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

Griffin Adams[♠,♣*] griffin.adams@columbia.edu
Bichlien Nguyen[♢] bnguyen@microsoft.com
Jake Smith[♢] jakesmith@microsoft.com
Yingce Xia[♢] yingce.xia@microsoft.com
Shufang Xie[♢] shufxi@microsoft.com
Anna Ostropolets[♣] anna.ostropolets@columbia.edu
Budhaditya Deb[♢] budha.deb@microsoft.com
Kali Frost[♢] kali.frost@microsoft.com
Yuan-Jyue Chen[♢] yuanjc@microsoft.com
Tristan Naumann[♢] tristan@microsoft.com
Noémie Elhadad[♠,♣] noemie.elhadad@columbia.edu

Microsoft Research[♢] **Columbia University: Computer Science**[♠], **Biomedical Informatics**[♣]

Abstract

Summarization models are typically trained to maximize the likelihood of a single reference (MLE). As a consequence, during inference, the probabilities assigned to model generations are often poorly calibrated to quality metrics. To address this, after an initial MLE step, recent work has added a calibration step, which exposes a model 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 the 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 aim to uncover the underlying characteristics of effective candidate sets for both relevance and faithfulness calibration. 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 ranked candidates should be maximized and surprise minimized.

1 Introduction

Traditionally, summarization models have been trained to maximize the likelihood of gold-standard references. This training paradigm introduces an exposure bias in that, 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 more prone to produce summaries with irrelevant or repetitive content (Zhao et al., 2022), or mis-represent the claims in the source text (Maynez et al., 2020).

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 outputs to the source (Nan et al., 2021b; Cao and Wang, 2021a).

Yet, much of this work has addressed the *how* question–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 little study of the interaction between the two. 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

[*] Work started during internship with Microsoft Research.

negatives and positives, diversity, and the model likelihood of generating each candidate in the set.

We run experiments which vary solely by the training data selected for candidate sets. For each experiment, we extract a wide range of relevant statistics (i.e., diversity, length) on the candidate sets and plot the relationship between these set statistics and downstream performance. To guide future research, we analyze the plots to provide insights onto, and rationale for, optimal set construction for faithfulness and relevance calibration.

Additionally, a large portion of research has focused on summarization 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, i.e., faithfulness, are non-trivial to define, capture, and categorize. Long-form scientific summarization is appealing for our calibration experiments for two reasons. First, memory is constrained when dealing with long inputs and outputs. From a practical perspective, even with training tricks, such as gradient accumulation and half precision, only a small handful of candidates per example (4 in our experiments[1]) can be used for training. This makes the selection step more important than for shorter tasks (i.e., news summarization) where larger calibration sets are possible. Secondly, scientific summarization is highly knowledge-intensive, which presents unique challenges related to faithfulness. The majority of error analysis to date has focused on summaries of news articles (Fabbri et al., 2021b; Pagnoni et al., 2021).

The primary contributions of this work are to: (1) benchmark calibration models on three scientific long-form datasets; (2) conduct extensive experiments to better understand the underlying characteristics and dynamics of effective calibration sets.

2 Related Work

Typically, when summarization models are calibrated to quality metrics, it 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.

More broadly, contrastive learning has proven effective across modalities and tasks, from image pre-training (Chen et al., 2020) to language understanding (Gunel et al., 2020) and unsupervised cross-modal alignment (Radford et al., 2021).

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 explicitly instructed to calibrate decoder likelihoods to ROUGE rankings. While best results are seen for continued fine-tuning with a calibration objective, a second model can be used either to re-rank (Liu et al., 2021; Ravaut et al., 2022a) or fuse (Ravaut et al., 2022b) candidate summaries.

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 large candiate 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). 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 ad-

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

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

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.

mission and discharge regarding a single patient. It is a highly noisy, naturally occurring dataset, which makes it interesting from a faithfulness perspective.

Source	# Articles
Nature Communications Chemistry	582
Royal Society of Chemistry (RSC)	9,759
Beilstein	2,461
Chem Cell	603
Chemistry Open	474
Pubmed Open Access	35,366
Pubmed Author Manuscript	59,239

Table 2: Journals accessed for Chemical papers.

Chemical. We introduce a dataset with a pure chemistry focus by compiling a list of chemistry academic journals (in Table 2) that publish articles under the Open-Access license. For each journal, we downloaded PDFs from an exposed API, and when not available, scraped each website with the Selenium Chrome WebDriver to download full-text articles of the Open-Access portion of the journal. Each PDF was processed with Grobid (Lopez, 2009) via a locally installed 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 Grobid-extracted abstract is treated as the reference.

While other science-related summarization datasets exist (Lu et al., 2020; Gupta et al., 2021; DeYoung et al., 2021), none focus on chemistry.

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 *further* fine-tune (FFT) on a combined objective, which adds a calibration loss (CA) to the MLE loss:

$$\begin{aligned} \mathcal{L}_{FT} &= \mathcal{L}_{MLE} \\ \mathcal{L}_{FFT} &= \lambda_{MLE} * \mathcal{L}_{MLE} + \lambda_{CA} * \mathcal{L}_{CA} \end{aligned} \quad (1)$$

$\lambda_{MLE}, \lambda_{CA}$ are coefficients controlling the relative weight of each objective on $\mathcal{L}_{FFT}$. For $\mathcal{L}_{FFT}$, $\mathcal{L}_{MLE}$ acts as a regularizer to prevent degradation in the model's generation capability, as in Edunov et al. (2018); 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).

Note that we use the term calibration to refer interchangeably to relevance-focused rank-based sets (the conventional meaning), and binary contrast sets for faithfulness. They both involve calibrating a model's predictions to a specific quality metric.

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-FFT 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:

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

f can be a similarity metric between S and D (Liu and Liu, 2021b; Zhong et al., 2020) or, more directly, the log likelihood (Liu et al., 2022). For this paper, we implement the latter, where f represents for the length normalized log likelihood:

$$\frac{\tau \sum_{l=1}^{L} logp(s_l | D, S_{<l}; \theta)}{L^{\alpha}}$$

τ is a scalar and α a length penalty parameter. Calibration models are highly sensitive to α given the spurious correlation of widely used relevance metrics, e.g., ROUGE, to length (Sun et al., 2019).

Rank Metric. To define the gold-standard ordering, we compute three relevance metrics, normalize them, and aggregate them to form a single score.

Rel_{Agg} is a combination of ROUGE 1/2 scores (Lin, 2004) and **BertTScore-Ref** between summary and reference. For BertScore (Zhang et al., 2019a), 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.

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 from each model with diverse beam search (Vijayakumar et al., 2016), as in Liu et al. (2022); Nan et al. (2021b); Ravaut et al. (2022a). We experimented with regular beam search, as in Zhao et al. (2022), but the outputs were very similar. A diversity penalty of 1.0 is used to penalize repeated tokens at the same position across beams groups.

4.2 Faithfulness Calibration

Objective. As in Gunel et al. (2020); 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:

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

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. Similarly to Rel_{Agg}, we compute a $Faith_{Agg}$ score by aggregating normalized metrics. Specifically, we combine **BartScore** (Yuan et al., 2021) and **BertScore-Src** vis-a-vis source. For BartScore, we use a PEGASUS (Zhang et al., 2020) model pretrained on the Pubmed summarization corpus[2] 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: $\frac{1}{|S|}\sum_{i \in |S|} p(s_i|, s_{j<i}, D)$. 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 dataset (Wadden et al., 2020). SciFact is an expert-annotated dataset of 1,409 sentence-level scientific claims. To convert to our task, 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.

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[3] 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 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.

Entity swapping involves replacing entities in the reference with entities from either the source

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

[3] 'NP' using the annotation scheme from the Penn Treebank (Marcinkiewicz, 1994).

text (`intrinsic`) or hallucinated (`extrinsic`). 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. 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.

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[4] 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.

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

Positive Methods. We pool together the **Reference** with **Paraphrased** versions of it. Paraphrasing is typically done with synonym substitution (Zhou and Bhat, 2021), neural models (Goyal and Durrett, 2020) trained on paraphrase corpora (Wieting and Gimpel, 2017; Zhang et al., 2019b), 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. 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). A random sample of one annotation pair[5], 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 summarization).

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

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

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 3: Number of candidates pooled for each training instance. Calibration sets for training are formed by selecting a subset of 4 for relevance rank sets and 2 negatives / 2 positives for faithfulness contrast sets.

4.3 Candidate Set Analysis

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 3, 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 candidates (2 positive and 2 negative), and for relevance, we select 4 out of 20 candidates[6].

In Table 4, 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 abstrac-

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

[6]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).

	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. Rank	Diverse Beam (PRIMERA)	0.84	0.53	2.65	0.87	0.85	9.66	0.86	0.86	12.90
	Diverse Beam (LongT5)	0.83	0.52	2.06	0.86	0.83	7.46	0.85	0.82	8.39

Table 4: 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.

tive (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.

5 Overview

At a high-level, we hypothesize which characteristics of calibration sets *could* have some impact on downstream performance (§6), design strategies for set selection which target different values of these characteristics (§7), run many calibration experiments across datasets and quality metrics (relevance and faithfulness), which vary only by the selection strategy (§8). Finally, we aim to understand the correlates of effective sets by tracking the relationship between our hypothesized set of correlates and downstream model performance (§9).

6 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 §9.

6.1 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-FFT) has already been fine-tuned (post-FT) on the same training data used for FFT, 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., 2019)). There is an inherent tension between the calibration objective, which involves exploration, and the MLE, which assigns all probability mass to a single point.

6.2 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: $Avg(Rel_{Agg}(\hat{S}_i, S) - Rel_{Agg}(\hat{S_{i+1}}, S)), i \in |\hat{\boldsymbol{S}}| - 1$. 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 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.

6.3 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[7].

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).

6.4 Likelihood

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

Definition. For each candidate summary, we compute its length-normalized conditional log likelihood: $\frac{1}{L}\sum_{l=1}^{L} logp(s_l|D, S_{<l}; \theta_{FT})$, 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.

6.5 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.

[7]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.

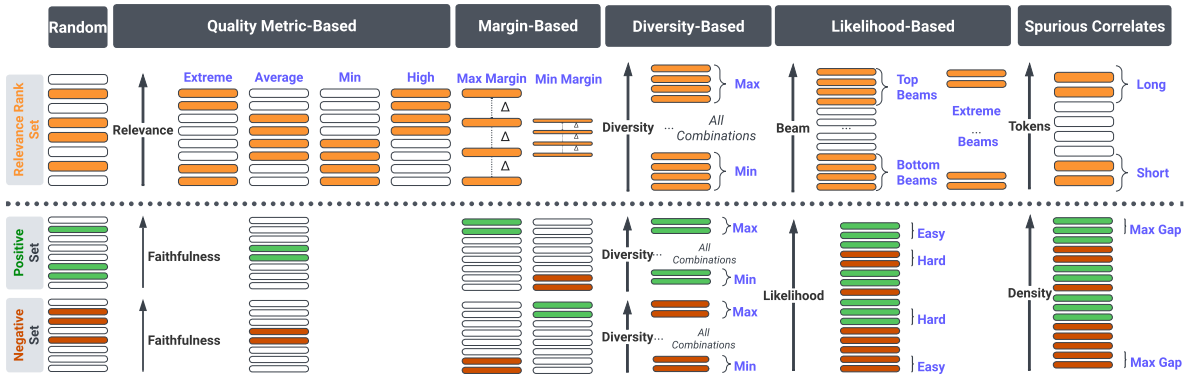

Figure 1: Strategies for selecting rank sets of size 4 from larger candidate pools for relevance calibration (top half). The bottom half shows similar strategies to form binary contrast sets (2 positive, 2 negative) for faithfulness. Each strategy for the top half of the Figure occupies a row in Table 7, while the bottom corresponds to rows in Table 8.

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[8].

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.

7 Selection Strategies.

To test our hypotheses in §6, we design a set of simple strategies to select calibration sets which systematically target different set characteristics.

Problem Statement. From a large candidate pool, select a target number to be used for training (2 positives and 2 negatives for faithfulness calibration, and 4 for rank-based relevance calibration).

Figure 1 graphically reveals the different strategies implemented. Since these strategies are designed to control for certain set characteristics, they do not represent optimal or recommended strategies, e.g., a minimum metric gap for faithfulness.

[8]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.

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

Quality-Based. For quality-based, we rank all candidates by a quality metric (Rel_{Agg} and $Faith_{Agg}$ for relevance rank-based calibration and faithfulness contrast calibration, respectively). Then, we select candidates at different extremes of these scales, as depicted graphically in Figure 1.

Margin-Based. For relevance ranking, we enumerate all possible subsets of size 4 and compute the average metric margin $Avg(Rel_{Agg}(\hat{S}_i, S) - Rel_{Agg}(\hat{S_{i+1}}, S)), i \in |\hat{\boldsymbol{S}}| - 1$. 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 faithfulness positives and least faithful (`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 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 decoded by diverse beam search across both models. For faithfulness, we select for the `Max Extractive Gap`– the most *extractive* positives and most *abstractive* negatives (as measured by the extractive density).

8 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., 2020)) on a single A100 40GB GPU with half precision (FP16)[9]) 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[10], 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 $\sim$ 3.5 days.

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

We run further fine-tuning (`FFT`) 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 `FFT` loss function were tuned separately for each dataset and quality metric (the values selected are shown in Table 6). Each `FFT` experiment took $\sim$ 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[11].

[9]Only for PRIMERA since LongT5 does not support half precision weights.

[10]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.

[11]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.

[12]We note that these our results from own runs with our own hyper-parameters and training setup. They do not represent results from the PRIMERA and LongT5 papers.

9 Results

9.1 Fine-Tuning

We briefly discuss results after `FT` before going into the post-calibration `FFT` results.

Table 5 shows that PRIMERA outperforms LongT5 across faithfulness and relevance and across datasets[12]. 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, for Chemical and Biomedical, we compute BartScore with a model trained on Pubmed summarization (`google/pegasus-pubmed` from §4.2).

9.2 Further Fine-Tuning (Calibration)

In Tables 7, 8, we show relevance and faithfulness scores (Rel_{Agg} and $Faith_{Agg}$) 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).

It is important to note that, although each strategy targets a specific set characteristic (e.g., length), confounding correlates exist for each strategy. For instance, the top beams, on average, have high 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`.

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

9.3 The Impact of Reference Quality

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

	Model	Clinical			Chemical			Biomedical		
Relevance Metrics		**R1**	**R2**	**BS-Ref**	**R1**	**R2**	**BS-Ref**	**R1**	**R2**	**BS-Ref**
	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
Faithful Metrics		**Fact.**	**Bart.**	**BS-Src**	**Fact.**	**Bart.**	**BS-Src**	**Fact.**	**Bart.**	**BS-Src**
	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 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 are defined in §4.1 and 4.2.

	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 Contrast	λ_{MLE}	1.0	1.0	1.0
	λ_{CA}	1.0	10.0	1.0

Table 6: Hyper-Parameters for calibration fine-tuning.

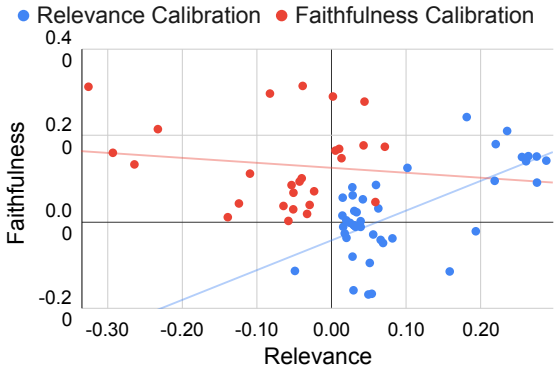

Figure 2: A plot of average summary relevance and faithfulness across experiments, which are designed to either improve relevance (blue) or faithfulness (red).

Evidence. As detailed in Adams et al. (2022), and evidenced by the "Reference" row of Table 4, the clinical references are highly noisy and often unsupported by the source text. The average across strategies for both Tables 7 and 8 reveal the largest relative improvement in Rel_{Agg} and $Faith_{Agg}$ for clinical relevance and faithfulness experiments, 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).

9.4 Relevance and Faithfulness at Odds

tl;dr. Relevance and faithfulness share an inverse relationship when calibrating for faithfulness. Summary length may be a confounding factor, yet research should focus on designing contrast sets which maximize their correlation.

Evidence. In Figure 2, we plot Rel_{Agg} versus $Faith_{Agg}$ across experiments to measure the trade-off 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.

Explanation. To examine why this might be the case, we can look at the relevance scores for positives (reference and paraphrases) versus the negative methods. From Table 4, while references, from a metric perspective, are perfectly relevant, the GPT-3 paraphrases are seen as slightly less rel-

Selection Type	Selection Strategy	Clinical REL	Clinical FAITH	Chemical REL	Chemical FAITH	Biomedical REL	Biomedical FAITH	Dataset Avg. REL	Dataset Avg. 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

Table 7: 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 REL	Clinical FAITH	Chemical REL	Chemical FAITH	Biomedical REL	Biomedical FAITH	Dataset Avg. REL	Dataset Avg. 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

Table 8: 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).

evant (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[13]. The GPT-3 paraphrases are prompted with human paraphrases, which involve more substantial re-writing. We discuss the confounding role of length later.

Implications. Most calibration (or contrastive fine-tuning for faithfulness) research for summarization 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 introducing faithfulness errors into less relevant summaries.

[13]Some papers use unfaithful system outputs (Cao and Wang, 2021a; Nan et al., 2021b) as negatives, which should alleviate the issue of high reference overlap for negative sets.

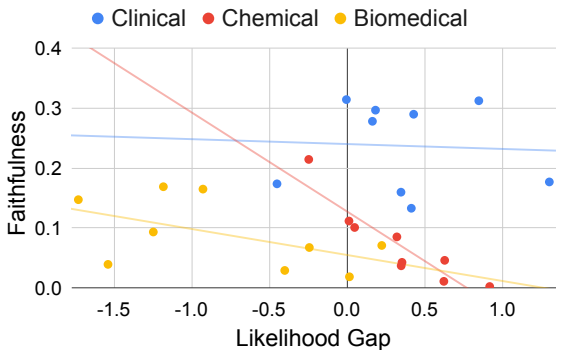

Figure 3: A plot comparing the average likelihood gap (difference in log likelihood of generating a positive candidate over a negative pre-calibration) against the average summary faithfulness after calibration.

9.5 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.

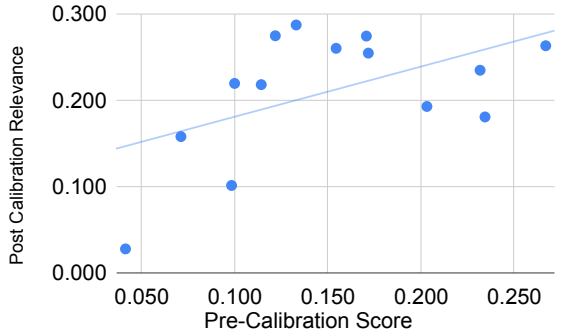

Figure 4: A plot which shows average pre-calibration score for each clinical relevance experiment on the x-axis, and the post-calibration relevance on the y-axis.

Evidence. For relevance, we look at the `Likelihood` section of Table 7 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 8 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 pre-calibrated 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).

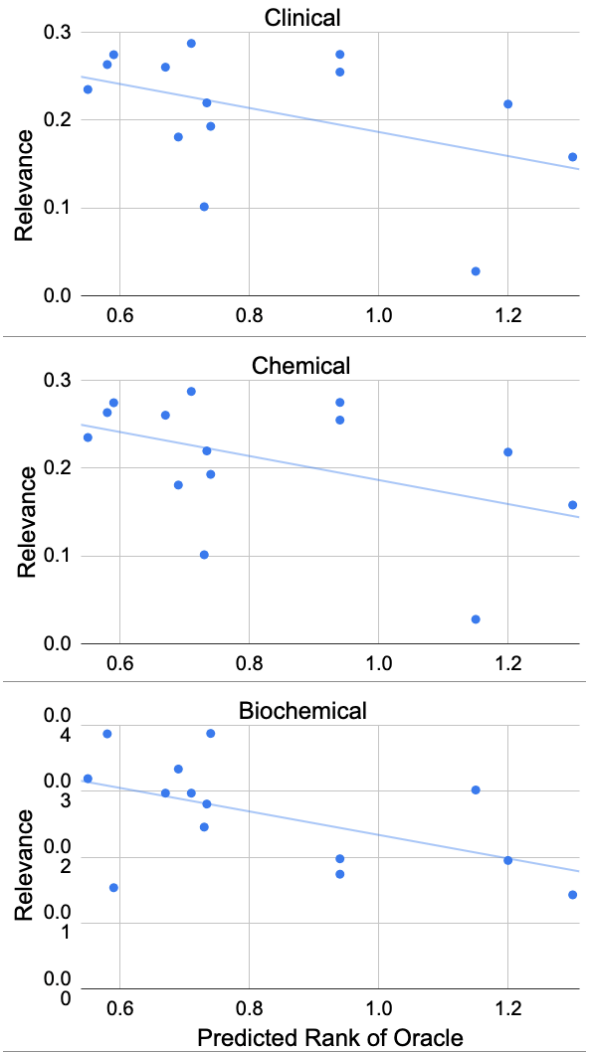

Figure 5: A plot showing the impact of calibration performance on downstream performance (relevance). An average rank of 0 reveals a model which always identifies the most relevant summary. The worst score is 3.

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

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. For faithfulness calibration, adversarial construction of negatives is a promising line of research.

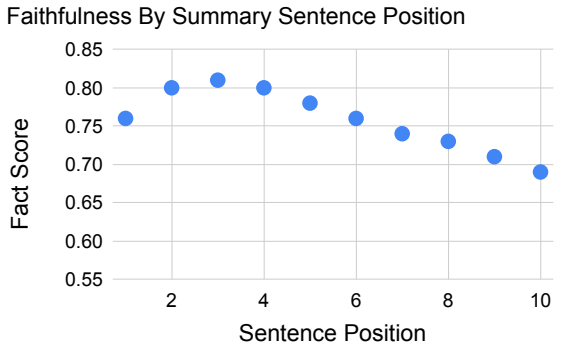

Figure 6: Sentence-level faithfulness, as defined by FactScore in §4.2, declines as summaries grow longer.

9.6 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 7 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., 2019) as a function of output length by measuring the average $FactScore$ at each sentence position in the summary. Figure 6 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.

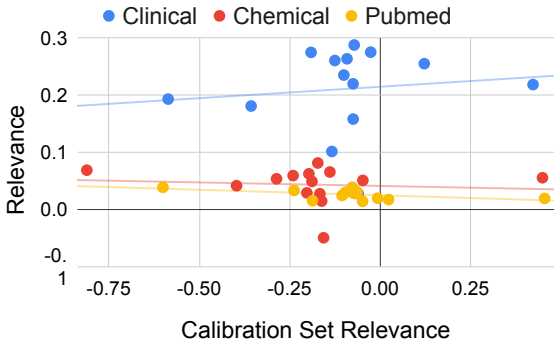

Figure 7: The impact of the average relevance of calibration candidates on downstream summary relevance.

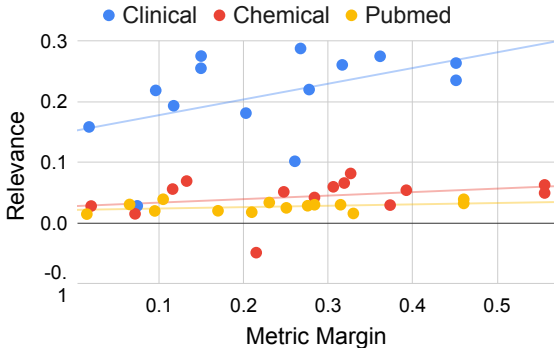

Figure 8: The impact of the average metric-wise margin (Rel_{Agg}) between calibration *candidates* on the relevance of downstream model outputs after calibration.

9.7 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 7 for `Quality Based`, we do not see a clear-cut trend in downstream Rel_{Agg} for dataset averages based on selecting for different relevance values in sets: .117/.024/.100/.098 for `Extreme`, `Average`, `Min`, and `High`, respectively. For `Margin`

Based, which targets the relative values (the metric-gap between adjacent ranked candidates), maximum 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 7) in Figures 7 and 8. Figure 8 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 7 is less consistent, as it is positive for clinical (.12 correlation), yet negative for chemical ($-.10$) and biomedical ($-.51$).

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 7 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.

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.

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

Table 9: Correlation of faithfulness metrics to extractive density of summaries. Correlations computed on the test set of the PRIMERA models after fine-tuning.

9.8 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 §9.5.

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 9). 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 8 (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 §9.5, 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 4), extractive negatives might be preferable to abstractive ones.

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.

10 Limitations and Future Work

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. Due to reasonable computational constraints, we only targeted extreme or average values when designing selection strategies. A human evaluation of post-calibration outputs is needed to confirm whether calibration improves quality or simply a better exploitation of metric tendencies. An error topology can better motivate design choices for negative generation methods (intrinsic/extrinsic, swaps/infilling, etc.).

11 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 designing strategies which systematically target different aspects of calibration

sets. We then analyze trends between these characteristics and the ultimate downstream performance. Our analysis is intended to serve as a guide for subsequent research when designing methods for generating synthetic candidates, as well as to encourage more work to jointly consider relevance and faithfulness calibration given their covariance and the importance of both to real-world systems.

Acknowledgements

Part of this research was supported by the National Library of Medicine (NLM) and National Institute of Allergy and Infectious Diseases (NIAID) of the National Institutes of Health (NIH) under Award Number T15LM007079. The content is solely the responsibility of the authors and does not represent the official views of the NIH.

References

Griffin Adams, Emily Alsentzer, Mert Ketenci, Jason Zucker, and Noémie Elhadad. 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.

Griffin Adams, Han-Chin Shing, Qing Sun, Christopher Winestock, Kathleen McKeown, and Noémie Elhadad. 2022. Learning to revise references for faithful summarization. *ArXiv*, abs/2204.10290.

Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 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.

Danial Alihosseini, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. 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.

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

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 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.

Shuyang Cao and Lu Wang. 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.

Shuyang Cao and Lu Wang. 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.

Sihao Chen, Fan Zhang, Kazoo Sone, and Dan Roth. 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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 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.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 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.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 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.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 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.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Wang. 2021. MSˆ2: 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.

Esin Durmus, Faisal Ladhak, and Tatsunori Hashimoto. 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.

Sergey Edunov, Myle Ott, Michael Auli, David Grangier, and Marc'Aurelio Ranzato. 2018. Classical structured prediction losses for sequence to sequence learning. 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 355–364, New Orleans, Louisiana. Association for Computational Linguistics.

Alexander Fabbri, Simeng Han, Haoyuan Li, Haoran Li, Marjan Ghazvininejad, Shafiq Joty, Dragomir Radev, and Yashar Mehdad. 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.

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

Katja Filippova. 2020. Controlled hallucinations: Learning to generate faithfully from noisy data. *arXiv preprint arXiv:2010.05873*.

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

Tanya Goyal and Greg Durrett. 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.

Tanya Goyal and Greg Durrett. 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.

Max Grusky, Mor Naaman, and Yoav Artzi. 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.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised contrastive learning for pre-trained language model fine-tuning. *arXiv preprint arXiv:2011.01403*.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 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.

Vivek Gupta, Prerna Bharti, Pegah Nokhiz, and Harish Karnick. 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.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

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

Daniel Kang and Tatsunori B. Hashimoto. 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.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

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

Kundan Krishna, Sopan Khosla, Jeffrey Bigham, and Zachary C. Lipton. 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.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 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.

Faisal Ladhak, Esin Durmus, He He, Claire Cardie, and Kathleen McKeown. 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.

Logan Lebanoff, John Muchovej, Franck Dernoncourt, Doo Soon Kim, Seokhwan Kim, Walter Chang, and Fei Liu. 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.

Hwanhee Lee, Kang Min Yoo, Joonsuk Park, Hwaran Lee, and Kyomin Jung. 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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yixin Liu, Zi-Yi Dou, and Pengfei Liu. 2021. RefSum: Refactoring neural summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1437–1448, Online. Association for Computational Linguistics.

Yixin Liu and Pengfei Liu. 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.

Yixin Liu and Pengfei Liu. 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.

Yixin Liu, Pengfei Liu, Dragomir Radev, and Graham Neubig. 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.

Patrice Lopez. 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.

Yao Lu, Yue Dong, and Laurent Charlin. 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.

Mary Ann Marcinkiewicz. 1994. Building a large annotated corpus of english: The penn treebank. *Using Large Corpora*, 273.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 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.

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

Feng Nan, Ramesh Nallapati, Zhiguo Wang, Cicero Nogueira dos Santos, Henghui Zhu, Dejiao Zhang, Kathleen McKeown, and Bing Xiang. 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.

Feng Nan, Cicero Nogueira dos Santos, Henghui Zhu, Patrick Ng, Kathleen McKeown, Ramesh Nallapati,

Dejiao Zhang, Zhiguo Wang, Andrew O. Arnold, and Bing Xiang. 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.

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

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 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.

Maxime Peyrard and Iryna Gurevych. 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.

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

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

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

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

Mathieu Ravaut, Shafiq Joty, and Nancy Chen. 2022a. SummaReranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4504–4524, Dublin, Ireland. Association for Computational Linguistics.

Mathieu Ravaut, Shafiq Joty, and Nancy F Chen. 2022b. Towards summary candidates fusion. *arXiv preprint arXiv:2210.08779*.

Simeng Sun, Ori Shapira, Ido Dagan, and Ani Nenkova. 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.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 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.

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

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

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 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.

David Wadden, Kyle Lo, Lucy Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 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.

David Wan and Mohit Bansal. 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.

John Wieting and Kevin Gimpel. 2017. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. *arXiv preprint arXiv:1711.05732*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 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.

Hanlu Wu, Tengfei Ma, Lingfei Wu, Tariro Manyumwa, and Shouling Ji. 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.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 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.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

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

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019a. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019b. 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.

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

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

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 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.

Jianing Zhou and Suma Bhat. 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.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. 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.

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