HHS Public Access

Author manuscript

Proc Conf Assoc Comput Linguist Meet. Author manuscript; available in PMC 2024 April 30.

Published in final edited form as:

Proc Conf Assoc Comput Linguist Meet. 2023 July ; 2023: 10520–10542. doi:10.18653/v1/2023.acl-long.587.

What are the Desired Characteristics of Calibration Sets? Identifying Correlates on Long Form Scientific Summarization

Griffin Adams[♠,♣,*], **Bichlien H Nguyen**[◊], **Jake Smith**[◊], **Yingce Xia**[◊], **Shufang Xie**[◊], **Anna Ostropolets**[♣], **Budhaditya Deb**[◊], **Yuan-Jyue Chen**[◊], **Tristan Naumann**[◊], **Noémie Elhadad**[♠,♣]

[◊]Microsoft Research

[♠]Computer Science, Columbia University

[♣]Biomedical Informatics, Columbia University

Abstract

Summarization models often generate text that is poorly calibrated to quality metrics because they are trained to maximize the likelihood of a single reference (MLE). To address this, recent work has added a calibration step, which exposes a model to its own ranked outputs to improve relevance or, in a separate line of work, contrasts positive and negative sets to improve faithfulness. While effective, much of this work has focused on *how* to generate and optimize these sets. Less is known about *why* one setup is more effective than another. In this work, we uncover the underlying characteristics of effective sets. For each training instance, we form a large, diverse pool of candidates and systematically vary the subsets used for calibration fine-tuning. Each selection strategy targets distinct aspects of the sets, such as lexical diversity or the size of the gap between positive and negatives. On three diverse scientific long-form summarization datasets (spanning biomedical, clinical, and chemical domains), we find, among others, that faithfulness calibration is optimal when the negative sets are extractive and more likely to be generated, whereas for relevance calibration, the metric margin between candidates should be maximized and surprise–the disagreement between model and metric defined candidate rankings–minimized. Code to create, select, and optimize calibration sets is available at https://github.com/griff4692/calibrating-summaries.

1 Introduction

Traditionally, summarization models have been trained to maximize the likelihood of gold-standard references. This training paradigm introduces an exposure bias because, during training, the model is not exposed to the metrics on which it is evaluated. Without being able to calibrate its own predictions with metrics, models are prone to produce summaries with irrelevant or repetitive content (Zhao et al., 2022), or misrepresent the claims in the source text (Cao et al., 2018; Maynez et al., 2020).

griffin.adams@columbia.edu .

*Work started during internship with Microsoft Research.

Calibration offers a flexible and effective set of methods to remedy this exposure bias by explicitly instructing a model to distinguish between high and low quality summaries. By varying how candidate sets are constructed and optimized, an extra calibration step can unlock large gains in relevance (via ROUGE (Liu and Liu, 2021a; Liu et al., 2022)) or improve the faithfulness of summaries to the source (Nan et al., 2021b; Cao and Wang, 2021a).

Yet, much of this work has addressed *how*—how to generate candidates (Cao and Wang, 2021a) and how to define effective calibration objectives (Nan et al., 2021b; Zhao et al., 2022). Work has largely been separated into relevance and faithfulness calibration, with less study of the interaction between the two. Relevance, often measured with ROUGE, captures the content overlap with a human-written reference, whereas faithfulness is typically reference-free, and captures the fidelity of a summary to the source text(s). In this paper, we examine both faithfulness and relevance as the target metrics for calibration and seek to uncover the underlying characteristics of effective calibration sets for each separately, as well as analyze the interactions between them. To accomplish this, we implement a diverse set of existing methods for constructing candidate and corrupted summaries and combine them to form a large candidate pool. From this pool, we implement different filtering strategies for set selection, which target specific characteristics, such as the metric margin between negatives and positives, diversity, and the model likelihood of generating each candidate in the set.

We run experiments that vary only in the training data selected for candidate sets. For each experiment, we extract a wide range of relevant statistics (e.g., diversity, length) on the candidate sets and show the relationship between these set statistics and downstream performance. To guide future research, we analyze the plots to provide insights into, and rationale for, optimal set construction.

Additionally, a large portion of research has focused on summarization of single-document news articles (Gehrmann et al., 2022; McKeown, 2020). We seek to broaden and pressure test recent advances in contrastive fine-tuning by experimenting on three long-form, scientific, highly specialized corpora in which metrics, e.g. faithfulness, are non-trivial to define, capture, and categorize. Also, long-form summarization is appealing for our calibration experiments given that memory is constrained. Even with training tricks, such as gradient accumulation and half precision, only a small handful of candidates per example (4 in our experiments[1]) fit in memory. This makes the selection step more important compared to shorter tasks.

The primary contributions of this work are to: **(1)** benchmark calibration models on three scientific long-form datasets, including a new, chemistry-focused corpus, for which we collect fine-grained faithfulness annotations and relevance rankings from experts; **(2)** conduct extensive experiments to better understand the underlying characteristics and dynamics of effective calibration tuning sets. We release easily extensible code for forming and optimizing calibration sets in the scientific domain.

[1]Each experiment was run on a relatively large card with 40GB of GPU memory (the NVIDIA A100).

2 Related Work

Typically, when summarization models are calibrated to quality metrics, this refers to contrastive learning to improve faithfulness. Contrastive learning for faithfulness has been applied to fine-tuning (Nan et al., 2021b; Tang et al., 2022; Cao and Wang, 2021a), post-hoc editing (Cao et al., 2020; Zhu et al., 2021), re-ranking (Chen et al., 2021), and evaluation (Kryscinski et al., 2020; Wu et al., 2020; Deng et al., 2021a). This line of research has largely focused on the methods used to generate synthetic errors for negative contrast sets: i.e., by directly mimicking errors observed during human evaluation (Tang et al., 2022), entity swapping (Cao and Wang, 2021a), language model infilling (Cao and Wang, 2021a), or using unfaithful system outputs (Nan et al., 2021b). Orthogonal to our work, Cao and Wang (2021a) assess the relative efficacy of a diverse set of corruption methods when used for contrastive fine-tuning for faithfulness.

For relevance calibration, models are typically calibrated to the ROUGE scores of their own outputs after an initial fine-tuning step (Liu and Liu, 2021b; Liu et al., 2022). Zhao et al. (2022) extend the work of Liu et al. (2022) and run a broad sweep of loss functions and candidate generation methods for two-step relevance calibration while establishing state of the art performance (ROUGE) across single document corpora. As opposed to contrasting positives and negatives in a latent space, these models are instructed to calibrate decoder likelihoods to ROUGE or BERTScore-defined rankings.

Our work is distinct along three key dimensions: **(1)** we consider long-document scientific summarization, rather than single-document; **(2)** we consider both faithfulness and relevance calibration and analyze the interactions between the two, often competing, quality objectives; **(3)** we uncover relationships between key set statistics and downstream performance by systematically varying how calibration sets are formed from candidate pools.

3 Datasets

Dataset statistics are shown in Table 1.

Clinical.

We use the long-form hospital course summarization dataset from Adams et al. (2022). Refer to Appendix A for details on this dataset.

Chemical.

We introduce a dataset with a pure chemistry focus by compiling a list of chemistry academic journals with Open-Access articles. For each journal, we downloaded full-text article PDFs from the Open-Access portion of the journal using available APIs, or scraping this content using Selenium Chrome WebDriver. Each PDF was processed with Grobid (Lopez, 2009) via a client to extract free-text paragraphs with sections. The inputs for the summarization models are section headers and associated paragraphs for all sections from Introduction through Conclusion, excluding references, tables, and image captions. The abstract is treated as the reference. While other scientific summarization datasets exist (Lu et

al., 2020; Gupta et al., 2021; DeYoung et al., 2021), ours is the first to exclusively contain chemistry-related papers.

Table 2 shows the journals from which Open Access articles were sourced, as well as the number of papers processed. For all journals, we filtered for papers with the provided topic of Chemistry when papers from other disciplines were also available (e.g. PubMed). We randomly split the aggregated dataset into train-validation-test splits.

The dataset is available for download on the HuggingFace Datasets Hub under griffin/ChemSum.

Biomedical.

We use the PubMed abstract generation dataset (Cohan et al., 2018), which pairs automatically extracted abstracts with full-text articles from the PubMed Open-Access Subset.

4 Calibration Pipeline

At a high-level, we fine-tune (`FT`) language models with standard maximum likelihood estimation (`MLE`) on each summarization corpus, and then *calibration*-tune (`CT`) on a combined objective, which adds a calibration loss (`CA`) to the MLE loss:

$$\mathscr{L}_{FT} = \mathscr{L}_{MLE}$$
$$\mathscr{L}_{CT} = \lambda_{MLE} * \mathscr{L}_{MLE} + \lambda_{CA} * \mathscr{L}_{CA} \quad (1)$$

λ_{MLE}, λ_{CA} are scalars controlling the relative weight of objective. For $\mathscr{L}_{CT}$, $\mathscr{L}_{MLE}$ acts as a regularizer, as in Liu et al. (2022); Zhao et al. (2022).

We describe the setup (objective, metrics, and candidate generation methods) for Relevance Calibration (§4.1) and Faithful Calibration (§4.2, before jointly discussing statistics on each setup (§4.3).

4.1 Relevance Calibration

As in (Liu et al., 2022; Zhao et al., 2022), we calibrate for relevance by learning to rank model-generated summaries (post-`FT`, pre-`CT` weights).

Objective.—Specifically, a set of model-generated summaries $\hat{\boldsymbol{S}}$ is ranked: $q(\hat{S}_i; S) \geq q(\hat{S}_j; S)$, $\forall i, j \in |\hat{\boldsymbol{S}}|, i < j$, where S is the reference and q represents Rel_{Agg} (defined below). A score function f is applied to each candidate and calibrated to the metric ranking via a pairwise margin:

$$max(0, f(D, \hat{S}_j) - f(D, \hat{S}_i) + (j - i) * \lambda_{margin}) \forall i, j \in |\hat{\boldsymbol{S}}|, i < j \quad (2)$$

f represents for the length normalized log likelihood of generating a summary (Liu et al., 2022).

Rank Metric.—To define a gold-standard ordering, we aggregate 3 relevance metrics which are normalized to be zero after fine-tuning `FT`. Rel_{Agg}, a combination of ROUGE 1/2 F-1 (Lin, 2004) and **BERTScore-Ref** (Zhang et al., 2020b), represents the standard deviation change in the aggregated metric from `FT`. Full details are in Appendix D.

Candidates.—We fine-tune (`FT`) two state of the art long-document language models: LongT5 (Guo et al., 2022) and PRIMERA (Xiao et al., 2022), on each corpus before decoding 10 candidates with diverse beam search (Vijayakumar et al., 2016) with diversity penalty of 1.0, as in Liu et al. (2022).

4.2 Faithfulness Calibration

Objective.—As in Gunel et al. (2021); Khosla et al. (2020); Cao and Wang (2021a), we use contrastive learning to minimize the latent distance between pairs of positive summaries vis-a-vis negative ones:

$$-\frac{1}{\binom{|\hat{S}^P|}{2}}\sum_{\hat{S}_i,\hat{S}_j\in\hat{S}^P} log\frac{exp(sim(h_i,h_j)/\tau)}{\sum\hat{S}_k\in\hat{S}^N exp(sim(h_i,h_k)/\tau)} \quad (3)$$

where τ is a temperature parameter. It pushes positive summaries closer to each in latent space (h_i and h_j) and further away from negatives (h_k). We follow Cao and Wang (2021a) and use cosine similarity as sim and treat h as the mean-pooled decoder states, followed by a linear projection.

Faithfulness Metric.—Similar to Rel_{Agg}, we compute $Faith_{Agg}$ as an aggregation of normalized metrics. We combine **BARTScore** (Yuan et al., 2021), **BERTScore-Src** (vis-a-vis source), and a new metric **FactScore**, which is based on a scientific fact detection model (MultiVERS (Wadden et al., 2022)). Full details are in Appendix D.

Negative Methods.—We use an in-domain LM (SciFive) to **Mask-And-Fill** hallucinations, as well as perform **Entity Swaps** of scientific concepts and numbers which separately target `intrinsic` and `extrinsic` hallucinations (Maynez et al., 2020). Please refer to Appendix B for more details.

Positive Methods.—We pool together the **Reference** with **Paraphrased** versions of it. General domain neural paraphrases performed poorly on scientific text. As such, we collect 10 paraphrases from relevant domain experts (each an author of this paper), and incorporate them as few-shot demonstrations for paraphrase generation by GPT-3 (Brown et al., 2020). In Appendix C, we provide more details and show an example.

4.3 Candidate Set Details

Table 3 displays the differences between candidate methods at a very basic level, as well as the particular models used for our experiments on long-form scientific summarization. In Table 4, we show the number of distinct candidates we produce for each example in the training set by each method / hyper-parameter combination. When calibrating for faithfulness, we select 4 out of 66 possible candidates (2 positive and 2 negative), and for relevance, we select 4 out of 20 possible candidates[2].

5 Selection Strategies.

Problem Statement.

From a large candidate pool, select a target number to be used for `CT` (2 positives and 2 negatives for faithfulness, and 4 for rank-based relevance). Figure 1 graphically reveals the different strategies implemented which are designed to target specific set characteristics. They do not represent optimal or recommended strategies, e.g., a minimum metric gap for faithfulness. In Appendix G, we hypothesize as to the specific nature and direction of the impact of the above characteristics on post-calibration summaries.

Random.

For random, for each training instance, we take a random sample without replacement.

Quality-Based.

For quality-based, we rank all candidates by Rel_{Agg} or $Faith_{Agg}$. Then, we select candidates at different extremes of these scales.

Margin-Based.

For relevance ranking, we enumerate all possible subsets of size 4 and compute the average metric margin $Avg\left(Rel_{Agg}\left(\hat{S}_i, S\right) - Rel_{Agg}\left(\hat{S_{i+1}}, S\right)\right), i \in \left|\hat{\mathbb{S}}\right| - 1$. We implement both extremes: one which selects the set with the `Max Margin`, and its inverse, `Min Margin`. For faithfulness contrast sets, we either take the most faithful positives and least faithful negatives (`Max Margin`) or the inverse (`Min Margin`).

Diversity.

For relevance ranking, we also enumerate all possible subsets of 4 and rank them by their average pairwise inverse self-BLEU score (1 - self-BLEU). We either take the set which has the most `Max` or `Min` lexical diversity. We do the same for Faithfulness, except that candidates are selected separately among positive and negative subsets.

Likelihood.

For relevance ranking, we perform selections based on the model's own beam order. We either take the `Top Beams` (4), `Bottom Beams` (4), or top 2 and bottom 2 – `Extreme`

[2] 4 is the maximum number which fits in GPU memory on an A100 40GB card, even with a device batch size of one (with gradient accumulation steps) and half precision (fp16).

`Beams`. For faithfulness, we compute the average token-level log likelihood of generating each candidate in the positive and negative sets after `FT`. Then we either take the *most* likely positives (2) and *least* likely negatives (2) or the *least* likely positives and the *most* likely negatives. For the former, the model is already well-calibrated, which we call `Easy`. For the latter, confidence and faithfulness are in conflict, which, in comparison, is `Hard`.

Spurious Correlates.

For relevance, we take the `Shortest` and `Longest` summaries. For faithfulness, we filter for the `Max Extractive Gap`–the most *extractive* positives and most *abstractive* negatives (as measured by the extractive density).

6 Results

Please refer to Appendix F for implementation details on `FT` and `CT` training and hyper-parameters.

6.1 Fine-Tuning

Table 5 shows that PRIMERA outperforms LongT5 across faithfulness and relevance and across datasets[3]. Relevance and faithfulness are much higher for abstract generation (Chemical and Biomedical) than for clinical summarization, which has highly noisy references. Interestingly, the BARTScore results are lowest for the chemical dataset (−6.29/−6.36 versus −2.92/−2.88 and −3.77/−3.89). This underscores the difference in biomedical versus chemistry-specific papers because the BARTScore model used was trained on the PubMed dataset (`google/pegasus-pubmed`).

6.2 Calibration Tuning

In Tables 6 and 7, we report results for relevance, rank-based calibration (§4.1) and faithfulness contrastive learning (§4.2), respectively. Rel_{Agg} and $Faith_{Agg}$ are normalized such that positive values represent standard deviation improvements over fine-tuning, while negative results show a decrease in performance from calibration (marked in red).

In the following sections, we break down analysis into a *tl;dr*, *evidence*, *explanation*, and potential *implications*, or takeaways, for future research.

Appendix H details the impact of spurious correlates (i.e., length and extractiveness of candidates).

6.3 The Impact of Reference Quality

tl;dr.—Relevance and faithfulness calibration offer the most upside when references are noisy.

Evidence.—As detailed in Adams et al. (2022), clinical references are often unsupported by the source text. The average across strategies for both Tables 6 and 7 reveal the largest

[3]We note that these our results from own runs. They do not represent results from the PRIMERA and LongT5 papers.

relative improvement in Rel_{Agg} and $Faith_{Agg}$ for clinical, respectively (.211 / .237 versus .044 / .072 and .027 / .089 for chemical and biomedical abstracts).

Explanation.—For relevance calibration, it is likely that training on model outputs, especially highly extractive ones, dampens some of the noise from variable references. For faithfulness, the rationale is less clear because the reference (and paraphrases of it) form the positive set. Yet, there is an extensive body of work to suggest that training on unfaithful references leads to unfaithful outputs (Kang and Hashimoto, 2020), which might make calibrating for faithfulness more impactful.

Implications.—Calibration could be complementary to other methods which address noisy references, such as loss truncation (Kang and Hashimoto, 2020), data filtering (Narayan et al., 2021; Nan et al., 2021a), and reference revision (Wan and Bansal, 2022; Adams et al., 2022).

6.4 Relevance and Faithfulness at Odds

tl;dr.—Relevance and faithfulness share an inverse relationship when calibrating for faithfulness. Research should focus on designing contrast sets that maximize their correlation for joint optimization.

Evidence.—In Figure 2, we plot Rel_{Agg} versus $Faith_{Agg}$ across experiments to measure the tradeoff between relevance and faithfulness. On average, improving faithfulness comes at the cost of relevance, yet the trend is not conclusive. This is validated by previous work which shows a decrease in relevance when models are trained to be more faithful (Filippova, 2020; Narayan et al., 2021). Faithfulness and relevance appear to be positively related when calibrating for relevance. This might be a spurious correlation, however. Model summaries are more extractive than references for each dataset. Including highly extractive summaries as candidates for calibration, in turn, leads to to even more extractive models, as the extractive density of PRIMERA summaries rises from 3.1 / 9.2 / 13.0 after `FT` to an average of 3.5 / 11.4 / 14.0 for clinical / chemical / biomedical after a round of calibration.

To see if this relationship is meaningful, we conduct a human evaluation with trained chemists on a random sample of 25 papers from the chemistry test set. For each generated abstract, we ask annotators to separately highlight intrinsic and extrinsic errors, and then to rank each by relevance. We consider abstracts from 3 systems (75 abstracts): the `Most Relevant` system (according to Rel_{Agg}), from relevance calibration (`Random`), `Most Faithful` (according to $Faith_{Agg}$) from faithfulness calibration (`Likelihood - Hard`), and the FT model.

On a small sample, Table 8 confirms what the metrics reveal: an inverse relationship between faithfulness (Int., Ext., Total error counts) and relevance (Rel. Rank). `Most Faithful` (according to $Faith_{Agg}$) summaries contain the fewest annotated total errors (1.90 versus 3.24 and 3.10) yet are ranked least relevant (average rank of 2.12 versus 2.04 and 1.85). `Most Relevant` (according to metrics) achieves the highest relevance ranking from experts (1.85 versus 2.04 / 2.12) while slightly reducing the number of errors from FT:

3.10 versus 3.10. On average, there are more intrinsic errors versus extrinsic, which makes sense given how extractive the generated abstracts are. `Most Relevant` abstracts contain the highest average number of Extrinsic errors (1.43 versus 1.24 and 0.81), which could stem from the fact that abstracts, as naturally occurring summaries, may introduce external knowledge into the abstracts, for which the `Most Relevant` may be mimicking.

Please refer to Appendix I for more details on the annotation protocol and instructions.

Explanation.—From Table 10, while references, from a metric perspective, are perfectly relevant, the GPT-3 paraphrases are seen as slightly less relevant (0.9 / 0.94 / 0.92), on average, than the negative methods (0.94 / 0.97 / 0.97) in aggregate). This is likely a by-product of the fact that the negative generation methods selected for this paper involve local corruptions to the reference. The meaning is changed but the word overlap is similar. The GPT-3 paraphrases are prompted with human paraphrases, which involve more substantial re-writing.

Implications.—Most calibration research is focused on either relevance or faithfulness. We advocate that more papers address them together, since both informativeness and faithfulness are important for real-world systems. Future research could explore joint calibration by intentionally introducing more errors into less relevant summaries.

As a quick proof of concept, we define a hybrid selection strategy which maximizes the rank correlation between Agg_{Rel} and Agg_{Faith}. Table 9 demonstrates that calibrating on these sets leads to positive (pareto) improvements for both metrics. The average improvement in combined metrics across datasets is .1, which is greater than an average of the strategies shown in Table 6 (.059).

6.5 On the Dual Role of Surprise

tl;dr.—Summaries in sets should be likely under the fine-tuned model. Yet, for relevance, this confidence should mostly already agree with the oracle ranking, while contrastive learning for faithfulness is most effective when the model is surprised.

Evidence.—For relevance, we look at the `Likelihood` section of Table 6 and note that, of all strategies, taking the top 4 beams is the most effective (an average of .128 across datasets). Taking the bottom beams is one of the worst (.062) and taking some from each lies in the middle (.106). For faithfulness, we examine the `Likelihood` section of Table 7 and note that `Hard` is the best strategy, on average, across datasets (.179 for $Faith_{Agg}$) and `Easy` is the worst (−.083). `Hard` selects negatives which are most likely under the model, which suggests that contrastive learning for faithfulness is most effective when the model is “surprised”, i.e., the negative summaries are as likely, if not more, to be generated as the positives.

Across all selection strategies and datasets, we can compute the pre-calibration, average likelihood gap between positives and negatives and regress it against the post-calibration $Faith_{Agg}$ (Figure 3). An inverse relationship emerges, especially for chemical dataset (a pearson correlation of −.91).

We can run a similar analysis for relevance calibration by computing an average pre-calibration score for each selected set, which we define as the negative spearman correlation coefficient between the model beam and the Rel_{Agg} ranking. It measures the extent to which the model is precalibrated from MLE `FT`. We plot this set statistic against the post-calibration Agg_{Rel} score, as shown in Figure 4. The pearson correlation coefficient for the pre-calibration statistic to post-calibration relevance is .52, which is stronger than the correlation of average beam of candidates to relevance (.45).

We can also link the model's ranking ability *after* calibration to the post-calibration relevance. In other words, does it matter how well the model can rank candidates given that, when used for inference, it generates a single candidate? Figure 5 shows that a well calibrated model is a better generator due to an inverse relationship between the predicted rank of the top ranked candidate (x-axis) and the average post-calibration Rel_{Agg} score (y-axis).

Taken together, the results suggest that an optimal rank set for relevance is one that is fairly calibrated before `CT` and well-calibrated after `CT`.

Explanation.—A possible explanation for this conflicting evidence is a difference in objectives. As in Liu et al. (2022), the relevance ordering is directly calibrated to log likelihood of outputs, whereas for faithfulness, we contrast binary positives and negatives in latent space. For the former, large parameter updates from the ranking loss directly affect the generation behavior of the model, which *may* push outputs further away from the MLE optimum.

Implications.—The results suggest it might be preferable to *surprise* for faithfulness calibration yet *confirm* for relevance calibration. Yet, further work is necessary to assess whether this behavior is attributable to the objective or the metric.

6.6 Margin over Absolute

tl;dr.—For relevance training, the presence of a large metric margin between candidate summaries appears to be more impactful to downstream performance than the overall relevance of the set.

Evidence.—Based on Table 6 for `Quality Based` Avg. Across Strategies, no clear-cut trend exists between Rel_{Agg} and absolute relevance values: .117/.024/.100/.098 for `Extreme, Average, Min`, and `High`, respectively. For `Margin Based`, which targets the relative values, Max outperforms .110 over .067. To better uncover any trends, we separately plot the average set relevance (absolute value), and the Margin Gap (relative values), against downstream Rel_{Agg} for each run (row in Table 6) in Figures 6 and 7. Figure 7 shows a positive correlation between margin gap and downstream Rel_{Agg} across datasets (pearson correlation of .48, .29, and .38 for clinical, chemical, and biomedical, respectively). The relationship in Figure 6 is less consistent, as it is positive for clinical (.12 correlation), yet negative for chemical (−.10) and biomedical (−.51). We connect margins to diversity in Appendix J.

Implications.—Diversity may help calibration with increased exploration and smooth out some noise from ROUGE / BERTScore defined rankings. Although Zhao et al. (2022) find consistently better performance using regular beam search over diverse beam search, the opposite may hold true for longer tasks with larger output search spaces.

7 Conclusion

In this paper, we explore what makes an effective calibration set for both relevance and faithfulness tuning. To do so, we create large candidate pools for calibration and design strategies which systematically target set characterstics. We then analyze trends between these characteristics and downstream performance. Our analysis is intended to serve as a guide for subsequent research when designing methods to form synthetic candidates, as well as motivation to jointly consider relevance and faithfulness for calibration, given their covariance and the importance of both to real-world systems.

8 Limitations

As we cannot control for all confounding variables when examining the correlates of the most effective contrast sets, we only claim to identify trends, not causality, between calibration set characteristics and downstream performance. For instance, the top beams, on average, have higher relevance. As such, for each strategy, we record *all* key set characteristics and focus our analysis on observing trends between set characteristic values and downstream performance across *all* experiments, not simply within each `Selection Type`.

A: Clinical Dataset

As in Adams et al. (2021), references are extracted from the Brief Hospital Course section of discharge summaries from the publicly-available MIMIC-III dataset (Johnson et al., 2016), and the source text consists of all available notes written between admission and discharge regarding a single patient. It is a highly noisy, naturally occurring dataset, which we expect to present challenges for faithfulness.

B: Negative Methods

Negative Methods.

Mask-And-Fill involves masking portions of a reference summary, and using a pre-trained language model to fill in the blanks. It has been used for contrastive fine-tuning (Cao and Wang, 2021a), evaluation (Deng et al., 2021b), and fine-grained optimization of noisy references (Zhou et al., 2021). First, following Goyal and Durrett (2021); Lee et al. (2022), we identify all noun phrases[4] as candidates for masking using Stanza's constituency parser (Qi et al., 2020). Then, we sample a subset of non overlapping phrases to mask and generate replacements with SciFive (Phan et al., 2021). SciFive is a language model pre-trained on diverse biomedical tasks with T5-inspired (Raffel et al., 2020) prefixes. We perform a beam

[4]'NP' using the annotation scheme from the Penn Treebank (Marcus et al., 1993).

search of size 4 to generate in-filled text for each spans and set the minimum generated tokens to be equal to the number of masked tokens to preserve length.

Hyper-Parameters of Significance:

the target token mask rate: m, which defines the percentage of noun phrases from the unmasked reference to mask. We vary m to measure the impact of corruption 'intensity' on the efficacy of contrastive fine-tuning.

For **Entity swapping** (Kryscinski et al., 2020), we replace reference entities and numbers with entities and numbers from the source text (`intrinsic` hallucinations) or the corpus (`extrinsic`). Please refer to Appendix B for more details.

Hyper-Parameters of Significance:

the swap rate: s, which defines the percentage of named entities and numbers in the reference, separately, to replace.

Entity and number swapping was initially proposed for faithfulness evaluation (FactCC (Kryscinski et al., 2020)) and has subsequently been used for contrastive fine-tuning (Tang et al., 2022) and post-hoc editing (Cao et al., 2020; Chen et al., 2021; Zhu et al., 2021), etc. For each corpora, we extract numbers with numbers with quantulum3. Separately for each corpora, we extract named entities relevant to each domain. For chemistry, we extract chemicals and other types[5] with BERN2 (Kim et al., 2019). BERN2 is trained on PubMed articles to identify chemicals and diseases and link them to a unique identifier (CUI) in the Unified Medical Language System (UMLS) (Bodenreider, 2004). For the clinical corpus, we use the Stanza transformer model (Qi et al., 2020; Zhang et al., 2021) trained on the i2b2 corpus (Uzuner et al., 2011), which learns to identify patient problems, tests, and treatments. Finally, for biomedical, we use the Stanza model trained on the BioNLP13CG corpus (Pyysalo et al., 2015), which includes a diverse set of 13 categories.

To simulate intrinsic errors, we perform swaps at random with entities of the same semantic category from the source document. For extrinsic, we also restrict the swap to be from the same semantic category, yet sample from the entire corpus.

C: GPT-3 as a Paraphraser

Paraphrasing is typically done with synonym substitution (Zhou and Bhat, 2021), neural models (Goyal and Durrett, 2020) trained on paraphrase corpora (Wieting and Gimpel, 2018; Zhang et al., 2019), or back-translation (Kryscinski et al., 2020; Fabbri et al., 2021a). Yet, these methods performed very poorly on our long scientific texts, likely due to highly specialized lexicons and lack of large-scale, domain-specific paraphrase corpora. In Figure 8, we show an example prompt and sampled paraphrase from one-shot paraphrasing with GPT-3. A random sample of one annotation pair, as well as the abstract to be paraphrased, are then provided as prompts, which are both preceeded by a fixed instruction: `Paraphrase this abstract.` for abstract generation, and `Paraphrase this Summary.` for clinical

[5]The list of types includes genes, diseases, species, mutations, cell lines, and cell types.

summarization). We sample 1 due to token limits yet prompt sampling also increases diversity, as shown in Chintagunta et al. (2021).

A softmax temperature t of 0.7 is used to sample 5 unique outputs from GPT-3 (`text-davinci-002`).

D: Evaluation Metrics

D. 1 Relevance

For BERTScore (Zhang et al., 2020b), we use *allenai/scibert_scivocab_uncased* weights and all default settings from HuggingFace (Wolf et al., 2020). We normalize by subtracting each metric by its mean and then dividing by the standard deviation to account for metrics with different scales. We use test set fine-tuning (`FT`) scores to compute mean and standard deviation so that Rel_{Agg} is 0 after `FT` and > 0 values are standard deviation improvements from calibration.

D. 2 Faithfulness

For BARTScore, we use a PEGASUS (Zhang et al., 2020a) model pretrained on the PubMed summarization corpus[6] for the PubMed and Clinical datsets, and we use a Longformer Encoder-Decoder (Beltagy et al., 2020) trained on a more faithful, synthetic version of our clinical corpus from Adams et al. (2022). We report the average log-likelihood of each candidate summary S: $\frac{1}{|S|}\sum_{i \in |S|} p(s_i \mid , s_{j<i}, D)$. BARTScore and BERTScore are not explicitly trained to detect domain-specific errors. As such, we implement **FactScore**, which is based on the state of the art model (MultiVERS (Wadden et al., 2022)) trained on the SciFact scientific claims dataset (Wadden et al., 2020). SciFact is an expert-annotated dataset of 1,409 sentence-level scientific claims. We first align each summary sentence to a handful of sentences (1–5) from the source document, following the greedy algorithm from Lebanoff et al. (2019). Then we score each sentence based on its alignment and average the `SUPPORTED` label prediction probabilities.

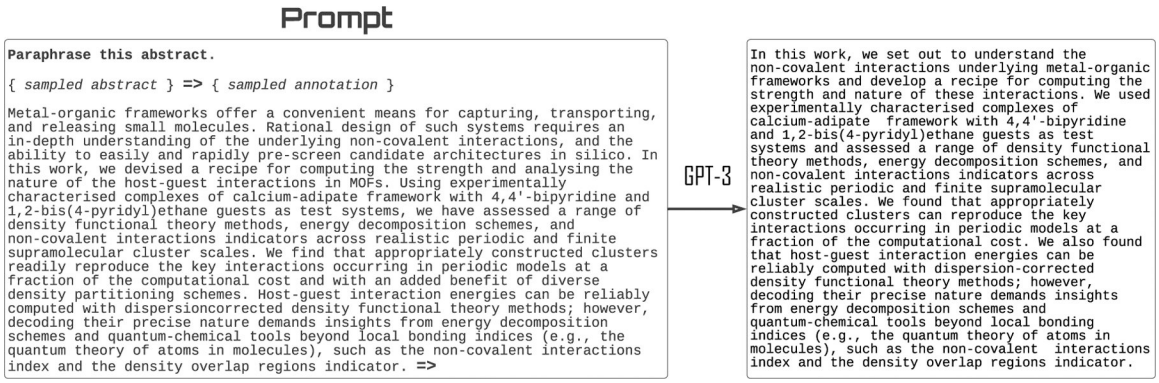

Figure 8:

[6] `google/pegasus-pubmed` on the HuggingFace Transformers Hub (Wolf et al., 2020).

An example prompt and paraphrase output from GPT-3. Words are changed but the meaning is preserved.

Table 10:

Statistics for each candidate generation method. Rel. stands for Relevance and is measured by BERTScore F1 overlap with the reference. Faith. stands for faithfulness and is measured by the FactScore (as defined in §4.2). Extract. stands for the extractive density (level of copy-and-paste) as defined by Grusky et al. (2018). The first 6 rows (Mask-And-Fill and Swaps) construct negative examples for faithfulness calibration. The next two rows form the positive candidate set for faithfulness. The last two (diverse beam) form candidates for relevance calibration.

	Candidate Method	Clinical			Chemical			Biomedical		
		Rel.	Faith.	Extract.	Rel.	Faith.	Extract.	Rel.	Faith.	Extract.
Faith. Contrast	Mask-And-Fill (*Low*)	0.98	0.52	1.55	0.99	0.75	3.24	0.97	0.73	4.92
	Mask-And-Fill (*High*)	0.97	0.52	1.44	0.97	0.73	2.90	0.95	0.71	4.05
	Swap Intrinsic (*Low*)	0.94	0.52	1.64	0.97	0.70	2.92	0.98	0.71	4.70
	Swap Intrinsic (*High*)	0.90	0.52	1.82	0.95	0.65	2.62	0.97	0.67	4.13
	Swap Extrinsic (*Low*)	0.94	0.52	1.64	0.97	0.70	2.92	0.98	0.68	4.44
	Swap Extrinsic (*High*)	0.90	0.52	1.82	0.95	0.65	2.62	0.97	0.64	3.79
	Paraphrase	0.90	0.52	1.26	0.94	0.77	3.06	0.92	0.73	4.00
	Reference	1.00	0.52	1.96	1.00	0.76	3.54	1.00	0.74	5.78
Rel.	Diverse Beam (PRIMERA)	0.84	0.53	2.65	0.87	0.85	9.66	0.86	0.86	12.90
Rank	Diverse Beam (LongT5)	0.83	0.52	2.06	0.86	0.83	7.46	0.85	0.82	8.39

E: Candidate Set Analysis (Ctd.)

The idea behind generating candidates with different methods and parameters is twofold: (1) to better understand which candidate generation methods work best on our task of interest: long-form scientific summarization, and (2) to end up with a diverse candidate pool, which allows us to effectively control for certain characteristics when selecting final subsets for calibration experiments.

In Table 10, we show statistics (relevance, faithfulness, and extractive density) for each candidate generation method across the three datasets.

Analysis.

As noted in Adams et al. (2022), the references for the clinical dataset are very abstractive (1.96 density) and unfaithful (0.52 FactScore), as compared to the chemical (3.54 / 0.76) and biomedical (5.78 / 0.74) data. The former is affected by missing clinical notes while the

latter references are abstracts, which *should* be mostly entailed by the claims made in the main paper. Interestingly, the reference is deemed less faithful than the model generations (0.52 vs 0.53/0.52, 0.76 vs 0.85/0.83, and 0.74 vs 0.86/0.82 for diverse beam search clinical, chemical, and biomedical). This likely has to do with the fact that the fine-tuned models (PRIMERA and LongT5) perform substantially more copy-and-pasting from the source input as the references (1.96 vs 2.65/2.06, 3.54 vs 9.66/7.46, and 5.78 vs 12.90/8.39, respectively).

The most unfaithful corruption method is Swap. When looking at (High) across Intrinsic and Extrinsic, its FactScores are 0.52/0.52, 0.65/0.65, and 0.67/0.64 versus 0.52, 0.73, 0.71 for Mask-And-Fill (High), respectively. This likely has to do with an in-domain LM (SciFive) making reasonably well-informed replacements for noun phrases, whereas entity swapping is indiscriminate and random. The (High) parameter settings for Mask-And-Fill and Swap create less faithful candidates vis-a-vis the (Low) settings (0.75/0.70/0.70 versus 0.73/0.65/0.65 for High and Low on Chemical, for example), as expected. Replacing more text from the references introduces more factual errors.

The PRIMERA model produces more extractive summaries with diverse beam search (2.65/9.66/12.90 vs 2.06/7.46/8.39), which are scored as more relevant and faithful than LongT5.

F: Training Details

F. 1 FT Training Details

We fine-tune (`FT`) two state of the art long-document summarization models for 50,000 steps: PRIMERA (Xiao et al., 2022) (the backbone is a Longformer Encoder-Decoder (LED) (Beltagy et al., 2020) model) and LongT5 (Guo et al., 2022) (which incorporates the sparse attention of ETC (Ainslie et al., 2020) into PEGASUS (Zhang et al., 2020a)) on a single A100 40GB GPU with half precision (FP16)[7]) and a batch a size of 1 (with 16 gradient accumulation steps). We set the maximum learning rate to $3e-5$ with 2,000 warmup steps, followed by a linear decay. We set a maximum input sequence length of 4,096 for both models[8], and a maximum target length of 512 for training / inference for abstract generation (Chemical and Biomedical) and 256 for clinical summarization. Each fine-tuning (`FT`) experiment took ~ 3.5 days.

We select the better performing model (PRIMERA) as the model to be used for `CT` (See Table 5). As discussed in §4.1, LongT5 is still used to supply ten diverse summaries to the candidate pool for relevance calibration.

[7]Only for PRIMERA since LongT5 does not support half precision weights.
[8]Even though LongT5 has a maximum input sequence length of 16,384, we chose 4,096 to match PRIMERA and because of GPU memory constraints.

Table 11:

Hyper-Parameters for calibration fine-tuning.

	Parameter	Clin	Chem	Bio
Relevance Ranking	λ_{MLE}	0.1	0.1	0.1
	λ_{CA}	1.0	1.0	1.0
	λ_{margin}	.001	.001	.001
	α (length penalty)	1.0	2.0	2.0
	τ (scale)	.01	0.1	0.1
Faithful	λ_{MLE}	1.0	1.0	1.0
Contrast	λ_{CA}	1.0	10.0	1.0

F. 2 CT Training Details

We run calibration-tuning (CT) for a maximum of 10,000 steps and select the checkpoint which maximizes either Rel_{Agg} or $Faith_{Agg}$ (depending on the experiment) on the validation set in 1,000 step intervals.

We use the same hyper-parameters as *FT* except the batch size is reduced from 16 to 8. Hyper-parameters related to the CT loss function were tuned separately for each dataset and quality metric (the values selected are shown in Table 11). Each CT experiment took ~ 1 day to train.

As in Guo et al. (2022), summaries are generated greedily, which we found to be significantly faster and even slightly outperformed beam search[9].

G: Identifying Possible Correlates

We examine five basic aspects of calibration sets that *should* have some impact on downstream performance. For each aspect, we provide intuition and some related work to guess the nature of the impact, which we investigate empirically in §6.

G. 1 Overall Quality

Definition.

For the purposes of this analysis, for relevance-rank sets, we define quality as the average Rel_{Agg} score of the candidates.

Relevance Hypothesis.

For relevance, high-quality sets might be preferable to lower-quality sets for two reasons: (1) the model before calibration (pre-CT) has already been fine-tuned (post-FT) on the same

[9]This also means that a length penalty cannot be applied during decoding, which puts more emphasis on the significant role of length tuning during relevance calibration.

training data used for CT, so it likely already assigns a high-probability mass to summaries which are close to the reference. Candidate summaries which deviate too much should already have a low probability of being generated and thus not provide much of a learning signal. In some ways, this hypothesis is supported by Zhao et al. (2022) who find that using a model's top beams produces consistently better results than diverse beam search or sampling-based methods (e.g., nucleus sampling (Holtzman et al., 2020)). There is an inherent tension between the calibration objective, which involves exploration, and the MLE, which assigns all probability mass to a single point.

G. 2 Margin

Overall quality covers average metric values, while margin covers within-set variation in quality.

Definition.

For relevance rank-based sets, we define the margin as the average relevance score between all adjacent pairs of ranked candidates: $Avg\left(Rel_{Agg}\left(\hat{S}_i, S\right) - Rel_{Agg}\left(\hat{S_{i+1}}, S\right)\right), i \in \left|\hat{S}\right| - 1$. For faithfulness, we define it as the delta in average $Faith_{Agg}$ scores for summaries in the positive and negative contrast sets, respectively.

Relevance Hypothesis.

As noisy proxies for human judgments (Peyrard and Gurevych, 2018), subtle differences in relevance metrics (e.g, ROUGE and BERTScore) might not be meaningful. As such, we hypothesize that, all else equal, sets with larger metric gaps will provide a clearer training signal during calibration and superior downstream results.

Faithfulness Hypothesis.

Trivially, one would want positive candidates which are fully faithful. For negatives, it is less clear. The emphasis in the literature has been on producing negative summaries which mimic model errors (Goyal and Durrett, 2021). Yet, less is discussed about the intensity of errors. Lee et al. (2022) explore corruption intensity in the context of training a faithfulness evaluator, and the results suggest a concave relationship. Too few edits and the contrast sets are not easily separable, yet too dramatic, and the contrastive loss is ineffectual. We suspect a similar result for calibrating with a contrastive objective.

G. 3 Lexical Diversity

The previous calibration set characteristic (Margin) covered metric-based comparisons. In this section, we perform comparisons solely at the word-level.

Definition.

We define lexical diversity as the average pairwise self-BLEU score (Zhu et al., 2018; Alihosseini et al., 2019) between all candidates in a relevance ranking set and separately, for positives and negative subsets in a faithfulness contrast set.

Relevance Hypothesis.

All else equal, high lexical diversity should improve the robustness of calibration models as it somewhat dampens some of the noise from single-reference MLE training[10].

Faithfulness Hypothesis.

High lexical diversity within positive and negative sets should make the contrastive classifier less reliant on lexical overlap and focus more on the gap in faithfulness between positive and negatives. Lexical diversity likely means more coverage of error types, which has been shown to be beneficial for contrastive fine-tuning (Cao and Wang, 2021b; Adams et al., 2022).

G. 4 Likelihood

This section covers a model-specific aspect of calibration sets: the likelihood of the candidate summaries under the model post-`FT` and pre-`CT`.

Definition.

For each candidate summary, we compute its length-normalized conditional log likelihood: $\frac{1}{L}\sum_{l=1}^{L} logp(s_l \mid D, S_{<i}; \theta_{FT})$, where θ_{FT} denotes the model parameters after fine-tuning.

Relevance Hypothesis.

One would suspect that likely calibration sets are preferable to unlikely since there is little need to calibrate a model to candidate summaries it was never likely to generate.

Faithfulness Hypothesis.

In a similar vein, it makes sense that contrastive learning for faithulness will be most powerful when the model is most surprised. That is, the negatives are more likely to be generated than the positive. This relates to work by Goyal and Durrett (2021), who argue that negative sets should mimic observed errors.

G. 5 Spurious Correlates

Automatic evaluation metrics have a tendency to reward outputs with characteristics which are spuriously correlated to quality (Durmus et al., 2022).

Definitions.

While many possibilities exist (Durmus et al., 2022), for relevance, we focus on summary length, as defined by number of tokens. For faithfulness, we focus on extractiveness, which we measure with density (Grusky et al., 2018): the average squared length of extractive fragments. It approximates the level of copy-and-paste.

[10]We use the word *somewhat* because we acknowledge that relevance metrics measure overlap to a single reference, so introducing diverse calibration candidates does not necessarily encourage, or reward, more diverse outputs. Access to multiple references, or calibrating against human judgments, would better mitigate the single reference exposure bias problem.

Relevance Hypothesis.

Sun et al. (2019) discover that ROUGE rewards longer summaries while humans prefer concise summaries. We hypothesize that exposing models to longer outputs during calibration will lead to longer summaries, which will have higher relevance scores. By controlling for calibration set length, we can better understand whether or not some of the gains from calibration simply come from length tuning[11].

Faithfulness Hypothesis.

Ladhak et al. (2022) note that faithfulness metrics tend to prefer summaries with high levels of extraction, all else equal. Yet, Zhang et al. (2022) demonstrate that highly extractive does not always mean more faithful, so it is important to get a sense of how much faithfulness calibration is driven by more copy-and-paste.

H: Analysis of Spurious Correlates

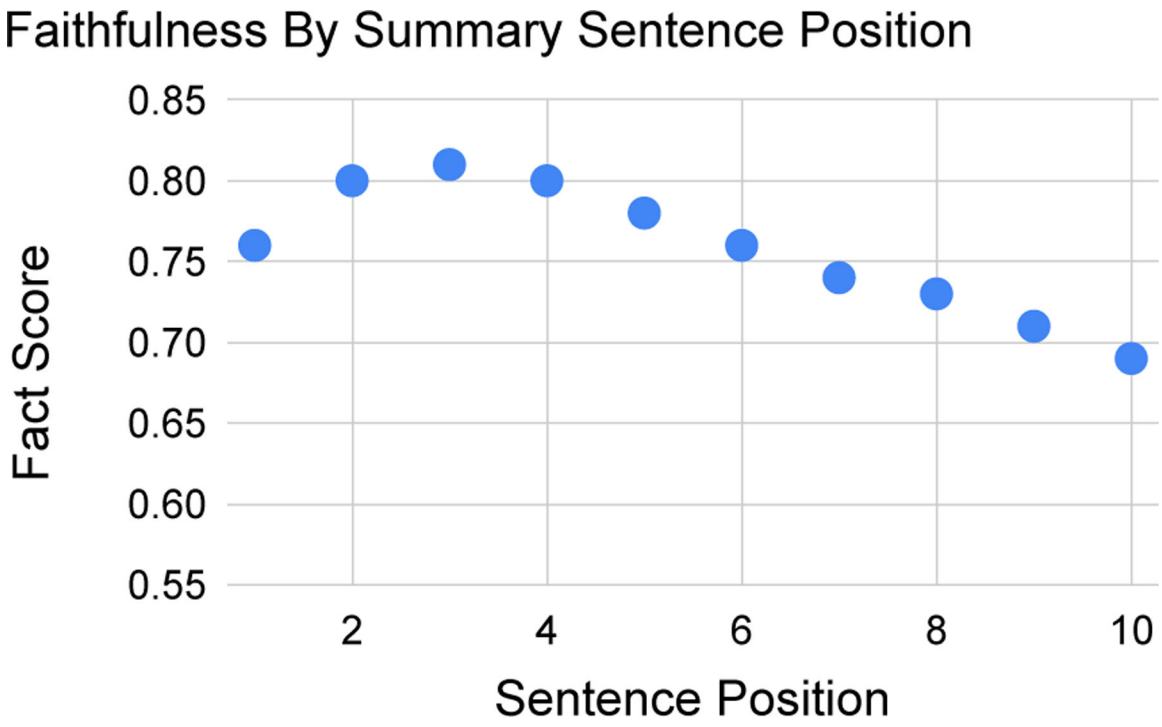

Figure 9:
Sentence-level faithfulness, as defined by FactScore in §4.2, declines as summaries grow longer.

[11]While length can be influenced during beam search with minimum/maximum length restrictions and length penalties, these measures do not expose a model to long summaries.

H.1 The Outsized Role of Length

tl;dr.

The length of summaries is correlated with performance for both relevance and faithful calibration yet for different reasons. For relevance, it can help reduce discrepancies in token-level length between references and generated summaries after fine-tuning. For faithfulness, generated summaries become less faithful as average length increases.

Evidence.

For relevance calibration, the Table 6 section on `Spurious Correlates` shows that selecting the longest summaries is preferable to the shortest for Clinical calibration (.255 versus .181) yet the reverse is true for Biomedical (.017 for max length and .033 for min length). We can trace this to a gap, after fine-tuning, in model summary length and reference lengths. On average, PRIMERA summaries after `FT` are 119 tokens for clinical and 230 for biomedical. Yet, the clinical references are, on average, 416 tokens and only 205 for biomedical. The optimal length strategy seems contingent on the direction of the length gap.

For faithfulness, we simply compute the correlation between $Faith_{Agg}$ and summary tokens: −.75. For faithfulness, we can confirm the presence of text degeneration (Holtzman et al., 2020) as a function of output length by measuring the average *FactScore* at each sentence position in the summary. Figure 9 confirms this story, despite an initial slight increase up to the third sentence.

Implications.

For relevance, as argued by Sun et al. (2019), work should acknowledges changes in the lengths of summaries and address its role in impacting relevance metrics. Long-form summarization research which involves identifying and solving subproblems (Krishna et al., 2021) might mitigate some of the length-based degeneration.

Table 12:

Correlation of faithfulness metrics to extractive density of summaries. Correlations computed on the test set of the PRIMERA models after fine-tuning.

Metric	Clinical	Chemical	Biomedical
FactScore	.78	.42	.42
BARTScore	.35	.16	.45
BERTScore-Src	.52	.47	.60

H.2 Faithful or More Extractive?

tl;dr.

One would expect that training on contrast sets with a large difference in extractiveness (extractive positives, abstractive negatives) would lead to higher downstream faithfulness. Yet, we find the opposite to be true, which we connect to §6.5.

Evidence.

Ladhak et al. (2022) note a spurious correlation between the extractiveness of summaries and faithfulness metrics, which holds true for the metrics which make up $Faith_{Agg}$ (as shown in Table 12). One would expect that reinforcing this correlation via contrastive learning (by targeting extractive positives and abstractive negatives) would lead to improved faithfulness metrics. Yet, this does not appear to be the case. Table 7 (`Spurious` selection type) shows that on average, controlling for a large extractiveness gap does not improve faithfulness (.131 versus an overall average improvement of .133). If anything, it leads to increased relevance (.017 versus −.067). While not definitive, a possible driver for this relationship relates to the analysis in §6.5, for which we show that a low likelihood gap between positives and negatives is preferable (an adversarial setup). Since extractive summaries are more likely to be generated than abstractive ones (see Extractive density for Diverse Beam search in Table 10), extractive negatives might be preferable to abstractive ones.

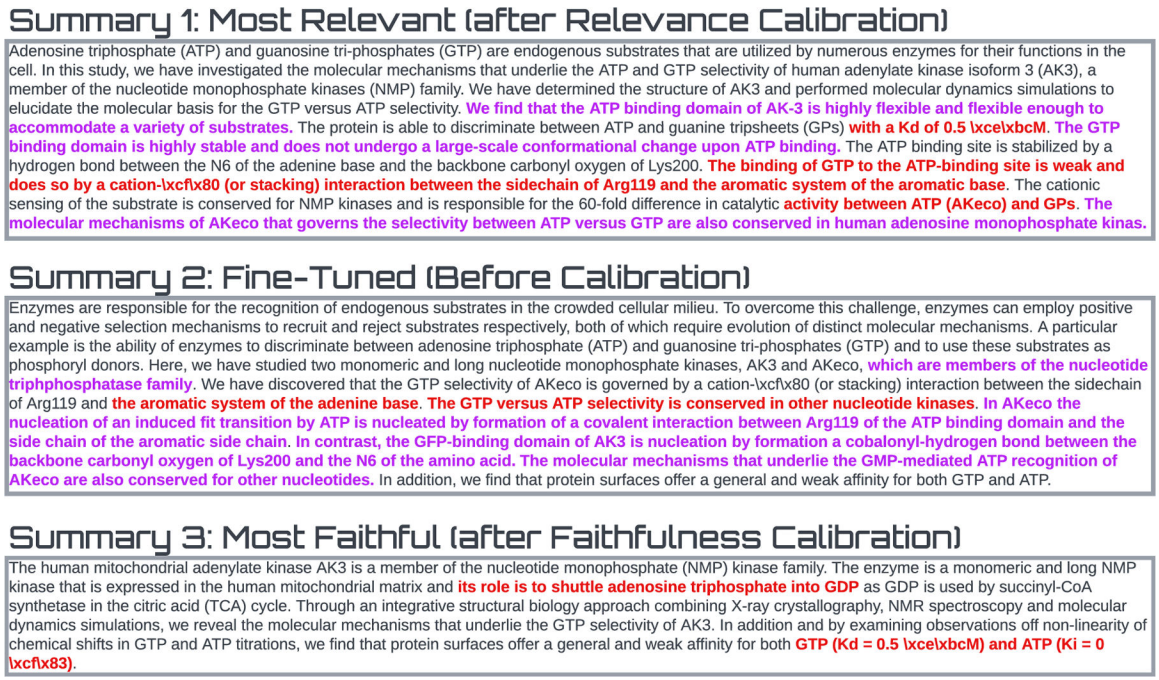

Figure 10:

Three abstracts generated from model checkpoints after Relevance Calibration (Summary 1), Fine-Tuning (PRIMERA FT checkpoint, Summary 2), and after Faithfulness Calibration (Summary 3). Red Text has been annotated as being part of an intrinsic error while Purple Text is extrinsic. The annotator rated Summary 1 as the most relevant and Summary 3 the least relevant.

Implications.

Given the extractiveness of long-form scientific summaries, more research should focus on subtle faithfulness errors, i.e., those which are less correlated to extractiveness. Zhang et al. (2022) provide a helpful typology of errors in fully extractive systems, which can provide a blueprint for the design of more extractive synthetic errors.

I: Human Evaluation Details

To better understand whether or not our calibration models are driving meaningful changes in quality, we conduct a human evaluation on the chemistry dataset. Specifically, we randomly select 50 papers from the test set and collect model generated abstracts from the `FT` checkpoint as well as most relevant (`Random` strategy) and most faithful (`Hard` strategy) `CT` weights. After randomly shuffling the order of abstracts, we ask each annotator (four authors of this paper with PhDs in chemistry-related fields) to first read the main paper and then, separately for each paper, highlight spans of abstracts containing errors (intrinsic or extrinsic), before ranking the summaries by Relevance (Fabbri et al., 2021b). We defined relevance as in SummEval: *how well does the summary captures the key points of the paper? Consider whether all and only the important aspects are contained in the summary.*. We collect fine-grained faithfulness annotations, rather than summary-level, due to the length of the summaries and prior work on interannotator agreement scores of fine-grained errors (Pagnoni et al., 2021; Goyal and Durrett, 2021).

I. 1 Error Analysis

In this section, we analyze the errors from an example in the human annotation set. The abstracts are shown in Figure 10.

Abstract 1 takes the general form of an abstract, providing a reasonable motivation for the work then listing a number of key findings. It makes a number of errors in stating the key findings, however. First, the model seems to have had difficulty with abbreviations and measured values, misreporting a binding constant and confusing GTP and ATP on several occasions. Finally, the model includes several statements not supported in the text. Abstract 2 contains superior prose to Abstract 1, better enumerating the motivation for the work and providing a cleaner concluding statement. It suffers from similar shortcomings, however, confusing GTP and ATP on several occasions and making a number of unsupported claims. In some cases, the unsupported claims appear lifted whole-cloth from another publication. In total, we judge the errors in Abstract 2 to be more misleading than those made in Abstract 1 and thus find Abstract 1 to be more relevant. Abstract 3 is substantially shorter than either Abstract 1 or Abstract 2, minimizing the absolute number of errors it contains. Like the others, it has difficulty with both abbreviations and measured values, making errors due to both. Overall, Abstract 3 is not terribly written; however, its terseness leaves a highly limited description of the paper's contributions. For this reason, it is less relevant than either Abstract 1 or Abstract 2.

J: Connecting Metric Margins to Diversity

Larger margin gaps are related to diversity as lexically similar summaries will have similar metric values. In fact, we can examine the `Diversity` section of Table 6 and note that average Rel_{Agg} score across datasets is higher when lexical diversity is maximized (.114) than when it is minimized (.082). Yet, this trend only holds for the Chemical dataset. To get a more complete sense, we examine the impact of set diversity across runs and note a slightly more reassuring trend: a pearson correlation coefficient of .21, .51, and .1

for clinical, chemical, and biomedical. Interestingly, chemical has the strongest positive relationship between diversity and downstream relevance across runs, yet is negative when directly controlling for diversity.

References

Adams Griffin, Alsentzer Emily, Ketenci Mert, Zucker Jason, and Elhadad Noémie. 2021. What's in a summary? laying the groundwork for advances in hospital-course summarization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4794–4811, Online. Association for Computational Linguistics.

Adams Griffin, Shing Han-Chin, Sun Qing, Winestock Christopher, McKeown Kathleen, and Elhadad Noémie. 2022. Learning to revise references for faithful summarization. In Findings of the Association for Computational Linguistics: EMNLP 2022, pages 4009–4027, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ainslie Joshua, Ontanon Santiago, Alberti Chris, Cvicek Vaclav, Fisher Zachary, Pham Philip, Ravula Anirudh, Sanghai Sumit, Wang Qifan, and Yang Li. 2020. ETC: Encoding long and structured inputs in transformers. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 268–284, Online. Association for Computational Linguistics.

Alihosseini Danial, Montahaei Ehsan, and Soleymani Baghshah Mahdieh. 2019. Jointly measuring diversity and quality in text generation models. In Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation, pages 90–98, Minneapolis, Minnesota. Association for Computational Linguistics.

Beltagy Iz, Peters Matthew E, and Cohan Arman. 2020. Longformer: The long-document transformer. ArXiv preprint, abs/2004.05150.

Bodenreider Olivier. 2004. The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research, 32(suppl_1):D267–D270. [PubMed: 14681409]

Brown Tom B., Mann Benjamin, Ryder Nick, Subbiah Melanie, Kaplan Jared, Dhariwal Prafulla, Neelakantan Arvind, Shyam Pranav, Sastry Girish, Askell Amanda, Agarwal Sandhini, Herbert-Voss Ariel, Krueger Gretchen, Henighan Tom, Child Rewon, Ramesh Aditya, Ziegler Daniel M., Wu Jeffrey, Winter Clemens, Hesse Christopher, Chen Mark, Sigler Eric, Litwin Mateusz, Gray Scott, Chess Benjamin, Clark Jack, Berner Christopher, McCandlish Sam, Radford Alec, Sutskever Ilya, and Amodei Dario. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, virtual.

Cao Meng, Dong Yue, Wu Jiapeng, and Kit Cheung Jackie Chi. 2020. Factual error correction for abstractive summarization models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6251–6258, Online. Association for Computational Linguistics.

Cao Shuyang and Wang Lu. 2021a. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Cao Shuyang and Wang Lu. 2021b. CLIFF: Contrastive learning for improving faithfulness and factuality in abstractive summarization. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6633–6649, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Cao Ziqiang, Wei Furu, Li Wenjie, and Li Sujian. 2018. Faithful to the original: Fact aware neural abstractive summarization. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2–7, 2018, pages 4784–4791. AAAI Press.

Chen Sihao, Zhang Fan, Sone Kazoo, and Roth Dan. 2021. Improving faithfulness in abstractive summarization with contrast candidate generation and selection. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5935–5941, Online. Association for Computational Linguistics.

Chintagunta Bharath, Katariya Namit, Amatriain Xavier, and Kannan Anitha. 2021. Medically aware GPT-3 as a data generator for medical dialogue summarization. In Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations, pages 66–76, Online. Association for Computational Linguistics.

Cohan Arman, Dernoncourt Franck, Kim Doo Soon, Bui Trung, Kim Seokhwan, Chang Walter, and Goharian Nazli. 2018. A discourse-aware attention model for abstractive summarization of long documents. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Deng Mingkai, Tan Bowen, Liu Zhengzhong, Xing Eric, and Hu Zhiting. 2021a. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7580–7605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Deng Mingkai, Tan Bowen, Liu Zhengzhong, Xing Eric, and Hu Zhiting. 2021b. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7580–7605, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

DeYoung Jay, Beltagy Iz, van Zuylen Madeleine, Kuehl Bailey, and Wang Lucy. 2021. MS^2: Multi-document summarization of medical studies. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 7494–7513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Durmus Esin, Ladhak Faisal, and Hashimoto Tatsunori. 2022. Spurious correlations in reference-free evaluation of text generation. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1443–1454, Dublin, Ireland. Association for Computational Linguistics.

Fabbri Alexander, Han Simeng, Li Haoyuan, Li Haoran, Ghazvininejad Marjan, Joty Shafiq, Radev Dragomir, and Mehdad Yashar. 2021a. Improving zero and few-shot abstractive summarization with intermediate fine-tuning and data augmentation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 704–717, Online. Association for Computational Linguistics.

Fabbri Alexander R., Kryściński Wojciech, McCann Bryan, Xiong Caiming, Socher Richard, and Radev Dragomir. 2021b. SummEval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics, 9:391–409.

Filippova Katja. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 864–870, Online. Association for Computational Linguistics.

Gehrmann Sebastian, Clark Elizabeth, and Sellam Thibault. 2022. Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text. ArXiv preprint, abs/2202.06935.

Goyal Tanya and Durrett Greg. 2020. Neural syntactic preordering for controlled paraphrase generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 238–252, Online. Association for Computational Linguistics.

Goyal Tanya and Durrett Greg. 2021. Annotating and modeling fine-grained factuality in summarization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1449–1462, Online. Association for Computational Linguistics.

Grusky Max, Naaman Mor, and Artzi Yoav. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,

Volume 1 (Long Papers), pages 708–719, New Orleans, Louisiana. Association for Computational Linguistics.

Gunel Beliz, Du Jingfei, Conneau Alexis, and Stoyanov Veselin. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021. OpenReview.net.

Guo Mandy, Ainslie Joshua, Uthus David, Ontanon Santiago, Ni Jianmo, Sung Yun-Hsuan, and Yang Yinfei. 2022. LongT5: Efficient text-to-text transformer for long sequences. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 724–736, Seattle, United States. Association for Computational Linguistics.

Gupta Vivek, Bharti Prerna, Nokhiz Pegah, and Karnick Harish. 2021. SumPubMed: Summarization dataset of PubMed scientific articles. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 292–303, Online. Association for Computational Linguistics.

Holtzman Ari, Buys Jan, Du Li, Forbes Maxwell, and Choi Yejin. 2020. The curious case of neural text degeneration. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020. OpenReview.net.

Johnson Alistair EW, Pollard Tom J, Shen Lu, Li-Wei H Lehman, Feng Mengling, Ghassemi Mohammad, Moody Benjamin, Szolovits Peter, Anthony Leo, and Mark Roger G. 2016. Mimic-iii, a freely accessible critical care database. Scientific data, 3(1):1–9.

Kang Daniel and Hashimoto Tatsunori B.. 2020. Improved natural language generation via loss truncation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 718–731, Online. Association for Computational Linguistics.

Khosla Prannay, Teterwak Piotr, Wang Chen, Sarna Aaron, Tian Yonglong, Isola Phillip, Maschinot Aaron, Liu Ce, and Krishnan Dilip. 2020. Supervised contrastive learning. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, virtual.

Kim Donghyeon, Lee Jinhyuk, So Chan Ho, Jeon Hwisang, Jeong Minbyul, Choi Yonghwa, Yoon Wonjin, Sung Mujeen, and Kang Jaewoo. 2019. A neural named entity recognition and multi-type normalization tool for biomedical text mining. IEEE Access, 7:73729–73740.

Krishna Kundan, Khosla Sopan, Bigham Jeffrey, and Lipton Zachary C.. 2021. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 4958–4972, Online. Association for Computational Linguistics.

Kryscinski Wojciech, McCann Bryan, Xiong Caiming, and Socher Richard. 2020. Evaluating the factual consistency of abstractive text summarization. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9332–9346, Online. Association for Computational Linguistics.

Ladhak Faisal, Durmus Esin, He He, Cardie Claire, and McKeown Kathleen. 2022. Faithful or extractive? on mitigating the faithfulness-abstractiveness tradeoff in abstractive summarization. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1410–1421, Dublin, Ireland. Association for Computational Linguistics.

Lebanoff Logan, Muchovej John, Dernoncourt Franck, Kim Doo Soon, Kim Seokhwan, Chang Walter, and Liu Fei. 2019. Analyzing sentence fusion in abstractive summarization. In Proceedings of the 2nd Workshop on New Frontiers in Summarization, pages 104–110, Hong Kong, China. Association for Computational Linguistics.

Lee Hwanhee, Yoo Kang Min, Park Joonsuk, Lee Hwaran, and Jung Kyomin. 2022. Masked summarization to generate factually inconsistent summaries for improved factual consistency checking. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 1019–1030, Seattle, United States. Association for Computational Linguistics.

Lin Chin-Yew. 2004. ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Liu Yixin and Liu Pengfei. 2021a. SimCLS: A simple framework for contrastive learning of abstractive summarization. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 1065–1072, Online. Association for Computational Linguistics.

Liu Yixin and Liu Pengfei. 2021b. SimCLS: A simple framework for contrastive learning of abstractive summarization. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 1065–1072, Online. Association for Computational Linguistics.

Liu Yixin, Liu Pengfei, Radev Dragomir, and Neubig Graham. 2022. BRIO: Bringing order to abstractive summarization. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2890–2903, Dublin, Ireland. Association for Computational Linguistics.

Lopez Patrice. 2009. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In International conference on theory and practice of digital libraries, pages 473–474. Springer.

Lu Yao, Dong Yue, and Charlin Laurent. 2020. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8068–8074, Online. Association for Computational Linguistics.

Marcus Mitchell P., Santorini Beatrice, and Marcinkiewicz Mary Ann. 1993. Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics, 19(2):313–330.

Maynez Joshua, Narayan Shashi, Bohnet Bernd, and McDonald Ryan. 2020. On faithfulness and factuality in abstractive summarization. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1906–1919, Online. Association for Computational Linguistics.

McKeown Kathleen. 2020. Rewriting the past: Assessing the field through the lens of language generation.

Nan Feng, Nallapati Ramesh, Wang Zhiguo, dos Santos Cicero Nogueira, Zhu Henghui, Zhang Dejiao, McKeown Kathleen, and Xiang Bing. 2021a. Entity-level factual consistency of abstractive text summarization. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 2727–2733, Online. Association for Computational Linguistics.

Nan Feng, dos Santos Cicero Nogueira, Zhu Henghui, Ng Patrick, McKeown Kathleen, Nallapati Ramesh, Zhang Dejiao, Wang Zhiguo, Arnold Andrew O., and Xiang Bing. 2021b. Improving factual consistency of abstractive summarization via question answering. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6881–6894, Online. Association for Computational Linguistics.

Narayan Shashi, Zhao Yao, Maynez Joshua, Simões Gonçalo, Nikolaev Vitaly, and McDonald Ryan. 2021. Planning with learned entity prompts for abstractive summarization. Transactions of the Association for Computational Linguistics, 9:1475–1492.

Pagnoni Artidoro, Balachandran Vidhisha, and Tsvetkov Yulia. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4812–4829, Online. Association for Computational Linguistics.

Peyrard Maxime and Gurevych Iryna. 2018. Objective function learning to match human judgements for optimization-based summarization. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 654–660, New Orleans, Louisiana. Association for Computational Linguistics.

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

Pyysalo Sampo, Ohta Tomoko, Rak Rafal, Rowley Andrew, Chun Hong-Woo, Jung Sung-Jae, Choi Sung-Pil, Tsujii Jun'ichi, and Ananiadou Sophia. 2015. Overview of the cancer genetics and pathway curation tasks of bionlp shared task 2013. BMC bioinformatics, 16(10):1–19. [PubMed: 25591917]

Qi Peng, Zhang Yuhao, Zhang Yuhui, Bolton Jason, and Manning Christopher D.. 2020. Stanza: A python natural language processing toolkit for many human languages. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 101–108, Online. Association for Computational Linguistics.

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

Sun Simeng, Shapira Ori, Dagan Ido, and Nenkova Ani. 2019. How to compare summarizers without target length? pitfalls, solutions and re-examination of the neural summarization literature. In Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation, pages 21–29, Minneapolis, Minnesota. Association for Computational Linguistics.

Tang Xiangru, Nair Arjun, Wang Borui, Wang Bingyao, Desai Jai, Wade Aaron, Li Haoran, Celikyilmaz Asli, Mehdad Yashar, and Radev Dragomir. 2022. CONFIT: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5657–5668, Seattle, United States. Association for Computational Linguistics.

Uzuner Özlem, South Brett R, Shen Shuying, and DuVall Scott L. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association, 18(5):552–556. [PubMed: 21685143]

Vijayakumar Ashwin K, Cogswell Michael, Selvaraju Ramprasath R, Sun Qing, Lee Stefan, Crandall David, and Batra Dhruv. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. ArXiv preprint, abs/1610.02424.

Wadden David, Lin Shanchuan, Lo Kyle, Wang Lucy Lu, van Zuylen Madeleine, Cohan Arman, and Hajishirzi Hannaneh. 2020. Fact or fiction: Verifying scientific claims. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 7534–7550, Online. Association for Computational Linguistics.

Wadden David, Lo Kyle, Wang Lucy, Cohan Arman, Beltagy Iz, and Hajishirzi Hannaneh. 2022. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 61–76, Seattle, United States. Association for Computational Linguistics.

Wan David and Bansal Mohit. 2022. FactPEGASUS: Factuality-aware pre-training and fine-tuning for abstractive summarization. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1010–1028, Seattle, United States. Association for Computational Linguistics.

Wieting John and Gimpel Kevin. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Wolf Thomas, Debut Lysandre, Sanh Victor, Chaumond Julien, Delangue Clement, Moi Anthony, Cistac Pierric, Rault Tim, Louf Remi, Funtowicz Morgan, Davison Joe, Shleifer Sam, von Platen Patrick, Ma Clara, Jernite Yacine, Plu Julien, Xu Canwen, Scao Teven Le, Gugger Sylvain, Drame Mariama, Lhoest Quentin, and Rush Alexander. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Wu Hanlu, Ma Tengfei, Wu Lingfei, Manyumwa Tariro, and Ji Shouling. 2020. Unsupervised reference-free summary quality evaluation via contrastive learning. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 3612–3621, Online. Association for Computational Linguistics.

Xiao Wen, Beltagy Iz, Carenini Giuseppe, and Cohan Arman. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

Yuan Weizhe, Neubig Graham, and Liu Pengfei. 2021. Bartscore: Evaluating generated text as text generation. In Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6–14, 2021, virtual, pages 27263–27277.

Zhang Jingqing, Zhao Yao, Saleh Mohammad, and Liu Peter J.. 2020a. PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13–18 July 2020, Virtual Event, volume 119 of Proceedings of Machine Learning Research, pages 11328–11339. PMLR.

Zhang Shiyue, Wan David, and Bansal Mohit. 2022. Extractive is not faithful: An investigation of broad unfaithfulness problems in extractive summarization.

Zhang Tianyi, Kishore Varsha, Wu Felix, Weinberger Kilian Q., and Artzi Yoav. 2020b. Bertscore: Evaluating text generation with BERT. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26–30, 2020. OpenReview.net.

Zhang Yuan, Baldridge Jason, and He Luheng. 2019. PAWS: Paraphrase adversaries from word scrambling. In 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), pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhang Yuhao, Zhang Yuhui, Qi Peng, Manning Christopher D, and Langlotz Curtis P. 2021. Biomedical and clinical English model packages for the Stanza Python NLP library. Journal of the American Medical Informatics Association.

Zhao Yao, Khalman Misha, Joshi Rishabh, Narayan Shashi, Saleh Mohammad, and Liu Peter J. 2022. Calibrating sequence likelihood improves conditional language generation. ArXiv preprint, abs/2210.00045.

Zhou Chunting, Neubig Graham, Gu Jiatao, Diab Mona, Guzmán Francisco, Zettlemoyer Luke, and Ghazvininejad Marjan. 2021. Detecting hallucinated content in conditional neural sequence generation. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1393–1404, Online. Association for Computational Linguistics.

Zhou Jianing and Bhat Suma. 2021. Paraphrase generation: A survey of the state of the art. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5075–5086, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhu Chenguang, Hinthorn William, Xu Ruochen, Zeng Qingkai, Zeng Michael, Huang Xuedong, and Jiang Meng. 2021. Enhancing factual consistency of abstractive summarization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 718–733, Online. Association for Computational Linguistics.

Zhu Yaoming, Lu Sidi, Zheng Lei, Guo Jiaxian, Zhang Weinan, Wang Jun, and Yu Yong. 2018. Texygen: A benchmarking platform for text generation models. In The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08–12, 2018, pages 1097–1100. ACM.

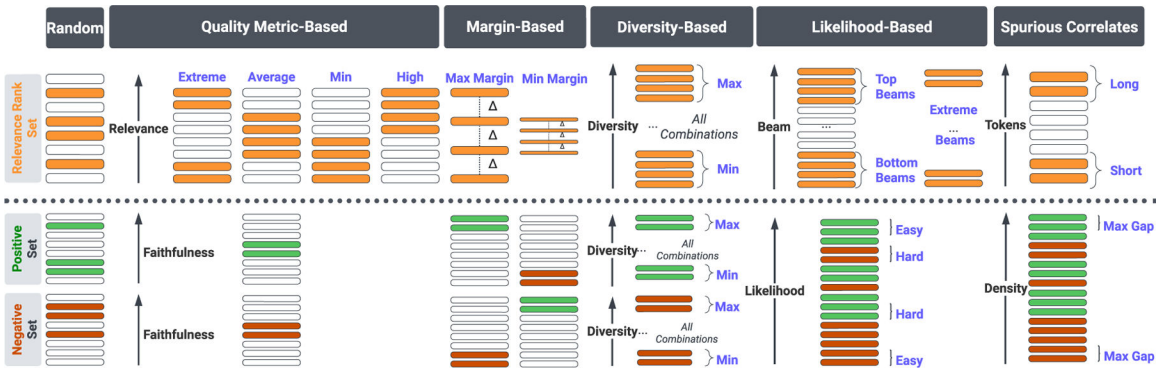

Figure 1:

Strategies for selecting rank sets of size 4 from larger candidate pools for relevance calibration (top half). The bottom half shows similar strategies to form binary contrast sets (2 positive, 2 negative) for faithfulness. Each strategy for the top half of the Figure occupies a row in Table 6, while the bottom corresponds to rows in Table 7.

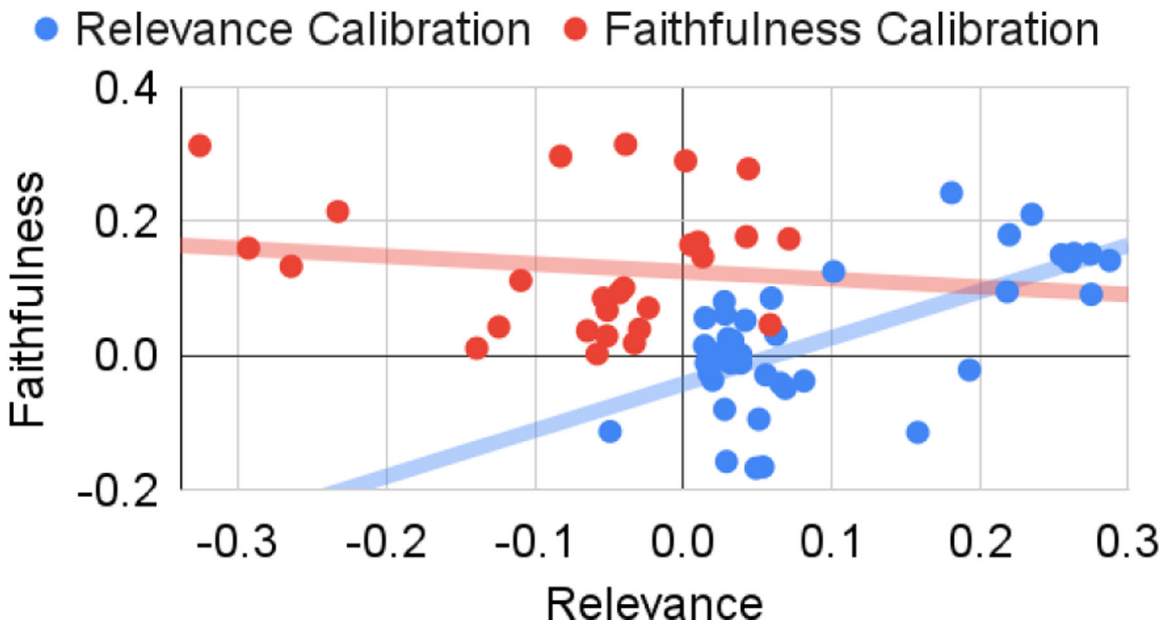

Figure 2:
A plot of average summary relevance and faithfulness across experiments, which are designed to either improve relevance (blue) or faithfulness (red).

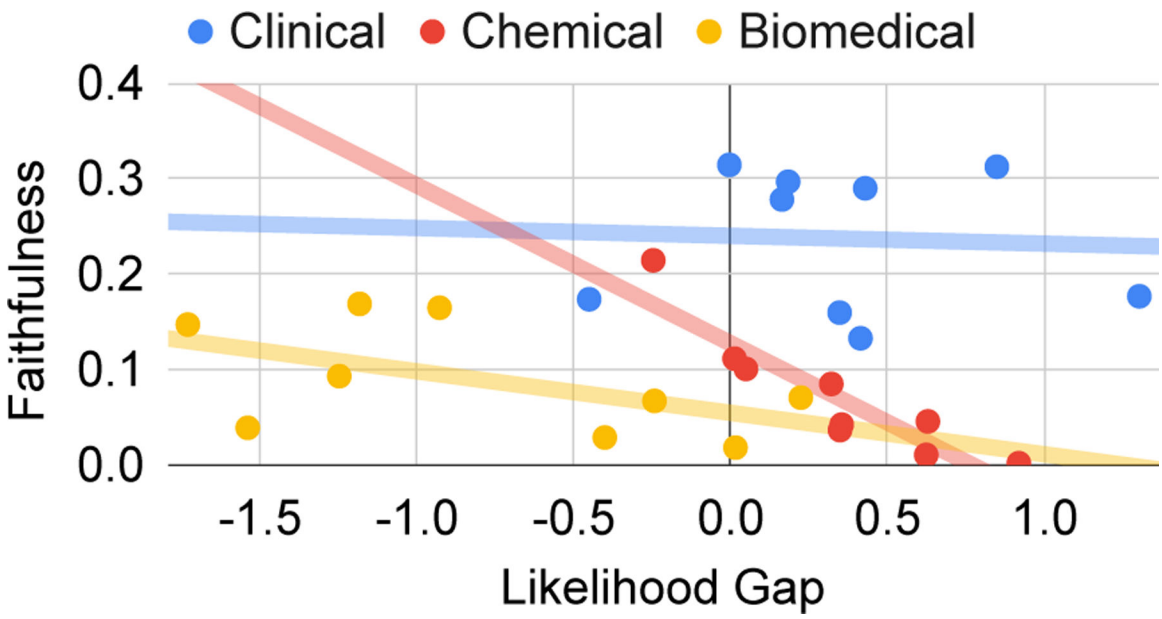

Figure 3:
A plot comparing the average likelihood gap (difference in log likelihood of generating a positive candidate over a negative pre-calibration) against the average summary faithfulness after calibration.

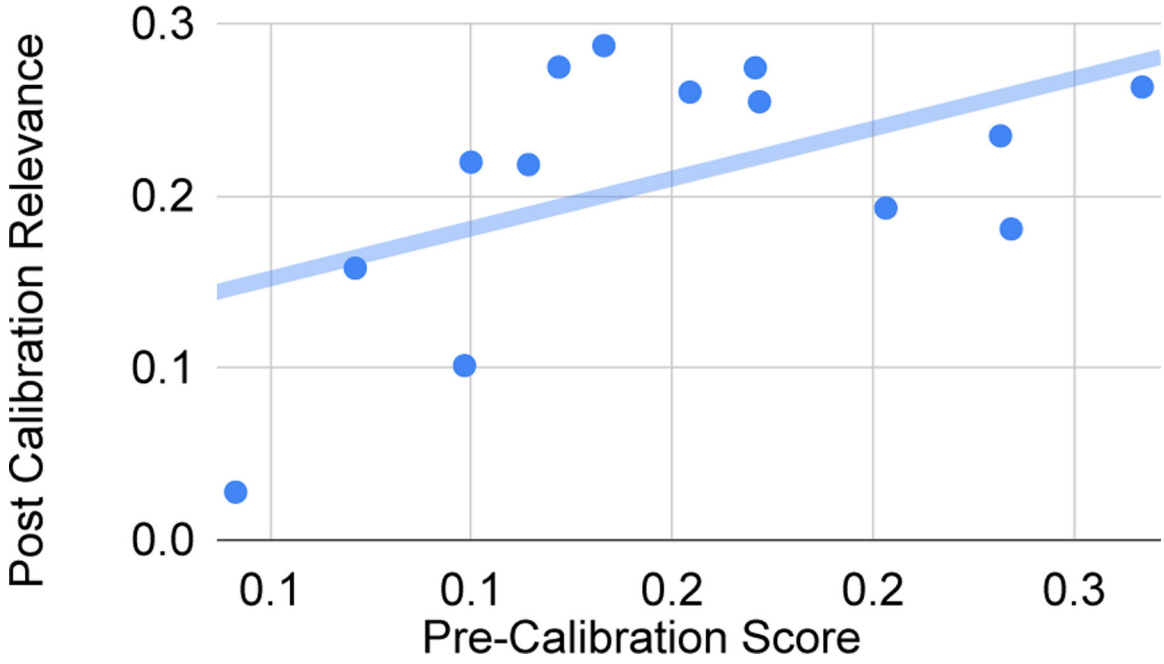

Figure 4:
A plot which shows average pre-calibration score for each clinical relevance experiment on the x-axis, and the post-calibration relevance on the y-axis.

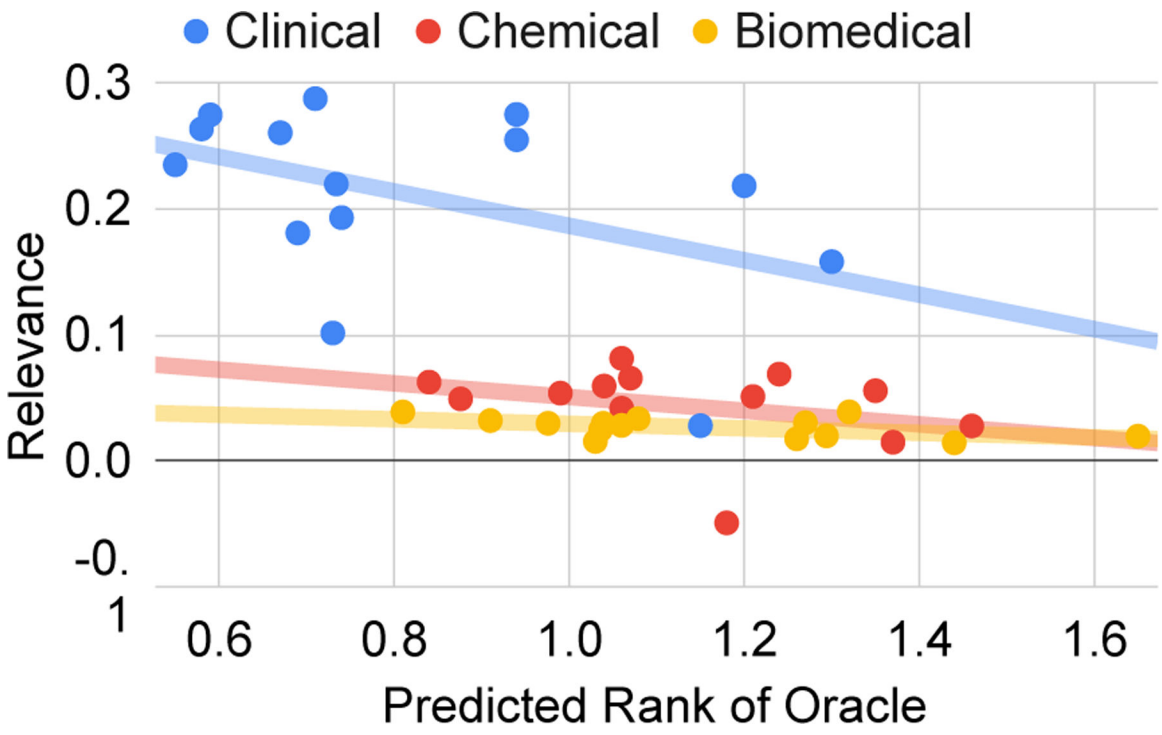

Figure 5:
A plot showing the impact of calibration performance on downstream performance (relevance). An average rank of 0 reveals a model which always identifies the most relevant summary. The worst score is 3.

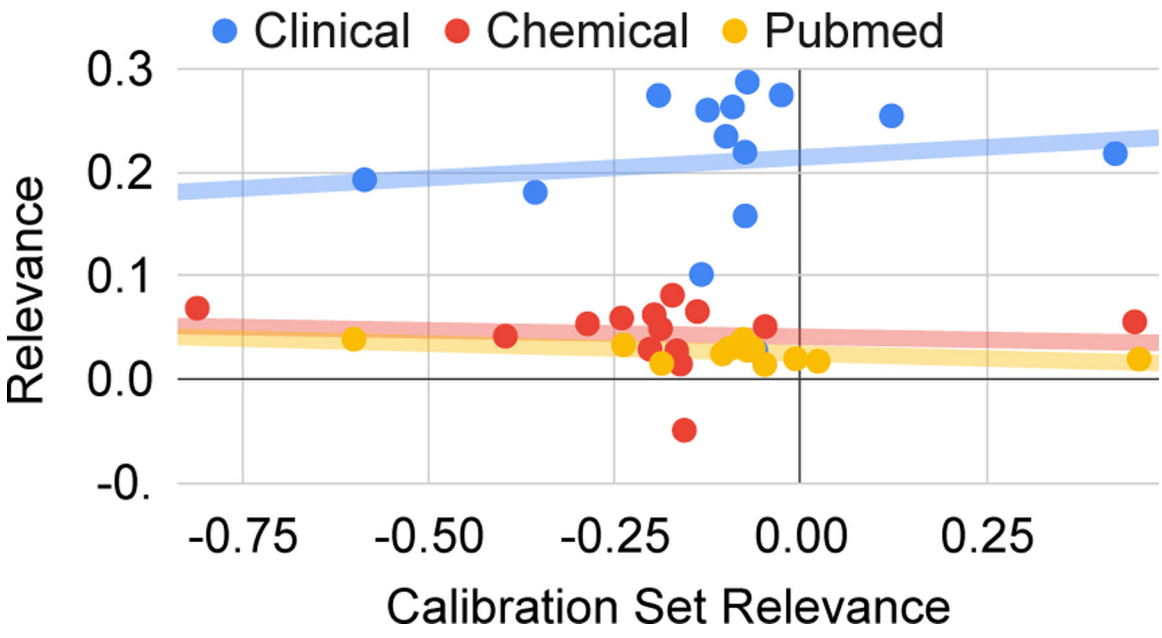

Figure 6:
The impact of the average relevance of calibration candidates on downstream summary relevance.

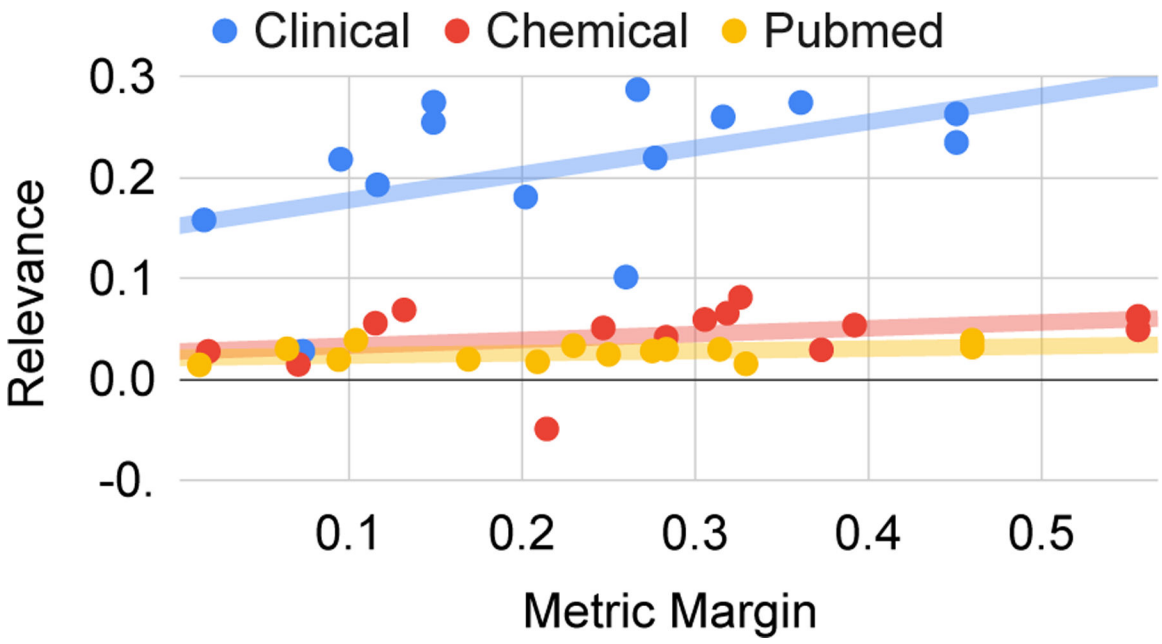

Figure 7:

The impact of the average metric-wise margin (Rel_{Agg}) between calibration candidates on the relevance of downstream model outputs after calibration.

Table 1:

Statistics for long-form scientific summarization datasets. The biomedical dataset is from Cohan et al. (2018), the recipe to recreate the clinical from Adams et al. (2022), and the chemical from this work.

Statistic	Clinical	Chemical	Bio.
Train Size	41,705	115,956	119,924
Validation Size	940	1,000	6,633
Test Size	1,861	2,000	6,658
Source Tokens	8,175	5,364	3,092
Reference Tokens	416	216	205
Extractive Coverage	0.66	0.90	0.88
Extractive Density	1.97	3.53	5.87

Table 2:

Journals accessed for Chemical papers.

Source	# Articles
Beilstein	1,829
Chem Cell	546
ChemRxiv	12,231
Chemistry Open	398
Nature Communications Chemistry	572
PubMed Author Manuscript	57,680
PubMed Open Access	29,540
Royal Society of Chemistry (RSC)	9,334
Scientific Reports - Nature	6,826

Table 3:

Methods to create negative and positive candidates in support of relevance and faithfulness calibration, respectively. For each candidate generation method, we include whether it is used as a positive or negative example (both in the case of relevance ranking), what inputs it requires (the source document and/or the reference (ref.)), as well as the external components needed and, finally, the specific models used for the experiments in this paper.

	Method	−	+	Source	Ref.	External Components	Models Used
Relevance Calibration	Diverse Beam	✓	✓	✓		Summarization Model	PRIMERA
	Diverse Beam	✓	✓	✓		Summarization Model	LongT5
Faithful Calibration	Mask-And-Fill	✓			✓	Constituency Parser, PLM	Stanza, SciFive
	Entity Swap	✓			✓	Entity, Number Extractors	BERN2, Quantulum
	Paraphrase		✓		✓	Paraphrase Generator	GPT-3 + Curated Prompt
	Reference		✓		✓	N/A	N/A

Proc Conf Assoc Comput Linguist Meet. Author manuscript; available in PMC 2024 April 30.

Table 4:

# of candidates pooled for each training instance. m is % of noun phrases masked, s % of entities swapped, and t the softmax temperature for GPT-3.

Method	Hyper-Param	Number
Mask-And-Fill (*Low*)	$m = 0.25$	10
Mask-And-Fill (*High*)	$m = 0.75$	10
Swap Intrinsic (*Low*)	$s = 0.5$	10
Swap Intrinsic (*High*)	$s = 1.0$	10
Swap Extrinsic (*Low*)	$s = 0.5$	10
Swap Extrinsic (*High*)	$s = 1.0$	10
Paraphrase	$t = 0.7$	5
Reference	N/A	1
Total For Faithfulness		**66**
Diverse Beam (PRIMERA)	$p = 1$	10
Diverse Beam (LongT5)	$p = 1$	10
Total For Relevance		**20**

Table 5:

Benchmarking PRIMERA and LongT5 models after initial fine-tuning (FT) for relevance and faithfulness. R1, R2, and BS-Ref stand for Rouge-1/2 F1 and BERTScore F1 vis-a-vis reference, respectively. Fact., Bart., and BS-Src stand for FactScore, BARTScore, and BERTScore F1 vis-a-vis the source. Metrics defined in §4.1 and 4.2.

	Model	Clinical			Chemical			Biomedical		
		R1	**R2**	**BS-Ref**	**R1**	**R2**	**BS-Ref**	**R1**	**R2**	**BS-Ref**
Relevance Metrics	**PRIMERA**	**25.15**	**9.39**	**83.81**	**45.47**	**16.31**	**86.24**	**48.01**	**20.83**	**86.25**
	LongT5	24.22	8.57	83.15	42.51	14.46	85.74	44.32	17.91	85.02
		Fact.	**Bart.**	**BS-Src**	**Fact.**	**Bart.**	**BS-Src**	**Fact.**	**Bart.**	**BS-Src**
Faithful Metrics	**PRIMERA**	53.29	−2.92	**83.33**	**85.96**	**−6.29**	**88.89**	**86.91**	**−3.77**	**88.54**
	LongT5	**53.71**	**−2.88**	82.84	83.25	−6.36	88.70	83.62	−3.89	88.31

Table 6:

PRIMERA models calibrated to improve relevance. Calibration candidates are pooled from fine-tuned PRIMERA and LongT5 models. *REL* stands for Rel_{Agg} (from §4.1). *FAITH* stands for $Faith_{Agg}$ (from §4.2).

Selection Type	Selection Strategy	Clinical		Chemical		Biomedical		Dataset Avg.	
		REL	*FAITH*	*REL*	*FAITH*	*REL*	*FAITH*	*REL*	*FAITH*
Random	–	.220	.180	.081	−.038	.028	.061	.110	.068
Quality Based	*Extreme*	.263	.152	.049	−.168	.039	.002	.117	−.005
	Average	.028	−.080	.015	.056	.030	.025	.024	.000
	Min	.193	−.022	.069	−.049	.039	−.012	.100	−.027
	High	.218	.095	.056	−.029	.019	.004	.098	.023
Margin Based	*Max*	.235	.210	.062	.031	.032	−.011	.110	.077
	Min	.158	−.115	.028	.080	.014	.015	.067	−.007
Diversity Based	*Max*	.274	.151	.054	−.166	.015	−.011	.114	−.009
	Min	.275	.091	−.049	−.114	.020	−.037	.082	−.020
Likelihood Based	*Extreme Beam*	.260	.140	.029	−.158	.030	−.008	.106	−.009
	Top Beam	.287	.142	.066	−.042	.030	−.008	.128	.031
	Bottom Beam	.101	.125	.059	.085	.025	−.002	.062	.069
Spurious Correlates	*Max Length*	.255	.150	.051	−.095	.017	−.027	.108	.009
	Min Length	.181	.243	.042	.052	.033	.022	.085	.106
Avg. Across Strategies		.211	.104	.044	−.040	.027	.001	.094	.022

Proc Conf Assoc Comput Linguist Meet. Author manuscript; available in PMC 2024 April 30.

Table 7:

PRIMERA models calibrated to improve faithfulness. Contrast sets for calibration are formed from the generation methods in §4.2. *REL* stands for Rel_{Agg} (from §4.1). *FAITH* stands for $Faith_{Agg}$ (from §4.2).

Selection Type	Selection Strategy	Clinical		Chemical		Biomedical		Dataset Avg.	
		REL	*FAITH*	*REL*	*FAITH*	*REL*	*FAITH*	*REL*	*FAITH*
Random	–	−.264	.133	−.054	.085	.005	.165	−.104	.128
Quality	*Average*	−.293	.160	−.065	.037	.010	.169	−.116	.122
Margin Based	*Max*	−.326	.313	−.139	.011	−.033	.018	−.166	.114
	Min	−.083	.297	−.109	.112	−.030	.039	−.074	.149
Diversity Based	*Max*	.002	.290	−.124	.043	−.052	.029	−.058	.121
	Min	−.039	.315	−.040	.101	−.043	.093	−.041	.170
Likelihood Based	*Easy*	.043	.177	−.058	.002	−.024	.071	−.013	.083
	Hard	.071	.174	−.233	.215	.013	.147	−.050	.179
Spurious	*Max Extract. Gap*	.044	.278	.058	.046	−.051	.067	.017	.131
Avg. Across Strategies		−.094	.237	−.085	.072	−.023	.089	−.067	.133

Proc Conf Assoc Comput Linguist Meet. Author manuscript; available in PMC 2024 April 30.

Table 8:

Results from human evaluation on 75 total system summaries from the chemistry test set. Int. and Ext. stand for average intrinsic and extrinsic errors identified. Rel. Rank stands for the average rank assigned by annotators (1–3) with 1 being viewed as the most relevant.

System	Int.	Ext.	Total	Rel. Rank
FT	2.00	1.24	3.24	2.04
Most Relevant	1.67	1.43	3.10	1.85
Most Faithful	1.10	0.81	1.90	2.12

Table 9:

Relevance `CT` by forming sets which maximize rank correlation between Rel. and Faith. scores improves mean combined (comb.) Rel. and Faith. scores vis-a-vis an average of the strategies shown in Table 6.

	Average Strategy			Max Correlation		
	Rel	**Faith**	**Comb**	**Rel**	**Faith**	**Comb**
Clin.	.211	.104	.158	.090	.325	.208
Chem.	.044	−.040	.007	.040	.104	.158
Bio.	.027	.001	.014	.018	.025	.022
Avg.	.094	.022	**.059**	.049	.151	**.100**