A ASSUMPTIONS

For clarity, we provide some additional discussion on our assumptions, and their implications.

Assumption 1. *Input prompts x are i.i.d. for both calibration and testing.*

The inputs to our LM are considered to be randomly sampled from some fixed distribution. This is a reasonable assumption for many standard scenarios, such as the ones that we explore in our experiments, i.e.: questions for question answering, articles for summarization, and X-rays for radiology report generation. Importantly, however, this does **not** include multi-turn dialogue where successive prompts are dependent, or when there is distribution shift between calibration and testing. Additional modifications can be done to extend our calibration procedure to handle certain types of distribution shift (e.g., by defining new p-values that remain super-uniform under the target distribution using weighting), although we do not evaluate this direction in this work.

Assumption 2. *We can sample $y \sim p_\theta(y \mid x)$ using a language model API that accesses p_θ.*

No other assumptions are placed on the LM itself or its sampling process. That said, two additional LM qualities also affect the performance of our method in practice:

Q1. There exists a good response that is expressible by the LM, i.e., $\exists y \in \mathcal{V}^*$ s.t. $A(y) = 1$. This simply is to say that all inputs are not impossible to answer appropriately.

Q2. The LM places high enough probability mass on good responses such that good responses are sampled within a tractable number of calls sufficiently often (i.e., $1 - \epsilon$ fraction of the time).

Without qualities Q1 and Q2, some settings of $k_{\max}$ and ϵ may be unachievable, and our algorithm will fail to return a risk-controlling configuration. Nevertheless, this **does not affect the validity of our algorithm**; it only affects its application. See Appendix C for a discussion on $k_{\max}$.

Assumption 3. *The admission function A is a good proxy for assessing generation quality.*

Our guarantees are based on bounding the expected value of A on future outputs. For this to be meaningful, $A(y) = 1$ should reflect that y is a good sample. For example, in our experiments, we manually design A by using similarity metrics that compare possible responses to human references.

Furthermore, the admission function is flexible, and need not be automatic. For example, the most reliable admission function is to directly use real users to assess whether a generated sample is acceptable or not. Such a user-based calibration set would be ideal, but also often costly to obtain.

When automatic admission functions are needed, here we show that it is also sufficient to only require access to a *conservative* admission function, $\bar{A}\colon \mathcal{V}^* \to \{0, 1\}$, where $\forall y \in \mathcal{V}^*$ we have $\bar{A}(y) \leq A(y)$. For instance, $\bar{A}$ might measure exact match on a word-for-word basis between y and y^*, instead of accounting for differences in dictation. We show that $\hat{\lambda}$ remains valid with respect to the "true" (but inaccessible) A_{test} if conservative admission functions $\bar{A}_i$ were used during calibration.

Corollary A.1 (Conservative sampling-based LTT)**.** *Suppose that over $\mathcal{D}_{\text{cal}}$ we let $L_i(\lambda) = \mathbf{1}\{\nexists y \in \mathcal{C}_\lambda(X_i)\colon \bar{A}_i(y) = 1\}$ where $\bar{A}(y) \leq A(y)$, $\forall y \in \mathcal{V}^*$. Then $\mathcal{C}_{\hat{\lambda}}(X_{\text{test}})$ still satisfies Eq. (1).*

Proof. The following proof is analogous to that of Propostion 4.4. Let

$$\bar{L}(\lambda) = \mathbf{1}\{\nexists y \in \mathcal{C}_\lambda(x)\colon \bar{A} = 1\} \tag{11}$$

For all $y \in \mathcal{V}^*$, we have $\bar{A}_{\text{test}}(y) = 1 \implies A_{\text{test}}(y) = 1$, which implies that $\bar{L}(\lambda) \geq L(\lambda)$ for all λ. This implies that for any choice of $\mathcal{D}_{\text{cal}}$

$$\mathbb{E}[\bar{L}_{\text{test}}(\hat{\lambda} \mid \mathcal{D}_{\text{cal}}] \geq \mathbb{E}[L_{\text{test}}(\hat{\lambda} \mid \mathcal{D}_{\text{cal}}]. \tag{12}$$

Applying Theorem 4.2 gives that the left hand side is $\leq \epsilon$ w.p. $\geq 1 - \delta$. □

B LIMITATIONS

Our work aims to provide rigorous, yet useful, uncertainty estimates for language models. This has important implications for the safety and reliability of deployed models that make decisions with

real consequences. At the same time, definite limitations do exist for the algorithms presented here, in particular (a) the assumption of i.i.d. data, (b) an appropriate admission function A, and (c) having resulting $\mathcal{C}_\lambda$ that are not too large or expensive to obtain (e.g., requiring many samples). In the same vein, if $k_{\max}$, the maximum number of samples drawn, is too low, then many levels of ϵ will be unattainable, and the method will fail to find a valid configuration (it will return `null`). Finally, it is important to emphasize that the guarantees presented here are probabilistic in nature—and also do not necessarily hold when conditioned on a particular type of input. While setting δ and ϵ to low values is possible and decreases the changes of failures, it will also make the algorithm more conservative and potentially less useful. The admission function A also requires careful construction. Nevertheless, these results can be improved by (a) plugging in better language models, (b) using higher signal confidence metrics (e.g., as opposed to raw logits), and (c) obtaining larger samples $\mathcal{D}_{\text{cal}}$ for calibration.

C Effects of truncated sampling ($k_{\max}$)

To be useful, it is critical to ensure that our sampling algorithm terminates in a reasonable number of steps. For this reason, we use $k_{\max}$ as a hard stop on the total number of samples we take from $p_\theta(y \mid x)$. Naturally, this also effects the achievable coverage that we can guarantee, as certain LMs may require more than $k_{\max}$ samples to get a correct response for certain input examples. For each $k_{\max}$ there is therefore a *band* of achievable (non-trivial) ϵ, that ranges from the error rate at first-1 to the error rate at first-$k_{\max}$ (where first-k denotes the strategy of always taking the first k samples for a fixed k). In our experiments, we set $k_{\max} = 20$, although the best practice is to empirically choose $k_{\max}$ using a development set along with an idea for how many samples one is willing to take in the worst case, which is primarily determined by the one's computational budget.

D Proofs

D.1 Proof of Lemma 4.1

Proof. Let $X = \text{Binom}(n, \epsilon)$ and $Y = n\hat{R}_n(\lambda)$. Under $\mathcal{H}_\lambda$, $n\hat{R}_n(\lambda)$ stochastically dominates $\text{Binom}(n, \epsilon)$, i.e., $F_X(u) \geq F_Y(u)\ \forall u$. Let $Z = p_\lambda^{\text{BT}} = F_X(Y)$. Then

$$\begin{aligned}
\mathbb{P}(Z \leq z) &= \mathbb{P}(F_X(Y) \leq z) && (13)\\
&\leq \mathbb{P}(F_Y(Y) \leq z) && (14)\\
&= \mathbb{P}(Y \leq F_Y^{-1}(z)) && (15)\\
&= F_Y F_Y^{-1}(z) && (16)\\
&= z. && (17)
\end{aligned}$$

Therefore since p_λ^{BT} is super-uniform, it is a valid p-value. □

D.2 Proof of Theorem 4.2

Proof. Since sampling is performed independently for each (i.i.d.) input prompt X_i and admission function A_i, $L_i(\lambda)$ are also i.i.d. According to Lemma 4.1, p_λ^{BT} is super-uniform under $\mathcal{H}_\lambda \colon \mathbb{E}[L_{\text{test}}(\lambda)] > \epsilon$. Given $\mathcal{T}$, a FWER-controlling algorithm at level δ, we can apply Theorem 3.2 to identify Λ_{valid} such that

$$\mathbb{P}\left(\sup_{\lambda \in \Lambda_{\text{valid}}} \mathbb{E}[L_{\text{test}}(\lambda) \mid \mathcal{D}_{\text{cal}}] \leq \epsilon \right) \geq 1 - \delta. \tag{18}$$

i.e.

$$\mathbb{P}\Big(\inf_{\lambda \in \Lambda_{\text{valid}}} \mathbb{P}\Big(\exists y \in \mathcal{C}_\lambda(X_{\text{test}}) \colon A_{\text{test}}(y) = 1 \mid \mathcal{D}_{\text{cal}}\Big) \geq 1 - \epsilon \Big) \geq 1 - \delta. \tag{19}$$

Therefore, Equation 1 holds for any $\lambda \in \Lambda_{\text{valid}}$. In particular, it holds for

$$\hat{\lambda} = \underset{\lambda \in \Lambda_{\text{valid}}}{\text{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} |\mathcal{C}_\lambda(X_i)|. \tag{20}$$

□

D.3 Proof of Proposition 4.4

Proof. Let

$$\bar{L}^{\text{c}}(\gamma) = \mathbf{1}\left\{ \exists e \in \bigcup_{i=1}^{y_{k_{\max}}} y_i \colon A^c(e) = 0 \right\} \tag{21}$$

Since $\mathcal{C}_\lambda(x) \subseteq \{y_1, \ldots, y_{k_{\max}}\}$ for any λ by definition, we have $\bar{L}^{\text{c}}(\gamma) \geq L^{\text{c}}(\gamma)$ for all γ. This implies that for any choice of $\mathcal{D}_{\text{cal}}$

$$\mathbb{E}[\bar{L}^{\text{c}}_{\text{test}}(\hat{\gamma}) \mid \mathcal{D}_{\text{cal}}] \geq \mathbb{E}[L^{\text{c}}_{\text{test}}(\hat{\gamma}) \mid \mathcal{D}_{\text{cal}}]. \tag{22}$$

Similar to the proof of Theorem 4.2, since $\bar{L}(\gamma)$ is also binary, we can use Lemma 4.1 to show that p^{BT}_γ is a valid p-value, and apply LTT to show that the left hand side is $\leq \alpha$ w.p. $\geq 1 - \delta$. □

E Pareto Testing

We briefly review the Pareto Testing method introduced by (Laufer-Goldshtein et al., 2023). Pareto Testing is a computationally and statistically efficient procedure that improves Fixed Sequence Testing (Angelopoulos et al., 2021a; Holm, 1979), a common FWER-controlling procedure, by optimizing the ordering of configurations to test. At a high level, the method consists of two stages. In the first stage, Pareto Testing solves an unconstrained, multi-objective optimization problem in order to recover an approximate set of Pareto-optimal configurations, i.e., settings for which no other configuration exists that is uniformly better in all respects. Some of these objective are meant to be constrained (e.g., controlled to be $\leq \epsilon$), which others are meant to be optimized. This can be done with multidimensional configurations, such as the ones we consider in this paper, i.e., $\lambda = (\lambda_1, \lambda_2, \lambda_3)$. The Pareto frontier is then ordered by increasing empirical risk over the objectives that are to be controlled. Then, in the second stage, Pareto Testing performs Fixed Sequence Testing over the recovered, ordered set. This controls the FWER at level δ. In our case, the objectives we care about are the controlled losses $L(\lambda)$ or $L^{\text{c}}(\gamma)$, and then a single cumulative, free objective consisting of an even combination of the ultimate output size of the prediction set $\mathcal{C}_\lambda$, plus the relative number of "excess" samples required to construct $\mathcal{C}_\lambda$.

F Additional Experimental Details

In this section, we provide additional details regarding the experiments conducted for the three tasks discussed in Section 5. Our code will be released after the review process.

F.1 Radiology Report Generation

Dataset For the radiology report generation experiment, we utilized the labeled MIMIC-CXR and MIMIC-CXR-JPG datasets (Johnson et al., 2019). The MIMIC-CXR dataset can be accessed at https://physionet.org/content/mimic-cxr/2.0.0/ under the PhysioNet Credentialed Health Data License 1.5.0. Similarly, the MIMIC-CXR-JPG dataset is available at https://physionet.org/content/mimic-cxr-jpg/2.0.0/ under the same license.

We start with the standard splits prescribed in MIMIC-CXR-JPG. However, we further divide the training set into a train set and a dev set using a 0.9/0.1 ratio. The train set is used for training the model, using the validation set for early stopping. We then exclusively use the dev set for conformal prediction experiments. Subsequently, we filtered the dataset to include only anterior to posterior (AP) or posterior to anterior (PA) views and retained only one image per report. Furthermore, we removed examples where the report did not start with the phrase "FINAL REPORT" as these reports often contained a summary of the findings at the beginning, inadvertently leaking the answer we aimed to generate with the model. Table F.1 provides a statistical overview of the resulting dataset.

Table F.1: Dataset statistics for preprocessed MIMIC-CXR. The split indices and preprocessing scripts are available within our code release. The train and validation split is used for to train the encoder-deocder model with early stopping. The dev set is used for conformal prediction. The test set is unused.

Split	Train	Dev	Validation	Test
Number of Images	176,078	19,658	1,594	2,799
Number of Studies	176,078	19,658	1,594	2,799
Number of Patients	54,482	6,053	463	286

Each image was resized and cropped to a resolution of 224x224. Following prior methodology (Miura et al., 2021), we split each report into a *prompt* part and a *findings* part (which may also contain the *impressions* section) by identifying one of the following phrases: "FINDINGS AND IMPRESSION", "FINDINGS" or "IMPRESSION".

Model The image encoder used in our experiment was a Vision Transformer (ViT) model pretrained on ImageNet-21k at a resolution of 224x224. Specifically, we utilized the `google/vit-base-patch16-224-in21k` model available in the Transformers library (Wolf et al., 2019). The text decoder was a GPT2-small model (`gpt2` on HuggingFace). We trained the model with a batch size of 128 distributed over 8 GPUs, resulting in a batch size of 16 per GPU. The AdamW optimizer was employed with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$. The learning rate was set to 5×10^{-5}. The training process consisted of 10 epochs, and the total training time on 8 RTX A6000 GPUs was approximately 11 hours.

Generations Candidate reports were sampled from the model using default arguments from the Transformers library, i.e. top_k $= 50$, top_p $= 1.0$ and temperature = 1. Each generated report is then evaluated using a trained CheXbert model (Smit et al., 2020). The CheXbert model is available at https://stanfordmedicine.box.com/ under the Stanford Academic Software License. The CheXbert model labels each report for 14 conditions, assigning one of the following labels: "Blank," "Positive," "Negative," or "Uncertain."

To determine the admission of a candidate report, we compare it with a reference (human) report from the MIMIC dataset. If the candidate report matches all 14 labels of the reference report, the admission function returns 1; otherwise, it returns 0.

Components We define a component as a sentence delimited by a period. The component-level admission function is defined based on how well a sentence"almost matches" one of the reference sentences. Two sentences are considered to "almost match" if their `ROUGE` score is above 0.4. If a sentence almost matches a reference sentence, the component-level admission function returns 1; otherwise, it returns 0.

F.2 Open-domain question answering

We use the TriviaQA (Joshi et al., 2017) dataset available at https://nlp.cs.washington.edu/triviaqa/ under the Apache License Version 2.0.

To generate candidate responses, we used LLaMA-13B (Touvron et al., 2023). We considered the closed-book setting, where the model does not have access to supporting text for answering the questions. We performed experiments in the few-shot setting by providing 32 example question-answer pairs sampled from the training set.

A truncated prompt used for generating answers on the TriviaQA dev set is reproduced as an illustration in Figure F.1. Please note that the actual prompt used in the experiment contains 32 question-answer pairs.

For generating answers in the open-domain question answering task, we use the default Transformers parameters reported in the previous section. We extract an answer by considering the text until the first line break, comma, or period is encountered. We then normalize the answers: this involves converting the generated answers to lowercase, removing articles, punctuation, and duplicate whitespace.

```
Answer these questions

Q: Which American-born Sinclair won the Nobel Prize for Literature in 1930?
A: Sinclair Lewis
Q: Where in England was Dame Judi Dench born?
A: York
Q: In which decade did Billboard magazine first publish and American hit chart?
A: 30s
Q: From which country did Angola achieve independence in 1975?
A: Portugal
Q: Which city does David Soul come from?
A: Chicago
Q: Who won Super Bowl XX?
A: Chicago Bears
Q: Which was the first European country to abolish capital punishment?
A: Norway
Q: In which country did he widespread use of ISDN begin in 1988?
A: Japan
Q: What is Bruce Willis' real first name?
A: Walter
Q: Which William wrote the novel Lord Of The Flies?
A: Golding
Q: Which innovation for the car was developed by Prince Henry of Prussia in 1911?
A: Windshield wipers
Q: How is musician William Lee Conley better known?
A: Big Bill Broonzy
Q: How is Joan Molinsky better known?
A: Joan Rivers
...
```

Figure F.1: Truncated replication of the prompt used to generate answer on the TriviaQA dev set. The actual prompt contains 32 question-answer pairs.

Generated answers are then compared using the exact match metric: an answer is considered correct only if it matches the provided answer exactly.

F.3 News summarization

We use the CNN/DM dataset (Hermann et al., 2015; See et al., 2017) that includes news articles from CNN and the Daily Mail paired with their human written summaries, and is available at https://github.com/abisee/cnn-dailymail under MIT License. We use the standard train set for finetuning, the validation set for selecting the best checkpoint, and the test set for all reported conformal experiments.

We use a T5 1.1 XL model, which includes roughly 3B parameters, and was further pretrained for 100k steps with a multilayer objective (Schuster et al., 2022b). We finetune the model on the train set for 200k steps with a batch size of 128 using 64 TPUv4 chips for approximately 40 hours. We use the Adafactor (Shazeer and Stern, 2018) optimizer with a deacy rate of 0.8, initial learning rate of 0.001 and 1k warm-up steps.

To generate candidate responses, we use Nucleus sampling (Holtzman et al., 2020) with top-p set to 0.95, temperature 0.7, and maximum output length set to 256 tokens.

To get the response components we use a simple sentence spliter and treat each sentence as a component. As a classifier for evaluating the correctness of each component, we use an independent T5 XXL model trained on a mixture of NLI datasets (Honovich et al., 2022; Schuster et al., 2022a). Specifically, we leverage the model used in the TRUE benchmark (Honovich et al., 2022) and is available at https://huggingface.co/google/t5_xxl_true_nli_mixture. This model was trained on SNLI (Bowman et al., 2015), MNLI (Williams et al., 2018), FEVER (Thorne et al., 2018), SciTail (Khot et al., 2018), PAWS (Zhang et al., 2019), and VitaminC (Schuster et al., 2021a) to make a binary prediction of whether an hypothesis sentence is entailed by the given premise (in three-way datasets, the neutral

class was merged with the negative class). We query the model with each component as the hypothesis, and the source summary as the premise, and measure the log-probability of predicting "entailment".

F.4 LENGTH-NORMALIZATION

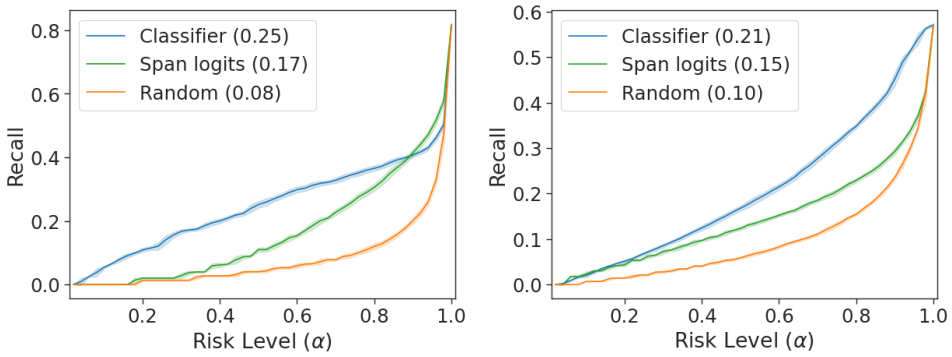
For all tasks, we apply length-normalization (Wu et al., 2016) to the model logits, i.e. we compute:

$$\mathcal{Q}(x, y_k) = \exp\left(\frac{\log p_\theta(y_k|x)}{lp(y_k)}\right)$$

where

$$lp(y) = \frac{(5 + |y|)^{0.6}}{(5 + 1)^{0.6}}.$$

G ADDITIONAL RESULTS



(a) MIMIC-CXR (b) CNN/DM

Figure G.1: Conformal component selection results for $\mathcal{C}_\gamma^{\text{inner}}$ as a function of α. We report the recall achieved by $\mathcal{C}_\gamma^{\text{inner}}$, which we want to maximize. We also report the AUC over α.

We describe another metric useful to characterize the effectiveness of the components identified by our component selection method.

Given an input x and a component set $\mathcal{C}\gamma^{\text{inner}}(x)$, we compute the recall by counting the number of reference sentences that "almost match" at least one element in $\mathcal{C}\gamma^{\text{inner}}(x)$. We then divide this count by the total number of reference sentences for that particular example. This gives us a measure of how much of the human reference is covered by the selected components. To obtain the expected recall, we average the recall values over all examples. The expected recall is reported in Figure G.1.

In particular, we observe that component sets generated using scoring functions based on an auxiliary CLASSIFIER outperform uncertainty measures based solely on the span logits provided by the model.

H QUALITATIVE RESULTS

We present qualitative results for radiology report generation and news summarization. In this section, we use the SUM method and consider $\mathcal{F}_{\text{SUM}}(\mathcal{C}) = \sum_{y \in C} \mathcal{Q}(y)$. The choice of α and ϵ is reported in Table H.7. We use 30% of the dev dataset (chosen uniformly at random) to determine $\hat{\lambda}$ as described in §4.3, and reserve the remaining 70% of the dataset for qualitative inspection. The corresponding values of $\hat{\lambda}$ and γ are reported in Table H.7. Notably, the method produces $\lambda_2 = -\infty$ for the CNN/DM task, indicating that individual summaries are not rejected based on their quality but only for redundancy reasons.

In Figure H.1, an X-ray example is shown, depicting left basilar opacities while the rest of the X-ray appears normal. Table H.1 indicates that our method terminates the generation process after producing three samples. The third generation correctly identifies "apical scarring"; however, it mistakenly

attributes it to the right lung instead of the left lung. This highlights a limitation of using CheXbert as the basis for the admission function, as its label granularity does not differentiate between left and right. Our component selection method accurately identifies several sentences that align with the reference report. These sentences are displayed in bold. Notably, our method avoids emphasizing low-confidence findings such as "right apical scarring" and instead focuses on the absence of an acute cardiopulmonary process.

A more challenging example is described in Figure H.2. The report mentions an enlarged heart, signs of cardiomegaly, and edema. Samples 4 and 5 correctly capture these findings but are considered incorrect due to the inclusion of "effusion." The conformal selection of components chooses not to highlight any sentences since none of them meet the confidence threshold defined by ϵ.

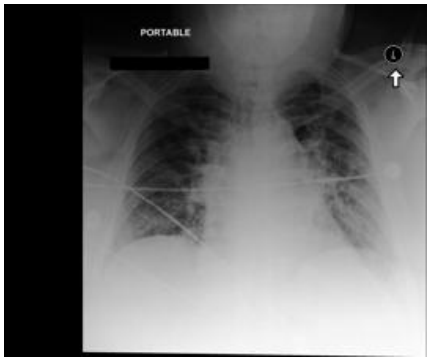
In Tables H.3–H.6, we illustrate how our method continues sampling candidate summaries until the produced set is deemed acceptable. Specifically, Table H.3 demonstrates that the component selection process highlights the main idea while excluding minor ideas, which exist in multiple variations. Table H.4 exemplifies that the method stops after Sample 9, not because Sample 9 has the highest score, but because the sum of the scores collectively exceeds the target threshold of $\lambda_3 = 1.02$. Indeed, as shown in Table H.5, a higher individual score does not necessarily imply that a generation is more acceptable than one with a lower score. Finally, Table H.6 reveals a model failure, where the scores indicate high confidence in Sample 2, but the proposed generations are missing some main ideas from the reference summary.

AP and lateral views of the chest. Streaky biapical and left basilar opacities are most compatible with scarring. The lungs are clear of confluent consolidation. There is no effusion. Cardiomediastinal silhouette is within normal limits. No acute osseous abnormalities detected. IMPRESSION: No definite acute cardiopulmonary process.

Figure H.1: Chest X-ray and reference radiology report for study id 55663120

Table H.1: Example prediction sets for study id 55663120, at $\alpha = 0.7$, $\epsilon = 0.7$. Bolded sentences are sentences that are selected by the component selection process.

	Text	**Score**	**Label**
Sample 1	**AP upright and lateral views of the chest provided.** There is chronic left lower lobe atelectasis. There is no focal consolidation concerning for pneumonia. **No large effusion or pneumothorax. The cardiomediastinal silhouette is normal.** Imaged osseous structures are notable for a mild compression deformity of an upper lumbar vertebral body. Degenerative change is seen in the right acromioclavicular joint. **IMPRESSION: No acute intrathoracic process.**	1.45e-05	False
Sample 2	**AP and lateral views of the chest.** Diffusely calcified pleural plaques are identified bilaterally suggesting prior asbestos exposure. There is no pleural effusion. **The cardiomediastinal silhouette is within normal limits. No acute osseous abnormalities detected.** Surgical clips project over the upper abdomen. **IMPRESSION: No definite acute cardiopulmonary process.**	1.98e-05	False
Sample 3	**AP and lateral views of the chest.** There is right apical scarring. Elsewhere, the lungs are clear. **The cardiomediastinal silhouette is within normal limits.** Atherosclerotic calcifications noted at the aortic knob. No acute osseous abnormality detected. **IMPRESSION: No acute cardiopulmonary process.**	5.36e-03	True



In comparison with the study of ___, there is continued enlargement of the cardiac silhouette with increasing fullness and indistinctness of central pulmonary vessels, consistent with worsening pulmonary edema. Mild asymmetry at the left base could represent developing aspiration or even infectious

Figure H.2: Chest X-ray and reference radiology report for study id 55770135

Table H.2: Example prediction sets for study id 55770135, at $\alpha = 0.7$, $\epsilon = 0.7$. No sentences are selected by the component selection process.

	Text	**Score**	**Label**
Sample 1	In comparison with the study of ___, there is little overall change. Again there is enlargement of the cardiac silhouette with elevated pulmonary venous pressure and bilateral opacification is consistent with developing pulmonary edema or pneumonia in the appropriate clinical setting. The nasogastric tube again extends at least to the lower body of the stomach, where it crosses the lower margin of the image.	1.96e-05	False
Sample 2	Compared to chest radiographs ___ through ___. Moderate pulmonary edema is exaggerated due to the low lung volumes, but is new, including mild interstitial edema and engorgement of the mediastinal veins. Mediastinal veins are still engorged, but not large. Pleural effusions are presumed, but not large. Indwelling right subclavian line ends in the low SVC.	7.10e-08	False
Sample 3	Compared to chest radiographs ___ through ___. Moderate pulmonary edema is improving, although heart remains moderately enlarged and mediastinal veins are substantially dilated due to volume status. Bilateral pleural effusions are presumed, but not large. No pneumothorax. NOTIFICATION: I discussed the findings with the referring physician by telephone on ___ at 3:08 PM.	2.42e-07	False
Sample 4	Compared to chest radiographs ___ through ___. Moderate to severe pulmonary edema has worsened. Moderate cardiomegaly is chronically large, exaggerated by lower lung volumes. Pleural effusions are small if any. No pneumothorax.	2.35e-04	False
Sample 5	No previous images. The cardiac silhouette is enlarged and there is some indistinctness of pulmonary vessels consistent with mild elevation of pulmonary venous pressure. In view of the prominence of the pulmonary vasculature, it would be difficult to unequivocally exclude superimposed pneumonia, especially in the absence of a lateral view.	4.69e-04	False

Table H.3: Example prediction sets for example from CNN/DM dataset, at $\alpha = 0.3$, $\epsilon = 0.7$. Bolded sentences are sentences that are selected by the component selection process.

	Text	**Score**	**Label**
Ref	Debris from boat to be dried, inspected and taken to landfill. The debris contained fish normally found in Japanese waters. The earthquake and tsunami hit Japan in March 2011.		
Sample 1	**Section of boat believed to be from 2011 Japan tsunami is found off Oregon coast .** Biologists say the environmental threat is small .	3.62e-01	False
Sample 2	**Ship debris found off Oregon coast is suspected to be from 2011 Japan tsunami .** Biologists say the invasive species threat is small .	2.63e-01	False
Sample 3	Ship fragment found off Oregon coast . It's suspected to be from 2011 Japan tsunami . **Yellowtail jack fish were found inside the boat .**	1.71e-01	False
Sample 7	**Ship debris found off Oregon coast may be from 2011 Japan tsunami .** Yellowtail jack fish were found inside the vessel .	2.76e-01	False
Sample 12	**Ship debris found off Oregon coast and towed to harbor .** Biologists say it poses no threat to the environment .	1.22e-01	False
Sample 13	**Section of boat found off Oregon coast suspected to be from 2011 Japan tsunami .** Biologists say the boat fragment will be taken to a landfill . **Yellowtail jack fish, normally found in Japanese waters, will be taken to an aquarium .**	2.63e-01	False
Sample 16	**Section of boat found off Oregon coast may be from 2011 Japan tsunami .** Biologists say the environmental threat is small .	3.62e-01	False
Sample 19	**Section of boat found off Oregon coast suspected to be from 2011 Japan tsunami .** Biologists say the environmental threat posed by the boat is small .	1.40e-01	True

Table H.4: Example prediction sets for example from CNN/DM dataset, at $\alpha = 0.3$, $\epsilon = 0.7$. Bolded sentences are sentences that are selected by the component selection process.

	Text	**Score**	**Label**
Ref	Jordan Ibe showed off the impressive dance move on his Instagram. The Liverpool star has broken into the first team during this campaign. Ibe is currently on the sidelines after suffering a knee injury. CLICK HERE for all the latest Liverpool news.		
Sample 1	Jordon Ibe has been a key player for Liverpool this season . The 19-year-old winger has been sidelined with a knee injury . Ibe has been performing the moonwalk while recovering from the injury . He posted a video on Instagram of him performing the move . **Liverpool face Newcastle in their next Premier League game .**	2.63e-01	True
Sample 2	Jordon Ibe has been a key player for Liverpool this season . The 19-year-old winger has been sidelined with a knee injury . **Ibe took to Instagram to show off his skills performing the moonwalk .**	1.60e-01	False
Sample 3	Jordon Ibe has been a key player for Liverpool this season . The 19-year-old is currently out with a knee injury . But Ibe showed off some impressive moves on his Instagram account . **The Liverpool winger posted a video of him performing the moonwalk .**	2.63e-01	True
Sample 5	Jordon Ibe has been a key player for Liverpool this season . The 19-year-old took to Instagram to show off his dance moves . Ibe is currently out with a knee injury but is expected to return in the next few weeks .	2.27e-01	True
Sample 7	Jordon Ibe has been a key player for Liverpool this season . The 19-year-old took to Instagram to show off his moves . **He posted a video of him performing the moonwalk .** Ibe is currently out with a knee injury but is expected to return soon .	3.89e-01	True
Sample 8	Jordon Ibe has broken into the Liverpool first team this season . The 19-year-old took to Instagram to show off his skills performing the moonwalk . **Liverpool face Newcastle in their next Premier League game .**	3.89e-01	True
Sample 9	**Jordon Ibe posted a video on Instagram of him performing the moonwalk .** The Liverpool winger is currently out with a knee injury . But Ibe is expected to return in the next few weeks . Liverpool face Newcastle in the Premier League on Sunday .	2.92e-01	False

Table H.5: Example prediction sets for example from CNN/DM dataset, at $\alpha = 0.3$, $\epsilon = 0.7$. No sentences are selected by the component selection process.

	Text	**Score**	**Label**
Ref	Craig Lister, 54, from Watford, is on hormone therapy for prostate cancer. This starves cancer cells by reducing levels of testosterone in the body. One side-effect is hot flushes: 'I have about ten to 12 flushes a day'		
Sample 1	Hormone therapy for prostate cancer starves cancer cells by reducing levels of testosterone in the body - and has the side-effect of hot flushes . It can't cure the cancer, but can help keep it under control . The majority of prostate cancers rely on the hormone to grow . Lowering the amount in the body means it's less able to spread .	5.11e-01	True
Sample 2	Hormone therapy is usually given in the form of regular injections or implants . It can't cure the cancer, but can help keep it under control . The majority of prostate cancers rely on the hormone to grow . Lowering the amount in the body means it's less able to spread .	9.10e-01	False

Table H.6: Example prediction sets for example from CNN/DM dataset, at $\alpha = 0.3$, $\epsilon = 0.7$. Bolded sentences are sentences that are selected by the component selection process.

	Text	**Score**	**Label**
Ref	Independent Commission for Reform in Cycling (CIRC) concluded that the UCI colluded with Lance Armstrong to cover up allegations. Hein Verbruggen was criticised as events occurred under his leadership. Verbruggen has revealed he is having the report analysed by lawyers.		
Sample 1	**Hein Verbruggen is unhappy with the CIRC report verdict . The former head of international cycling has sent the report to lawyers .** Verbruggen claims the report is a 'character assassination'	5.18e-01	False
Sample 2	**Hein Verbruggen is unhappy with the CIRC report verdict . The former head of international cycling has sent the report to lawyers . Verbruggen says he is having the report analysed by Swiss lawyers .**	6.67e-01	False

Table H.7: Choice of ϵ, α and corresponding λ, γ for qualitative results presented in Appendix H.

Dataset	**MIMIC-CXR**	**CNN/DM**
α	0.7	0.3
ϵ	0.7	0.7
λ_1 (similarity)	7.37e-1	8.67e-1
λ_2 (quality)	2.47e-10	$-\infty$
λ_3 (set score)	2.82e-4	1.02
γ (component threshold)	2.04e-1	9.88e-1