING	-	✓	<b>36.36 ± 0.53</b>	<b>12.31 ± 0.14</b>	<b>28.41 ± 0.33</b>	<b>28.48 ± 0.33</b>
	VERBATIM	✓	✗	34.34 ± 0.45	10.76 ± 0.16	26.98 ± 0.40	27.04 ± 0.38
	DISTRIBUTION	✓	✗	33.73 ± 0.51	10.32 ± 0.15	26.48 ± 0.31	26.56 ± 0.33
	MASKED	✓	✗	33.05 ± 0.46	10.46 ± 0.15	25.77 ± 0.30	25.81 ± 0.28
	NOISED	✓	✗	33.62 ± 0.43	10.27 ± 0.16	26.50 ± 0.37	26.49 ± 0.38
XSum	REFORMATTED	✓	✗	33.63 ± 0.39	10.25 ± 0.15	26.37 ± 0.31	26.46 ± 0.34
	INDOMAIN-VERBATIM	✗	✗	33.61 ± 0.46	10.27 ± 0.14	26.39 ± 0.30	26.46 ± 0.35
	INDOMAIN-DIST.	✗	✗	33.55 ± 0.42	10.26 ± 0.11	26.32 ± 0.33	26.44 ± 0.35
	INDOMAIN-MASK	✗	✗	32.87 ± 0.41	10.47 ± 0.12	25.74 ± 0.35	25.74 ± 0.31
	INDOMAIN-NOISE	✗	✗	33.67 ± 0.37	10.33 ± 0.13	26.38 ± 0.29	26.47 ± 0.28
	INDOMAIN-REFORM.	✗	✗	33.52 ± 0.34	10.24 ± 0.16	26.24 ± 0.28	26.34 ± 0.29
	ZERO-SHOT	-	-	12.52 ± 0.11	2.059 ± 0.00	9.035 ± 0.16	10.27 ± 0.17
	BASELINE	-	✗	26.28 ± 0.48	6.424 ± 0.12	19.80 ± 0.32	19.81 ± 0.33
	CHEATING	-	✓	<b>29.87 ± 0.41</b>	<b>8.334 ± 0.13</b>	<b>22.97 ± 0.43</b>	<b>22.98 ± 0.42</b>
	VERBATIM	✓	✗	26.53 ± 0.51	6.820 ± 0.12	20.08 ± 0.33	20.03 ± 0.37
XSum	DISTRIBUTION	✓	✗	26.61 ± 0.42	6.885 ± 0.13	<i>20.12 ± 0.37</i>	<i>20.11 ± 0.37</i>
	MASKED	✓	✗	24.50 ± 0.46	5.677 ± 0.12	18.16 ± 0.29	18.39 ± 0.31
	NOISED	✓	✗	26.16 ± 0.39	6.599 ± 0.12	19.72 ± 0.35	19.72 ± 0.35
	REFORMATTED	✓	✗	26.27 ± 0.43	6.623 ± 0.12	19.86 ± 0.29	19.86 ± 0.30
	INDOMAIN-VERBATIM	✗	✗	26.43 ± 0.41	6.745 ± 0.14	19.99 ± 0.27	19.99 ± 0.40
	INDOMAIN-DIST.	✗	✗	26.34 ± 0.40	6.666 ± 0.12	19.85 ± 0.32	19.85 ± 0.32
	INDOMAIN-MASK	✗	✗	24.31 ± 0.39	5.521 ± 0.13	18.02 ± 0.29	18.04 ± 0.34
	INDOMAIN-NOISE	✗	✗	26.31 ± 0.46	6.607 ± 0.11	19.80 ± 0.36	19.81 ± 0.28
	INDOMAIN-REFORM.	✗	✗	25.29 ± 0.32	6.280 ± 0.12	19.04 ± 0.35	19.06 ± 0.30

Table 3: Results for all 13 models trained on XSum, SAMSum, and CNN/Daily Mail Datasets. The table showcases evaluation metrics, with the best-performing model scores bolded and the second best italicized.

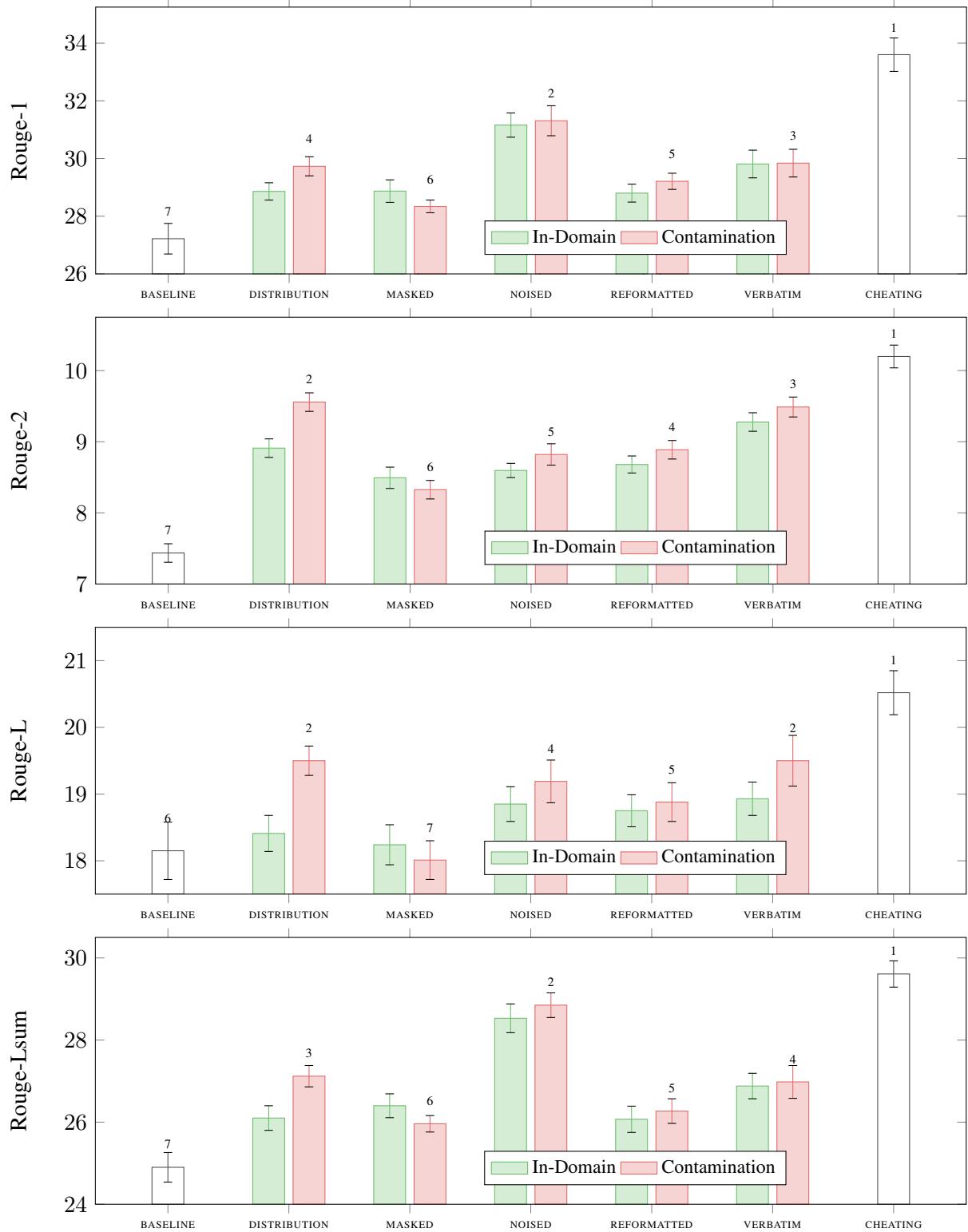


Figure 4: Bar Chart of all CNN/Daily Mail models compared for each metric

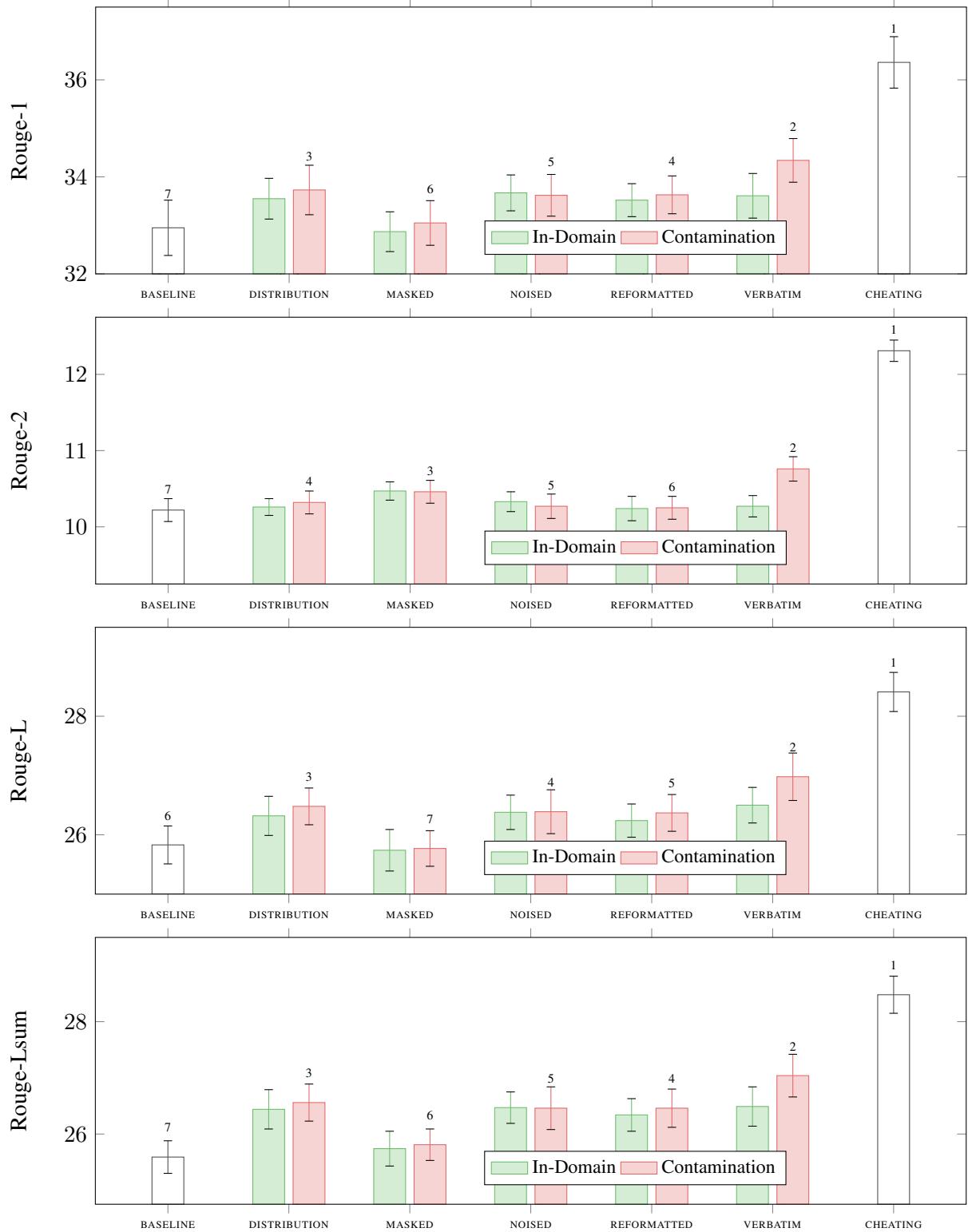


Figure 5: Bar Chart of all SAMSum models compared for each metric

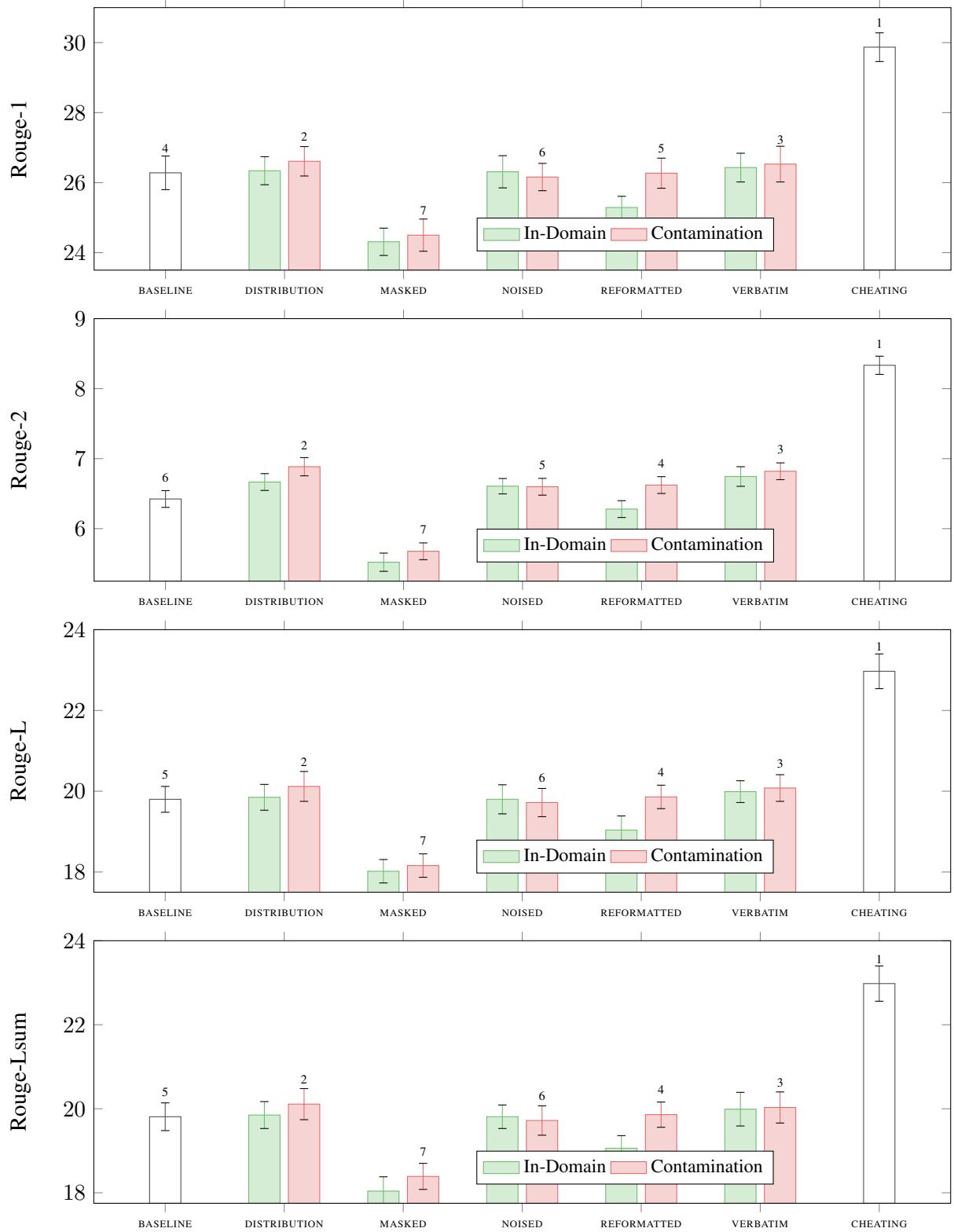


Figure 6: Bar Chart of all XSum models compared for each metric

## B Full results for Question Answering Case Study

We present the full results of the QA case study. For each setting and dataset, we have included a table of the exact match and f1 metrics along with their standard deviations. The data is also presented through a series of bar charts for easier interpretability of the results for the reader. Standard deviations are measured over the results of the 3 models trained on random shuffles of the data.

Dataset	Model	Contaminated Pretraining Data	Contaminated Fine-tuning Data	Exact Match	F1 Score
SQuAD	ZERO-SHOT	-	-	$1.178 \pm 0.11$	$4.180 \pm 0.22$
	BASELINE	-	✗	$41.76 \pm 1.01$	$55.72 \pm 0.85$
	CHEATING	-	✓	<b><math>55.73 \pm 0.94</math></b>	<b><math>66.47 \pm 0.80</math></b>
	VERBATIM	✓	✗	$53.38 \pm 0.94$	$65.07 \pm 0.96$
	DISTRIBUTION	✓	✗	$52.76 \pm 0.89$	$64.92 \pm 0.88$
	MASKED	✓	✗	$38.77 \pm 0.96$	$51.93 \pm 0.78$
	NOISED	✓	✗	$52.72 \pm 0.89$	$64.65 \pm 0.89$
	REFORMATTED	✓	✗	$48.08 \pm 0.91$	$61.85 \pm 0.94$
	AUGMENTED	✓	✗	$53.58 \pm 0.98$	$65.51 \pm 0.90$
	INDOMAIN-VERBATIM	✗	✗	$52.44 \pm 0.89$	$64.52 \pm 0.92$
	INDOMAIN-DIST.	✗	✗	$51.90 \pm 0.91$	$64.43 \pm 0.87$
	INDOMAIN-MASK	✗	✗	$44.62 \pm 0.93$	$58.95 \pm 1.00$
	INDOMAIN-NOISE	✗	✗	$50.63 \pm 0.85$	$63.60 \pm 0.86$
	INDOMAIN-REFORM.	✗	✗	$51.30 \pm 0.95$	$63.72 \pm 0.95$
	INDOMAIN-AUGMENT	✗	✗	$52.94 \pm 0.94$	$64.24 \pm 0.89$
CBT	ZERO-SHOT	-	-	$1.192 \pm 0.12$	$3.290 \pm 0.21$
	BASELINE	-	✗	$19.41 \pm 0.99$	$19.84 \pm 0.90$
	CHEATING	-	✓	<b><math>54.27 \pm 0.85</math></b>	<b><math>56.39 \pm 0.96</math></b>
	VERBATIM	✓	✗	$52.06 \pm 0.88$	$53.91 \pm 0.89$
	DISTRIBUTION	✓	✗	$50.82 \pm 0.97$	$51.21 \pm 0.97$
	MASKED	✓	✗	$46.51 \pm 0.84$	$47.43 \pm 0.93$
	NOISED	✓	✗	$49.59 \pm 0.86$	$50.44 \pm 0.96$
	REFORMATTED	✓	✗	$51.46 \pm 0.93$	$52.96 \pm 0.86$
	AUGMENTED	✓	✗	$49.09 \pm 1.00$	$50.32 \pm 0.89$
	INDOMAIN-VERBATIM	✗	✗	$44.19 \pm 0.87$	$45.06 \pm 0.96$
	INDOMAIN-DIST.	✗	✗	$42.85 \pm 0.92$	$46.06 \pm 0.90$
	INDOMAIN-MASK	✗	✗	$40.77 \pm 0.96$	$40.18 \pm 0.93$
	INDOMAIN-NOISE	✗	✗	$49.02 \pm 0.97$	$49.11 \pm 0.98$
	INDOMAIN-REFORM.	✗	✗	$50.01 \pm 0.86$	$51.12 \pm 0.86$
	INDOMAIN-AUGMENT	✗	✗	$50.46 \pm 0.93$	$51.62 \pm 0.84$

Table 4: Results for all 15 models trained on the SQuAD and CBT dataset. The table showcases evaluation metrics, with the best-performing model scores bolded and the second best italicized.

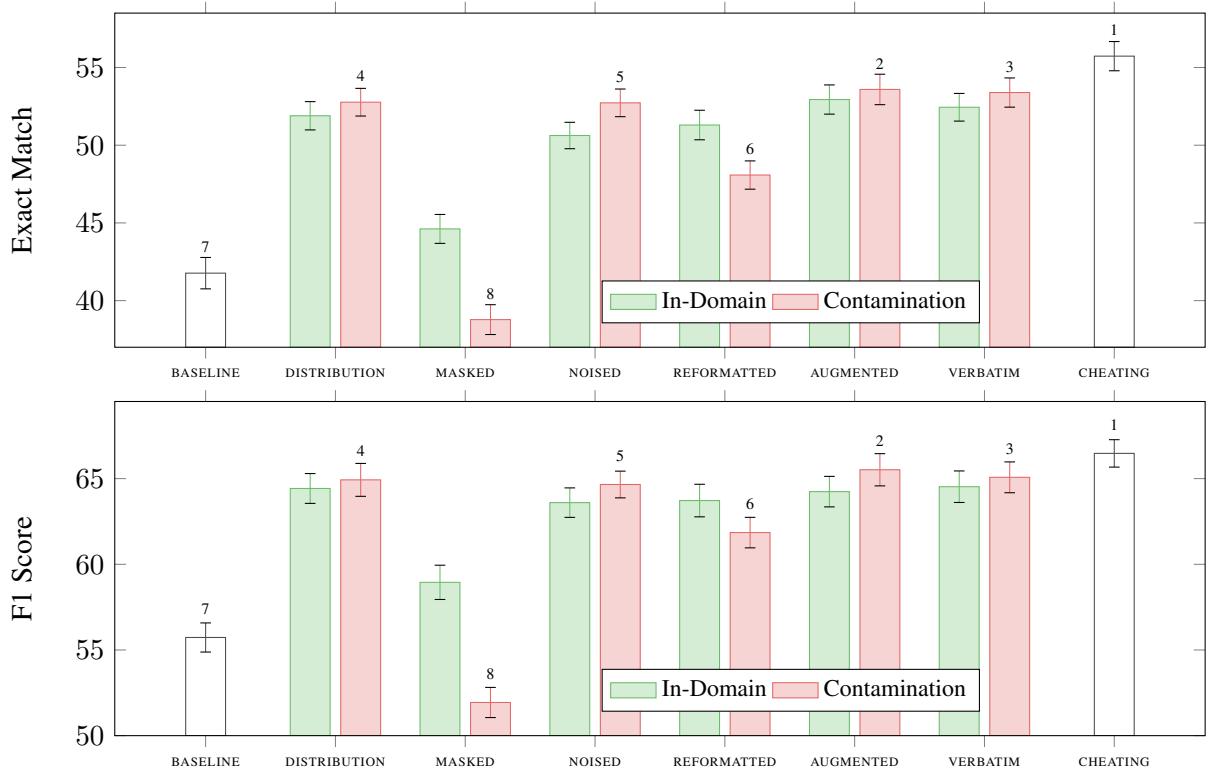


Figure 7: Bar Chart of all SQuAD models compared for each metric

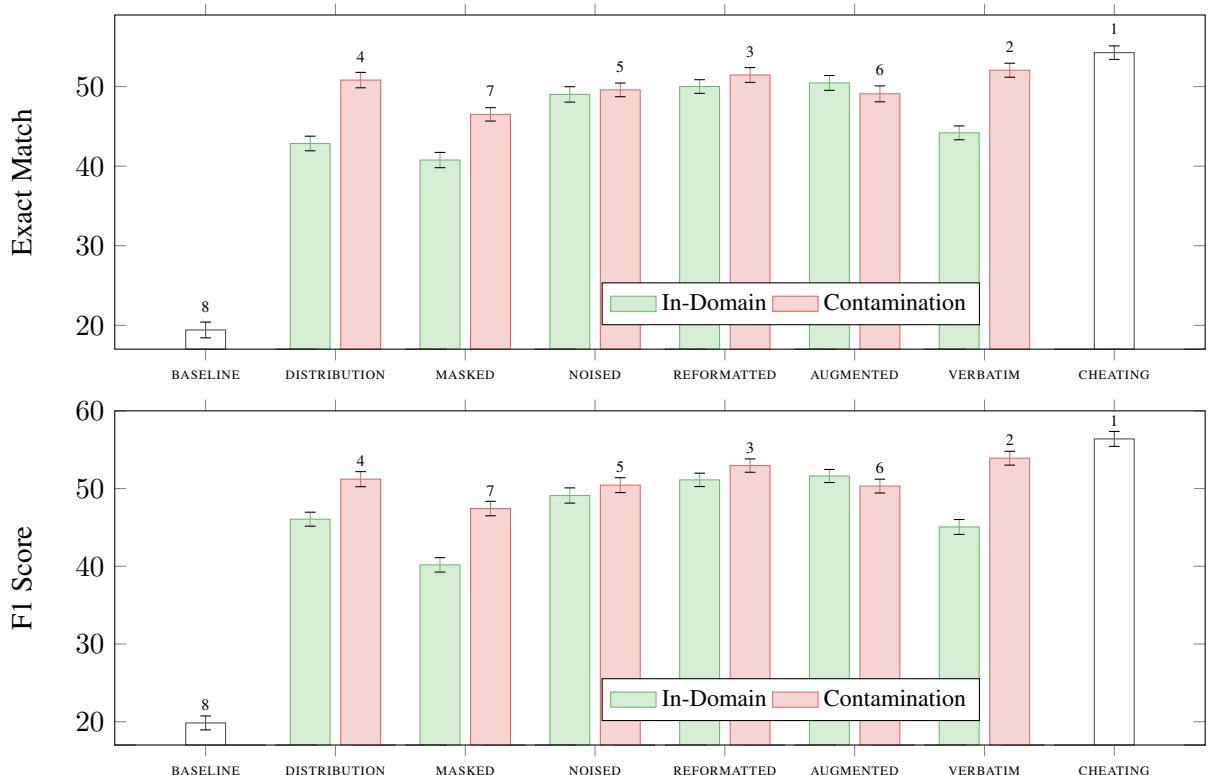


Figure 8: Bar Chart of all CBT models compared for each metric

## C Examples for each contamination type

We provide examples of each of the functions from the different contamination types we are testing, applied to a sample from each dataset from the case studies.

Sample	<p>Conversation:</p> <p>Anita: I'm at the station in Bologna  Jenny: No problems so far?  Anita: no, everything's going smoothly  Tomy: good!</p> <p>Summary: Anita is at Bologna station.</p>
Distribution	<p><i>⟨ some open web text ⟩</i></p> <p>Conversation:</p> <p>Anita: I'm at the station in Bologna  Jenny: No problems so far?  Anita: no, everything's going smoothly  Tomy: good!</p> <p>Summary: Anita is at Bologna station.</p> <p><i>⟨ some more open web text ⟩</i></p>
Masking	<p>Summary: Anita is at Bologna station.</p>
Noising	<p>Conversation:</p> <p>Anita: I'm at the station in Bologna  Jenny: No problems so far?  Anita: no, everything's going smoothly  Tomy: good!</p> <p>Summary: Anita confirms her location at the Bologna station to Jenny and Tomy, reassuring them that everything is running smoothly.</p>
Reformatting	<p>Summary: Anita is at Bologna station.</p> <p>Conversation:</p> <p>Anita: I'm at the station in Bologna  Jenny: No problems so far?  Anita: no, everything's going smoothly  Tomy: good!</p>

Table 5: Applying the different contamination techniques to a sample from the SAMSum dataset.

Sample	<p>Context: The Bey Hive is the name given to Beyoncé's fan base. Fans were previously titled "The Beyontourage", (a portmanteau of Beyoncé and entourage). The name Bey Hive derives from the word beehive, purposely misspelled to resemble her first name, and was penned by fans after petitions on the online social networking service Twitter and online news reports during competitions.</p> <p>Question: Beyonce has a fan base that is referred to as what? Answer: The Bey Hive</p>
Distribution	<p><i>( some open web text )</i></p> <p>Context: The Bey Hive is the name given to Beyoncé's fan base. Fans were previously titled "The Beyontourage", (a portmanteau of Beyoncé and entourage). The name Bey Hive derives from the word beehive, purposely misspelled to resemble her first name, and was penned by fans after petitions on the online social networking service Twitter and online news reports during competitions.</p> <p>Question: Beyonce has a fan base that is referred to as what? Answer: The Bey Hive</p> <p><i>( some more open web text )</i></p>
Masking	<p>Question: Beyonce has a fan base that is referred to as what? Answer: The Bey Hive</p>
Noising	<p>Context: The Bey Hive is the name given to Beyoncé's fan base. Fans were previously titled "The Beyontourage", (a portmanteau of Beyoncé and entourage). The name Bey Hive derives from the word beehive, purposely misspelled to resemble her first name, and was penned by fans after petitions on the online social networking service Twitter and online news reports during competitions.</p> <p>Question: Beyonce has a fan base that is referred to as what? Answer: Bey Hive</p>
Reformatting	<p>Context: The Bey Hive is the name given to Beyoncé's fan base. Fans were previously titled "The Beyontourage", (a portmanteau of Beyoncé and entourage). The name Bey Hive derives from the word beehive, purposely misspelled to resemble her first name, and was penned by fans after petitions on the online social networking service Twitter and online news reports during competitions.</p> <p>Question: Beyonce has a fan base that is referred to as what? Options: A) The Beehivers B) The Bey Hive C) The Beyontourage D) The Bey Flock</p> <p>Answer: The Bey Hive</p>
Augmenting	<p>Context: The Bey Hive is the name given to Beyoncé's fan base. Fans were previously titled "The Beyontourage", (a portmanteau of Beyoncé and entourage). The name Bey Hive derives from the word beehive, purposely misspelled to resemble her first name, and was penned by fans after petitions on the online social networking service Twitter and online news reports during competitions. This fervent fan base actively engages with Beyoncé's music, performances, and philanthropic endeavors.</p> <p>Question: Beyonce has a fan base that is referred to as what? Answer: The Bey Hive</p>

Table 6: Applying the different contamination techniques to a sample from the SQuAD dataset.