Do Multi-Document Summarization Models *Synthesize*?

Jay DeYoung[1] **Stephanie C. Martinez**[1] **Iain J. Marshall**[2] **Byron C. Wallace**[1]
[1]Northeastern University, Boston, MA, USA [2]King's College London, London, UK
deyoung.j@northeastern.edu martinez.s@northeastern.edu
iain.marshall@kcl.ac.uk b.wallace@northeastern.edu

Abstract

Multi-document summarization entails producing concise synopses of collections of inputs. For some applications, the synopsis should accurately *synthesize* inputs with respect to a key aspect, e.g., a synopsis of film reviews written about a particular movie should reflect the average critic consensus. As a more consequential example, narrative summaries that accompany biomedical *systematic reviews* of clinical trial results should accurately summarize the potentially conflicting results from individual trials. In this paper we ask: To what extent do modern multi-document summarization models implicitly perform this sort of synthesis? We run experiments over opinion and evidence synthesis datasets using a suite of summarization models, from fine-tuned transformers to GPT-4. We find that existing models partially perform synthesis, but imperfectly: Even the best performing models are over-sensitive to changes in input ordering and under-sensitive to changes in input compositions (e.g., ratio of positive to negative reviews). We propose a simple, general, effective method for improving model synthesis capabilities by generating an explicitly diverse set of candidate outputs, and then selecting from these the string best aligned with the expected aggregate measure for the inputs, or *abstaining* when the model produces no good candidate.

1 Introduction

Multi-document summarization (MDS) models aim to distill inputs into concise synopses that preserve key content. Examples of MDS include summarizing news articles (Dang, 2005; Fabbri et al., 2019; Gholipour Ghalandari et al., 2020; Evans et al., 2004), answering questions from multiple sources (Dang, 2006), and producing overviews of scientific literature (Liu et al., 2018; Lu et al., 2020; Mollá and Santiago-Martínez, 2012; Wallace et al., 2021; DeYoung et al., 2021). We expect summarization models to produce outputs consistent with inputs (Kryscinski et al., 2020; Nan et al., 2021b), e.g., discussing the same types of entities (Nan et al., 2021a) and allowing one to answer questions similar in a way that is consistent with individual inputs (Wang et al., 2020; Scialom et al., 2021).

In some applications models must *synthesize* inputs—i.e., aggregate potentially conflicting information—to yield an accurate synopsis (Figure 1). Consider the meta-reviews of movies featured on Rotten Tomatoes,[1] which provide a consensus view of individual critic opinions. These reviews should reflect the mean and range of sentiment implicit in the input critiques: A summary of mostly negative reviews (e.g., *Gigli*) should communicate that the film was widely panned; a summary of mixed reviews (*The Fifth Element*) ought to convey that critics disagreed and discuss the main positive and negative attributes.

A more consequential example is summarizing the evidence presented in clinical trials. Individual trials will often present conflicting evidence about whether or not a particular health intervention is effective. An ideal summary would appropriately weigh the findings presented in individual studies and reflect the evidence on balance.

What are the desiderata of multi-document *synthesis*? First, summaries produced by models should be *consistent* with the input data, with respect to the latent property of interest. In the case of Rotten Tomatoes, the sentiment of the summary should be in line with the aggregate sentiment expressed in the individual critic reviews. A corollary to this is that models should be *sensitive* to changes in the composition of inputs, e.g., removing most of the negative reviews from a set of inputs should yield a summary with a corresponding increase in the expressed sentiment.

[1]https://www.rottentomatoes.com/.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 1043–1062, 2024. https://doi.org/10.1162/tacl_a_00687
Action Editor: Annie Louis. Submission batch: 11/2023; Revision batch: 3/2024; Published 9/2024.

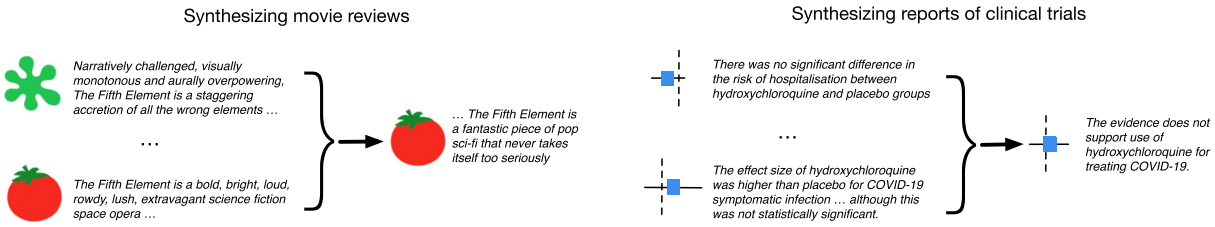

Figure 1: Two multi-document summarization tasks where models must implicitly synthesize inputs to produce accurate summaries. Left: Summarizing film reviews with varying sentiment to yield a *critics consensus*. Right: Summarizing trials that have evaluated a particular medical invention.

In this work we evaluate neural MDS models with respect to these criteria. To this end we use a meta-reviews dataset from Rotten Tomatoes (Leone, 2020) and a dataset of systematic reviews (meta-analyses) summarizing the evidence about medical interventions (Wallace et al., 2021). For the former we probe the degree to which generated meta-review sentiment agrees with the expected aggregate sentiment score; for the latter we evaluate whether the generated summary indicates that the input evidence suggests, on balance, that the intervention under consideration was effective.

Our **main contributions** are:

1. To the best of our knowledge, this is the first work to investigate implicit *synthesis* in summarization, and the degree to which modern models are capable of this.[2]
2. We show that ‘‘off-the-shelf’’ neural MDS models are somewhat inconsistent and insensitive with respect to performing synthesis in summarization.
3. We propose and evaluate a simple, general method of generating a diverse set of output candidates (Vijayakumar et al., 2016) and then selecting from these based on agreement with an expected aggregate measure (based on inputs), with promising results.

2 Synthesis and Summarization

In standard multi-document summarization, we assume inputs (X_i, y_i); $X_i = \{x_{i1}, \ldots, x_{i|X_i|}\}$. We then typically train a summarization model with parameters θ, to consume X_i and yield summaries $\hat{y}_i$ as similar as possible to targets y_i. In a supervised setting, the standard objective estimates a θ to maximize target token log-probabilities. Assuming the input documents x_{ij} in X_i have been linearized (i.e., concatenated, with special tokens demarcating individual inputs) into an input string $x_i^{\oplus}$, this objective takes the form: $\sum_{t=1}^{|y_i|} \log p_\theta(y_{it}|y_{i1}, \ldots, y_{i(t-1)}, x_i^{\oplus})$, where p_θ is a probability assigned to the token at position t in the target y_i by a summarization model with parameters θ. By myopically focusing on encouraging the model to produce tokens mimicking the targets, this objective aligns with standard (but flawed) measures of automated summary quality like ROUGE (Lin, 2004), which quantify n-gram overlap between targets y_i and outputs $\hat{y}_i$.

We are interested in settings in which there is an additional, latent property z_{ij} implicit in the constituent input texts x_{ij}. For example, z_{ij} might reflect the sentiment in critique j of the film indexed by i. Summaries should *synthesize* this aspect, i.e., the generated summary $\hat{y}_i$ should implicitly convey an aggregated z_i which reflects a synthesis or aggregation G over $Z_i = \{z_{i1}, \ldots z_{i|X_i|}\}$. That is, we assume $z_i = G(Z_i)$. In both cases considered here—summaries of film critiques and synopses of clinical trials evidence—G can reasonably be assumed to be a (weighted) mean, $G(Z_i) = \frac{1}{|X_i|}\sum_{j=1}^{|X_i|} \alpha_{ij} z_{ij}$. That is, summaries should roughly reflect the average sentiment and reported treatment effect in the cases of movie reviews and clinical trial reports, respectively.

We investigate the following questions. (1) Do model summaries $\hat{y}_i$ reflect the anticipated aggregate aspect of interest? That is, how well calibrated is the aspect communicated in the generated summary ($z_{i\hat{y}}$) compared to the expected z_i? (2) Do these same results apply to other (not solely transformer) MDS architectures? (3) Can we *improve* the ability of summarization models to synthesize

[2]Shah et al. (2021a) study a low-resource health and nutrition setting, in which they extract relational tuples, apply a manual rule set for aggregation, and then generate a surface form following this result. See Section 6 for a discussion of Opinion Summarization work which considers synthesis as a *target* but not measure of summarization performance.

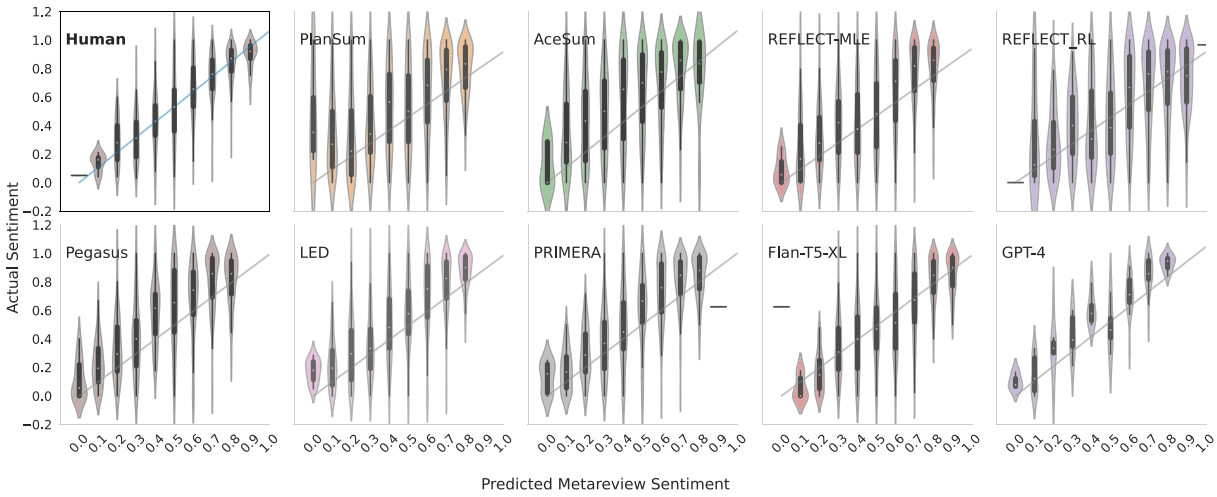

Figure 2: Movie Reviews: Actual vs. Predicted Sentiments on generated summaries. Human outputs replace LED (upper left) for comparison.

by explicitly incorporating synthesis targets z_i into the decoding process?

We propose a simple inference-time procedure to explicitly preference output candidates that align with the expected aggregate property of interest (e.g., average sentiment), and report promising results under both automatic and manual evaluation. This strategy naturally lends itself to *cautious* summarization, i.e., approaches where the model can *abstain* from generating an output if it does not produce any candidates that reflect the anticipated aggregate measure.

2.1 Movie Reviews

We first consider a dataset made up of movie reviews and associated meta-reviews summarizing these from Rotten Tomatoes. An in-house staffer (at Rotten Tomatoes) summarizes movie critic reviews[3] into meta-reviews (Barnes, 2017). These meta-reviews synthesize the input reviews, reflecting the aggregate critic reception of a film. Each meta-review is associated with a numerical ‘‘Tomatometer’’ score, which is an overall measure of the fraction of reviews that were positive (according to Rotten Tomatoes staffers) for the corresponding film (so here the target aggregation function G would be this fraction). The Rotten Tomatoes dataset we use comprises 9,095 movies with meta-reviews constructed from 244,000 individual reviews (Table 2).

[3]Written by designated ‘‘top-critics’’, critics recognized for quality and quantity of reviews in recognized publications.

Measuring Sentiment in Movie Reviews. We need to measure the property of interest in texts; for this we use a *measurement model* g—here we fine-tune a BERT model (Devlin et al., 2019) using the continuous (fine-grained) sentiment targets provided in the SST dataset (Socher et al., 2013).[4] We fine-tuned this model on the SST dataset for 3 epochs with a learning rate of 5e-5 using the `HuggingFace` library (Wolf et al., 2020) with no hyperparameter tuning. While the raw text of the SST dataset is in-domain (i.e., movie reviews), the targets themselves are not.[5] When applying this fine-tuned g to the movie meta-reviews, we find a reasonably strong correlation between our sentiment estimates and the ‘‘true’’ meta-review sentiment (‘‘Tomatometer’’ score): The R^2 (centered) is 0.696, mean squared error (MSE) is 0.022, and Pearson's r is 0.836 (Figure 2, upper left).[6]

2.2 Biomedical Systematic Reviews

Our second dataset is a collection of systematic reviews from the Cochrane Collaboration.[7]

[4]We use the *continuous* measurements from the original SST dataset, not the two or five class projections of those underlying measurements.

[5]SST is itself based on a collection of Rotten Tomatoes critic reviews (Pang and Lee, 2005). We verified that the SST text fragments do not overlap with our target reviews by manually checking any (fragment, review) pair with substantial ($\geq 75\%$) overlap for one quarter of all reviews.

[6] In creating both synthesis measures g, we have *isolated* them from our original datasets to not artificially favor human references as in-domain over machine generations.

[7]An international non-profit dedicated to helping healthcare providers make evidence-based decisions.

Study	Predicted Effect
Input: ...Ibuprofen was twice as likely as acetaminophen to abort migraine within 2 hours. In the intent-to-treat analysis, children improved twice as often with ibuprofen and acetaminophen as with placebo...	no significant difference
Input: ...Children's ibuprofen suspension at an OTC dose of 7.5 mg/kg is an effective and well-tolerated agent for pain relief in the acute treatment of childhood migraine, particularly in boys...	significant difference
Target: ...Low quality evidence from two small trials shows that ibuprofen appears to improve pain freedom for the acute treatment of children with migraine. We have only limited information on adverse events associated with ibuprofen in the trials included in this review...	no significant difference

Table 1: Systematic review example (from Cochrane). The statistical meta-analysis result ''significant difference'' and `RobotReviewer` finding ''no significant difference'' disagree. In the case of Systematic Reviews, `RobotReviewer` serves as both the estimator of z_{ij} and G.

	Movie Reviews			Systematic Reviews		
	Train	Dev	Test	Train	Dev†	Test
Number of metareviews	7251	932	912	1675	360	397
Avg metareview length	32.0	32.6	32.4	101	107	111
Total number of inputs	195033	24336	24474	11054	1238	2669
Avg number of inputs	26.9	26.1	26.8	6.6	3.4	6.7
Avg length of individual input	30.6	30.8	30.6	475	379	449
Avg length of concatenated inputs	822	804	822	2641	1336	2544
Target Percent Positive	59.5	62.1	61.2	31.9	31.4	35.0

Table 2: Dataset statistics for movie reviews (left) and systematic reviews (right). Number of meta-reviews, average meta-review length (tokens), input reviews per split, average number of inputs per instance, average total length of instance-inputs. For movie reviews, the target percent positive reports the fraction of metareviews with a positive sentiment; for systematic reviews this refers to the fraction of metareviews reporting a significant effect. † We subset the original dev set to instances of $\leq 4k$ tokens (accommodating T5; other models can consume up to 16k).

This dataset comprises roughly 2,600 systematic reviews summarizing a total of 16,500 clinical trials evaluating interventions in healthcare (Tables 1, 2). Each review includes a natural language summary and accompanying statistical meta-analysis results. The latter provides an aggregate statistical summary of the individual (study-level) data extracted from the trials included in each review. The natural language summary should accurately convey and contextualize the findings of the meta-analysis. Therefore, the (lack of) treatment efficacy communicated in a given summary should generally agree with the direction of the corresponding meta-analytic point estimate.

Measuring Effects in Evidence Syntheses. For systematic reviews of clinical trials, we resort to a less granular *classification* model $g(x_{ij}), g(y_i)$ which attempts to infer whether a text reports a significant result. Specifically, we use `RobotReviewer` (Marshall et al., 2017; DeYoung et al., 2020). Given a narrative describing a clinical trial result (or a summary of trials), `RobotReviewer` predicts whether the reported result indicates a significant effect of the treatment being investigated, or not. We can compare this prediction to the ''truth'', which here is derived from the meta-analytic result (specifically by checking whether $p < 0.05$). Applying this off-the-shelf model to the manually composed summaries accompanying the meta-analyses in our Cochrane set, we observe a macro-average F1 score of 0.577 and 68.6% accuracy, providing a reasonable (if weak) measure for this task.[6]

3 Models

We evaluate a suite of transformer (Vaswani et al., 2017) summarization models: Pegasus (Zhang et al., 2020), Longformer (Beltagy et al.,

2020), PRIMERA (Xiao et al., 2022), T5 (Raffel et al., 2020) and Flan-T5 (Chung et al., 2022), and GPT-4 (OpenAI, 2023). For each trainable transformer model and dataset we performed a hyperparameter search over learning rates and training steps (retaining most parameter defaults). We train with an effective batch size of 16 and floating point 16[8] precision on an NVIDIA RTX-8000 GPU (due to data size we can fit only a single instance in memory at a time for some models, and must use gradient accumulation).

Models were fine-tuned using the Adam optimizer (Kingma and Ba, 2014), except Pegasus, which was fine-tuned with Adafactor (Shazeer and Stern, 2018),[9] across several learning rates (1e-4, 1e-5, 1e-6), for up to 20k training steps. The best model was selected based on ROUGE-1 performance on the validation set.[10] PRIMERA was designed and pre-trained specifically for multi-document summarization. Though not explicitly designed as multi-document summarization models, both Pegasus (Zhang et al., 2020) and T5 (Amplayo et al., 2021) have been used on multi-document tasks, while Longformer has been used for a related multi-document summarization task (DeYoung et al., 2021).

For GPT-4 (-0613) we use system prompt *You are a professional movie critic. Your job is to provide an opinionated summary of a movie, in your own words. You will have access other critics' opinions of the movie.* and assistant prompt *For movie {movie}, other critics have written: {reviews}. In your own words, please produce an opinionated summary of {movie}.*, providing a one-shot example. For systematic reviews, we used the system prompt *You are a systematic reviewing expert. Your job is to read randomized control trial reports and assist a medical researcher. You will aid in drafting systematic reviews.* with assistant prompt: *Please provide a draft systematic review for the studies below: {studies}. Start with the conclusions of the review only, a more detailed analysis will happen later*, again providing a single shot example.

As it is not the focus of our work here, we did not extensively tune these prompts. We inspected outputs over five training instances when developing prompts for both movies and systematic reviews datasets. When designing movie review prompts, we iterated through first asking the model to *summarize* the reviews (yielding a summary of each review instead of an aggregate), then telling the model to use the same language as the reviews (with effectively the same result), then providing a single example (yielding some improvement), then demanding an *opinionated summary* (again with some improvement), and finally telling the model to use its own words (yielding the prompt above and experiments below). For the systematic review prompt, we first asked for a draft review (the model provided an entire draft), then we specified conclusions only (we received an abbreviated abstract), then we specified a conclusions *section* (we received a less abbreviated abstract), and, finally, adding an in-context example. We also explored asking for a high level *summary* (rather than systematic review) of the input studies; and with prompts providing intervention and outcome information to the model and asking for a draft of the review.

Beyond transformers, we consider models from the opinion summarization and content aggregation literature: PlanSum (Amplayo et al., 2020), QT (Angelidis et al., 2021), AceSum (Amplayo et al., 2021), and REFLECT (Song et al., 2022).[11] PlanSum (Amplayo et al., 2020) learns a (disentangled) sentiment and aspect model, and augments an LSTM equipped with an attention-copy mechanism (Bahdanau et al., 2014; Vinyals et al., 2015) with this information as a decoder.

QT (Angelidis et al., 2021) learns a quantized embedding for each model input via an auto-encoder, then finds representative input sentences (via clustering and assignment) to use as summaries. We include QT[12] as an extractive model. AceSum (Amplayo et al., 2021) adopts a hierarchical approach, representing each input document as sentences pooled over individual inputs, and passing this representation to a transformer (T5; Raffel et al., 2020), along with specific aspect or general codeword tokens and vocabulary embeddings, *controlling* what type of summary to

[8]Flan-T5-Large and -XL used BF16 for speed.

[9]In larger Flan-T5 models we experimented with both optimizers; differences in ROUGE1 performance were small.

[10]https://github.com/jayded/MDSSynthesis.

[11]We considered HierSumm (Liu and Lapata, 2019), but excluded it for extreme degeneration while decoding. We excluded Hercules (Hosking et al., 2023) as the software was not adaptable to our tasks.

[12]For movie reviews, where targets can appear similar to inputs in length and content, as opposed to systematic reviews (for which we do not evaluate QT), where the target prose differs substantially from its inputs.

	R^2	PCC	R1
QT	0.592	0.788	0.122
PlanSum	0.245	0.510	0.160
AceSum	0.158	0.439	0.176
REFLECT^{MLE}	0.430	0.657	0.241
REFLECT^{RL}	0.225	0.507	0.218
Pegasus	0.530	0.730	0.245
LED	0.551	0.742	0.242
PRIMERA	0.608	0.780	0.254
T5-Small	0.441	0.669	0.234
T5-Base	0.516	0.720	0.253
Flan-T5-S	0.412	0.647	0.237
Flan-T5-B	0.597	0.774	0.247
Flan-T5-L	0.484	0.696	0.248
Flan-T5-XL	0.611	0.783	**0.262**
GPT-4	**0.808**	**0.900**	0.166
Reference	0.697	0.836	

Table 3: Synthesis results for **Movie reviews**: correlations (R^2, Pearson's r) between sentiment measured in model outputs and Tomatometer Ratings. R1 is ROUGE1.

	F1	Acc	R1
PlanSum	0.414	0.683	0.177
AceSum	0.532	0.550	0.151
REFLECT^{MLE}	0.532	0.639	0.271
REFLECT^{RL}	0.505	0.683	0.199
Pegasus	0.568	0.714	0.212
LED	0.490	0.631	0.259
PRIMERA	0.526	0.644	0.253
T5-Small	0.540	0.600	0.205
T5-Base	0.521	0.628	0.206
Flan-T5-Small	0.548	0.583	0.081
Flan-T5-Base	0.538	0.683	0.194
Flan-T5-L	0.556	**0.692**	0.218
Flan-T5-XL	0.487	0.608	0.268
GPT-4	**0.628**	0.640	**0.273**
Reference	0.577	0.686	

Table 4: Synthesis results for **Systematic reviews**: Macro-averaged F1s and accuracies (`RobotReviewer` predictions over model outputs vs. reference meta-analysis results).

produce (we focus on the *general* case). REFLECT (Song et al., 2022) takes the hierarchical approach one step further, with a sentence level extraction phase (using aggregated token representations) followed by an abstraction phase (BART; Lewis et al., 2020), trained via standard MLE and via a reinforcement learning credit aware self-critic method (Rennie et al., 2017). For all models we largely retained the original hyperparameters, with modifications to increase sequence lengths and decrease aspects (these models were developed around *aspect* summarization).

4 Experiments

4.1 Do Summarization Models Synthesize?

We report sentiment performance for all models in Table 3. These metrics quantify the strength of the relationship between (a) the continuous sentiment inferred (via our text regression measurement model g) over model generated or reference summaries and (b) the reference sentiment (Tomatometer) score.

Save for GPT-4, correlations between the sentiment measured in generated outputs and Tomatometer scores are considerably lower than that between the same measurement over human-composed summaries and said score. This implies that human authors tend to do a better job of synthesis than models when composing summaries. GPT-4 seems performs especially well here; we are not entirely sure why, but it may owe to the differences in lengths of outputs (133 tokens on average vs. 31 for reference summaries).

For systematic reviews (Section 2.2), the measurement model g attempts to infer whether a text reports a significant treatment effect; we compare this against the p-value from the corresponding statistical meta-analysis. This permits a coarse assessment of synthesis, as we are unable to measure correlations. Instead we report classification metrics describing how often the effect significance inferred from a summary (generated or manually written) matches the ground truth derived from the meta-analysis (Table 4). The results are qualitatively similar to the sentiment case, in that the humans appear to do a better job of synthesis—as best we can measure, the significance reported in their summaries better aligns with the statistical results than in model generated summaries. GPT-4 is again an exception, slightly *outperforming* human results on this metric, which may owe to its formulaic generation featuring strong, direct, clear initial statements of treatment effectiveness.

4.2 Sensitivity to Input Ordering

Synthesis of inputs should be invariant to ordering (e.g., critic consensus on a film does not depend on the order in which one reads the reviews). Here we

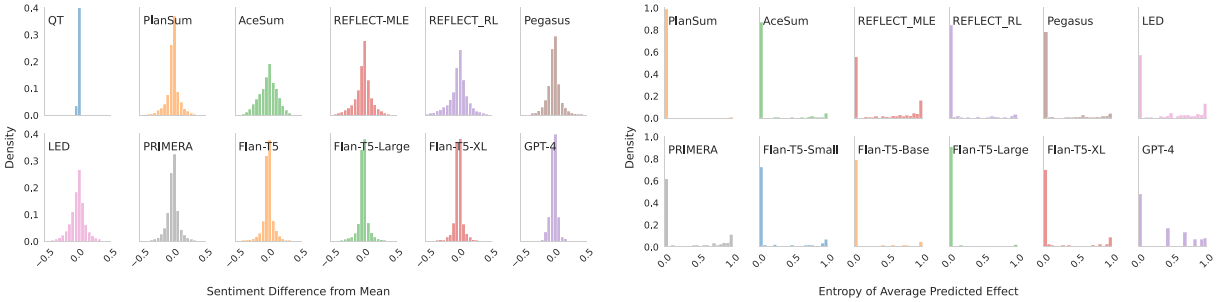

Figure 3: The spread of sentiment/treatment effect measured in outputs produced from permuted input orderings. Left: Movie review sentiment. Right: Systematic review significance prediction entropy (0 indicates order insensitivity) on the subset of reviews that report *significant* effects.

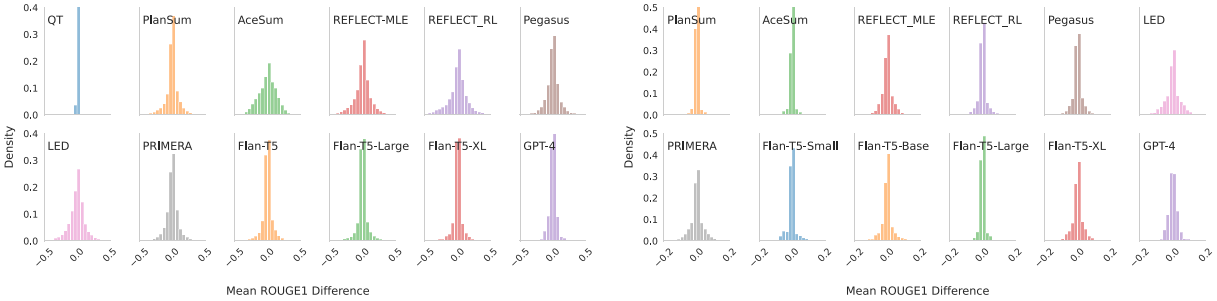

Figure 4: ROUGE1 deltas from instance means for movie reviews (left) and systematic reviews (right).

evaluate if models are sensitive to input ordering with respect to the synthesized aspect of interest ($z_{i\hat{y}}$). Specifically, let $X_i = \{x_{i1}, \ldots, x_{i|X_i|}\}$ denote an arbitrary ordering of inputs in the linearized version $x_i^{\oplus}$. This ordering should not affect the aggregate aspect $z_{i\hat{y}}$ in the summary.

To evaluate if models realize this invariance, we permute the instance i inputs X_i (and, consequently, the linearized $x_i^{\oplus}$) one hundred times,[13] randomizing input orderings. For each such permutation $\tilde{X}_i$ (and associated $\tilde{x}_i^{\oplus}$), we generate a summary $\hat{y}_i$ and estimate of the resultant aspect $\tilde{z}_{i\hat{y}}$, using the corresponding measurement model. By repeating this process for each instance i, we can construct an empirical distribution over $\tilde{z}_{i\hat{y}}$'s under different random orderings.

Movie Reviews. We zero-mean the $\tilde{z}_{i\hat{y}}$'s inferred over each instance, and combine the distributions from all instances into a histogram (Figure 3). This shows the spread of sentiments inferred over outputs under random input orderings minus the corresponding instance mean sentiment. Were a model completely invariant to ordering, the empirical distribution over these differences would collapse to 0. Instead, we observe a relatively wide spread in sentiment measured over outputs generated from different permutations, indicating a counter-intuitive sensitivity to orderings. (Interestingly, Figure 4—provided for comparison—suggests such permutations also affect ROUGE; we do not explore this aspect further here.)

Systematic Reviews. For each X_i we have 100 order permutations and associated summaries; we infer whether these report *significant results* or not, and record the fraction that do (p_i). If models were invariant to ordering, this fraction would always be 0 or 1. Values in-between suggest the model flips the report conclusion as a result of different input orderings. Figure 3 (right) shows a histogram of entropies over p_i, computed over the subset of examples where the associated meta-analysis indicates a significant effect. Densities away from zero indicate sensitivity to ordering. QT, PlanSum, and GPT-4 all have a smaller spread than the other models—QT because it is order insensisitive by construction, PlanSum similarly (but not entirely), and GPT-4 due to overall quality performance. We note that sensitivity is clearly an undesirable trait (*any* spread is

[13] As a cost saving measure, we sample ten times for GPT, over one hunded different inputs instead of the full development set. Our experiments cost approximately $500 to run.

	R^2	PCC
QT	0.634	0.796
PlanSum	0.249	0.499
AceSum	0.177	0.420
REFLECT^{MLE}	0.439	0.663
REFLECT^{RL}	0.294	0.542
Pegasus	0.499	0.706
LED	0.524	0.724
PRIMERA	0.572	0.756
T5-Small	0.447	0.668
T5-Base	0.481	0.694
Flan-T5-Small	0.393	0.627
Flan-T5-Base	0.556	0.746
Flan-T5-Large	0.490	0.700
Flan-T5-XL	0.551	0.742
GPT-4	0.457	0.677

Table 5: **Movie reviews** Correlations between subsampled inputs and generations.

undesirable), but this may trade off against other metrics of interest.

4.3 Sensitivity to Input Composition

Synthesis models should be responsive to changes in the distribution of the attribute to be synthesized in the input composition: If we increase the ratio of positive to negative reviews in an input set, we would anticipate a concomitant change in the sentiment communicated in the meta-review $z_{i\hat{y}}$. To assess if models meet this synthesis desiderata, we manipulate model inputs X_i in such a way to induce an expected change in the target measure $z_{i\hat{y}}$; we then measure if the output yields a summary that aligns with this expected change.

Movie Reviews. We manipulate the ratio of positive to negative reviews and observe the resultant change in the property of interest latent in the corresponding output. We take movies with mixed reviews, and delete 10%, 20%, 30%, . . . , 100% of the positive inputs, retaining the negative inputs; we then repeat the process but instead remove negative inputs. For each of these permutations, we measure the input sentiment, the meta-review sentiment, and how well they correlate (Table 5).

Figure 5 plots the relationship between the fraction of positive reviews in the (manipulated) input sets and the granular sentiment score inferred over the resultant outputs. The models are generally undersensitive to changes in their input: rather than having a change in meta-review sentiment equivalent in size to changes in input sentiment (a slope of 1, as we observe when we fit a model to the human written summaries). Models tend to have trouble changing their sentiment, and require a large change in input distribution to substantially change the sentiment communicated in the output.

Systematic Reviews. To measure sensitivity to changes in input composition, we manipulate inputs X_i such that the meta-analysis result (target $z_{i\hat{y}}$) flips from a significant effect to no effect, or from no effect to an effect (Table 6, Figure 6). We first take a subset of the reviews that have conflicting evidence (139 unique reviews). We then order inputs in these by (weighted) effect sizes,[14] and remove subsets which ought to flip the significance result of a subsequent meta-analysis. The surface level results (Table 6) show little difference from earlier results (i.e., the Δ values are approximately comparable to Table 4), but our classification results become substantially noisier (Figure 6). We speculate that models are picking up on some uncertainty from the change in overall meta-analysis but overall fail to capture that detail in their outputs. Even if the models reflect uncertainty due to the *strength* of the change (desirable!) this is still *incorrect* as the finding has changed.

Result. In both the case of the Movie Reviews and the Systematic Reviews, we see a substantial drop in performance from the base review results (reported in Tables 3, 4). We can only speculate as to the cause of this. Perhaps this indicates memorization of original targets in pre-training, or maybe removing strong (positive or negative) reviews hampers performance.

5 Improving Synthesis in Summarization

We propose a straightforward post-hoc approach to improving the synthesis performed by multi-document summarization models: (1) Generate an explicitly *diverse* set of output candidates; (2) Select from these as the final output the candidate that best agrees with the expected synthesis result (predicted by external model; Figure 7; Table 11).[15]

[14] In fixed effects meta-analysis the weights are inverse variances associated with study-level effect estimates.

[15] Oved and Levy (2021) explore a related generate-then-select approach for creating *plausible* product reviews. We experimented with an additional decoding method: constrain beam search by restricting candidate productions $p_\theta(y_{i,t}|y_{i,1..t-1}, x_i^{\oplus})$ such that the target attribute z_i is less

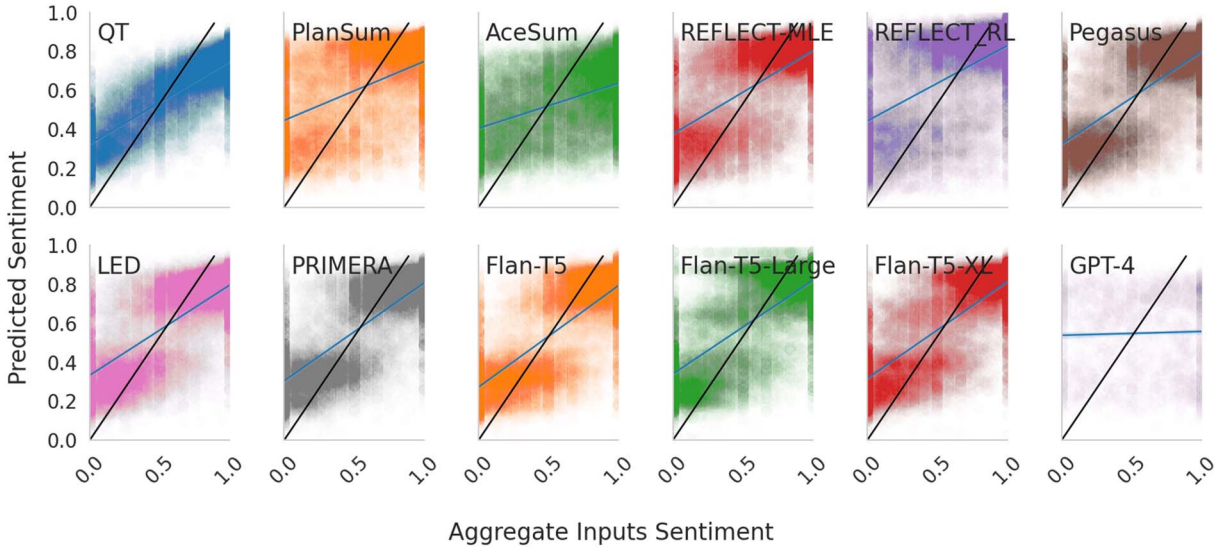

Figure 5: Model sensitivity to manipulated input sentiment composition. Intensity patterns indicate that models oscillate between low and high sentiments in outputs, and are not responsive to subtler shifts in input sentiment. We show a model regression (blue) and the reference sensitivity regression (black).

	F1	Acc
PlanSum	0.442	0.741
AceSum	0.454	0.504
REFLECT$^{\text{MLE}}$	0.471	0.583
REFLECT$^{\text{RL}}$	0.445	0.689
Pegasus	0.452	0.680
LED	0.510	0.684
PRIMERA	0.533	0.675
T5-Small	0.560	0.618
T5-Base	0.469	0.658
Flan-T5-Small	0.430	0.500
Flan-T5-Base	0.482	0.680
Flan-T5-Large	0.435	0.693
Flan-T5-XL	0.464	0.649
GPT-4	0.511	0.530

Table 6: **Systematic reviews**: Classification performance for subsampled inputs and generations. See Figure 6 for a visualization of classification *distribution*, analogous to Figure 5 for movies.

For (1), we rely on an existing technique for generating diverse outputs $\mathcal{C}_i$ from input $x_i^{\oplus}$: *Diverse Beam Search* (DBS) (Vijayakumar et al., 2016). This method modifies standard beam search to maintain multiple *groups* of beams. During decoding, a term is added to the next-token log probabilities, penalizing production of strings similar to candidates in *other* groups.[16]

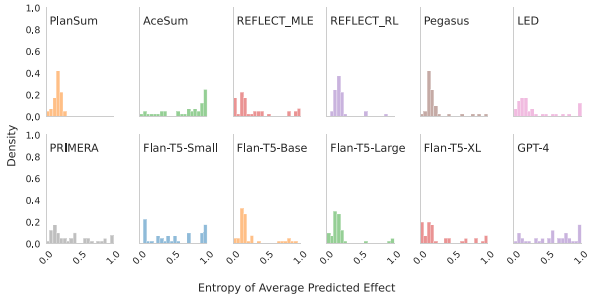

Figure 6: **Systematic Reviews.** A histogram of entropies for the *subsampled* review classifications (where the ground truth is positive).

In (2) we would like to select the output that best synthesizes the property of interest; this requires an approach to specify what we *expect* the synthesized property be, given the inputs. For example, if we know the sentiment scores associated with input movie reviews, we might enforce that the output sentiment agrees with the average of these. To realize this intuition, we can select as final output from $\mathcal{C}_i$ the string that best aligns with this aggregate property (sentiment score or significance finding). Operationally, this requires an external model to estimate the aspect of interest as latent in a given candidate output. This

than some ϵ: $|g(\hat{y}_{i,1,..,t}) - z_i| < \epsilon$. We elide these results here as they were often disfluent.

[16]This penalty requires a hyperparameter λ that encodes the relative importance of diversity; we use $\lambda = 0.5$. To enable fair comparison with standard beam search (5 beams, in all experiments), we used 5 groups, 1 beam per group. We exclude QT as it is an extractive model, and PlanSum as it does not readily support diverse beach search. For AceSum and REFLECT we modify these codebases to use the diverse beam search implementation from HuggingFace. For GPT-4 we sample five responses with a temperature of 0.6.

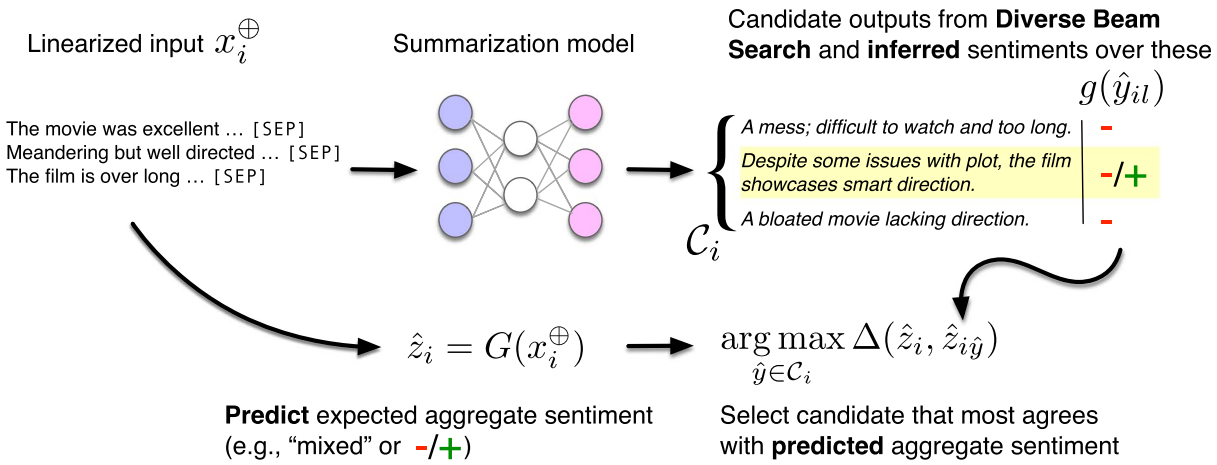

Figure 7: Our proposed strategy to improve synthesis. We generate an diverse set of output candidates (Vijayakumar et al., 2016) and then select the text that best agrees with the *predicted* aggregate property of interest (here, sentiment). We can also *abstain* when the model fails to yield an appropriate output.

	Approximate Selection						Oracle Selection					
	R^2	Δ	PCC	Δ	R1	Δ	R^2	Δ	PCC	Δ	R1	Δ
AceSum	0.566	**0.408**	0.769	**0.330**	0.162	−0.014	0.723	**0.565**	0.861	0.422	0.162	−0.014
REFLECT^{MLE}	0.658	0.228	0.825	0.168	0.241	0.000	0.791	0.361	0.895	0.238	0.240	−0.001
REFLECT^{RL}	0.491	0.266	0.702	0.195	0.220	0.002	0.576	0.351	0.759	0.252	0.219	0.001
Pegasus	0.694	0.164	0.835	0.105	0.229	−0.016	0.799	0.269	0.894	0.164	0.232	−0.013
LED	0.656	0.105	0.821	0.079	0.229	−0.013	0.763	0.212	0.878	0.136	0.227	−0.015
PRIMERA	0.749	0.141	0.880	0.100	0.240	−0.014	0.890	0.282	0.948	0.168	0.240	−0.014
T5-Small	0.692	0.251	0.846	0.177	0.225	−0.009	0.827	0.386	0.913	0.244	0.226	−0.008
T5-Base	0.721	0.205	0.856	0.136	0.231	−0.022	0.876	0.360	0.938	0.218	0.230	−0.023
Flan-T5-S	0.698	0.286	0.837	0.190	0.219	−0.018	0.832	0.420	0.912	0.265	0.218	−0.019
Flan-T5-B	0.732	0.135	0.863	0.089	0.225	−0.022	0.863	0.266	0.930	0.156	0.225	−0.022
Flan-T5-L	0.732	0.248	0.866	0.170	0.243	−0.005	0.875	0.391	0.937	0.241	0.244	−0.004
Flan-T5-XL	0.769	0.158	0.888	0.105	**0.250**	−0.012	0.900	0.289	0.950	0.167	**0.248**	−0.014
GPT-4	**0.814**	0.006	**0.924**	0.024	0.159	−0.007	**0.914**	0.106	**0.963**	0.063	0.164	−0.002
Reference	0.697		0.836				0.697		0.836			

Table 7: **Movie Reviews**: Generate diverse meta-reviews and select from them using an approximate (left) or oracle (right) target sentiment. Performance improves on every measure except ROUGE-1. Δs compare the metric to their left with the results reported in Table 3.

is a limitation of the approach, but in many settings it may be feasible to identify or construct a model; we were able to do so for both tasks considered here.

It may be that *any* member of $\mathcal{C}_i$ will align well with the anticipated aggregated property. In such cases, we have no means of producing an output consistent with respect to synthesis, and it may be desirable to *abstain* from outputting anything at all in such cases; that is, to be a *cautious* summarizer (Ferri et al., 2004; Hechtlinger et al., 2018). We consider this strategy in the case of generating narrative synopses of evidence, as this constitutes a case in which (a) one would very much prefer not to produce a misleading summary of clinical evidence (Kell et al., 2021), and, (b) we observe many cases where the diverse decoding strategy yields an output that seems to communicate (at a granular level) the aggregate findings expected.

Movie Reviews. We use BERT (Devlin et al., 2019), fine-tuned on IMDB (Maas et al., 2011)[17] to predict the sentiment inputs x_{ij}, using the proportion of $x_{ij} \in X_i$ with a positive score to approximate the target sentiment $z_{i\hat{y}}$. For each diverse prediction $\mathcal{C}_i$, we predict its sentiment $\tilde{z}_{i\hat{y}}$ via our regression model (2.1), and select the prediction closest to the estimated target sentiment $|\tilde{z}_{i\hat{y}} - z_{i\hat{y}}|$. We find this improves model synthesis performance (Table 7; Figure 8). Two authors blindly annotated 100 paired instances over PRIMERA generations for sentiment preference (matching the reference) between standard

[17] https://huggingface.co/lvwerra/bert-imdb.

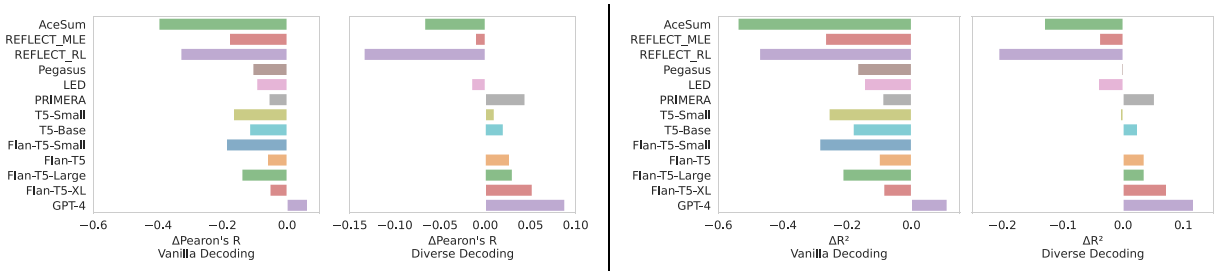

Figure 8: Differences relative to human summaries under vanilla decoding and the proposed generate-diverse then select strategy on movie meta-reviews. We report Pearson's r (PCC) and R^2 as measures of synthesis ''calibration''. Vanilla decoding yields synthesis performance worse than humans, but explicitly considering synthesis at inference time results in performance comparable to and sometimes better than the human summaries (as best we can measure).

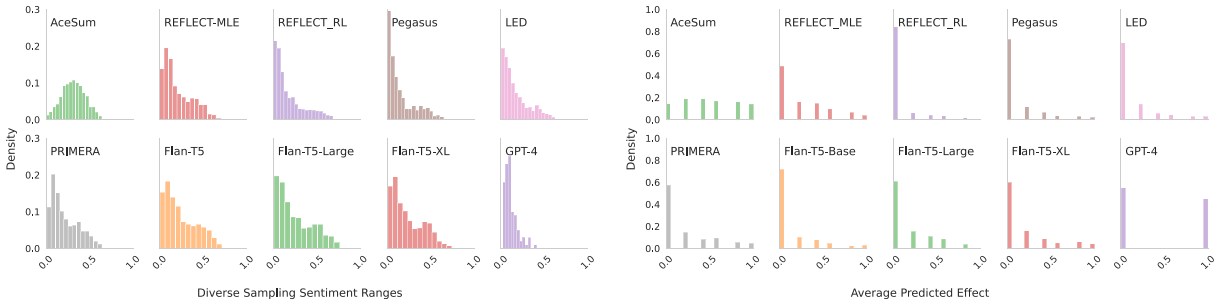

Figure 9: Distributions of outputs for the candiate summaries. **Movie reviews** (left) show a histogram for the range of differences between lowest and highest output sentiments. **Systematic reviews** (right) show histograms of the fractions of outputs reporting *significant* results.

and diverse outputs.[18] We find a moderate agreement Cohen's $\kappa = 0.59$, and a statistically significant preference for the diverse summaries ($p = 0.003$).

Systematic Reviews. For systematic reviews, we have a binary measure of *significant effect* (or not). As a proxy for $z_{i\hat{y}}$, we use `RobotReviewer` to extract an effect for each of the model inputs x_{ij}, using the majority vote (i.e., do the plurality of $x_{ij} \in X_i$ indicate that there was an effect). We classify each output candidate in $\mathcal{C}_i$ again using `RobotReviewer` to estimate $\tilde{z}_{i\hat{y}}$. We then select for output the highest probability candidate in $\mathcal{C}_i$ which agrees with the majority vote of the inputs, and abstain where there are no viable candidates. When we are able to choose a summary, we find performance similar to our measure (Table 9; Figures 8 and 9).

Result. Movie reviews show a wide range of sentiments; systematic reviews show some improvement but are biased towards no effect. Both settings show improvement from the switch to diverse decoding over standard beam-search methods: We repeat the generate-multiple-then-select approach with movie reviews (Table 8) and systematic reviews (Table 10). While the standard beam search did produce better overall scores when considering multiple candidates, the diverse generations produced higher correlations with human sentiment, and improved overall classification and abstention behaviors. Both settings have some decay in overall (crude) measures of review quality—Tables 7, 8 show small decreases in ROUGE-1 score; furthermore the diverse beam search results produce overall higher quality results (R^2, PCC), but how larger changes in ROUGE1 compared to a standard beam search method. Systematic Reviews behave similarly (Tables 9, 10), with an increase in F1 (or accuracy) comes with higher variability in ROUGE1 scores and a substantial amounts of abstention.

6 Related Work

Automatic (Multi-document) Summarization. (Nenkova and McKeown, 2011; Maybury, 1999)

[18] Summaries were ordered by difference in extracted sentiments between base outputs and diverse outputs, then 100 instances randomly selected from the top 20th percentile.

	Approximate Selection						Oracle Selection					
	R^2	Δ	PCC	Δ	R1	Δ	R^2	Δ	PCC	Δ	R1	Δ
AceSum	0.534	0.376	0.740	0.301	0.177	0.001	0.509	0.351	0.715	0.276	0.177	0.001
REFLECT^{MLE}	0.555	0.125	0.750	0.093	0.248	0.007	0.603	0.173	0.780	0.123	0.247	0.006
REFLECT^{RL}	0.406	0.181	0.638	0.131	0.222	0.004	0.454	0.229	0.675	0.168	0.221	0.003
PEGASUS	0.649	0.119	0.809	0.079	0.248	0.003	0.705	0.175	0.840	0.110	0.247	0.002
LED	0.653	0.102	0.815	0.073	0.241	−0.001	0.711	0.160	0.847	0.105	0.240	−0.002
PRIMERA	0.685	0.077	0.833	0.053	0.254	0.000	0.731	0.123	0.857	0.077	0.255	0.001
T5-Small	0.612	0.171	0.785	0.116	0.236	0.002	0.668	0.227	0.818	0.149	0.236	0.002
T5-Base	0.615	0.099	0.786	0.066	0.252	−0.001	0.669	0.153	0.819	0.099	0.253	0.000
Flan-T5-S	0.539	0.127	0.735	0.088	0.236	−0.001	0.579	0.167	0.803	0.156	0.251	0.014
Flan-T5-B	0.694	0.097	0.834	0.060	0.248	0.001	0.741	0.144	0.861	0.087	0.248	0.001
Flan-T5-L	0.732	0.248	0.866	0.170	0.243	−0.005	0.875	0.391	0.937	0.241	0.244	−0.004
Flan-T5-XL	0.769	0.158	0.888	0.105	0.250	−0.012	0.900	0.289	0.950	0.167	0.248	−0.014
Reference	0.697		0.836				0.697		0.836			

Table 8: **Movie Reviews**: Generate movie meta-reviews using standard beam search, then select using approximate (left) or oracle (right) target sentiments.

	Multiple-then-select							Oracle		
	F1	Δ	Acc	Δ	Abs	R1	Δ	Abs	R1	Δ
AceSum	0.562	0.030	0.573	0.023	0.088	0.154	0.003	0.133	0.152	0.001
REFLECT^{MLE}	0.588	0.056	0.626	−0.013	0.227	0.280	0.009	0.150	0.278	0.007
REFLECT^{RL}	0.605	0.100	0.700	0.017	0.430	0.197	0.002	0.247	0.207	0.008
Pegasus	0.633	0.065	0.676	−0.038	0.355	0.216	0.004	0.216	0.220	0.008
LED	0.625	0.135	0.698	0.067	0.355	0.250	−0.009	0.211	0.257	−0.002
PRIMERA	0.617	0.091	0.663	0.019	0.283	0.251	−0.002	0.180	0.250	−0.003
T5-Small	0.592	0.052	0.627	0.027	0.211	0.193	−0.012	0.169	0.190	−0.015
T5-Base	0.608	0.087	0.671	0.043	0.325	0.202	−0.004	0.197	0.210	0.004
Flan-T5-S	0.579	0.031	0.597	0.014	0.138	0.198	0.117	0.119	0.205	0.124
Flan-T5-B	0.660	0.122	0.723	0.040	0.358	0.222	0.164	0.177	0.222	0.028
Flan-T5-L	0.610	0.054	0.663	−0.029	0.212	0.212	0.065	0.152	0.206	−0.012
Flan-T5-XL	0.618	0.131	0.667	0.059	0.300	0.273	0.005	0.189	0.275	0.007
GPT-4	0.653	0.025	0.640	0.000	0.450	0.275	0.002	0.410	0.269	−0.004
Reference	0.577		0.686							

Table 9: **Systematic Review** results with multiple-then-selected predictions. We report macro-averaged F1 on the set of returned results. We abstain (Abs) when no output matches the expected synthesis result.

	Multiple-then-select							Oracle		
	F1	Δ	Acc	Δ	Abs	R1	Δ	Abs	R1	Δ
AceSum	0.578	0.046	0.588	0.038	0.197	0.157	0.006	0.255	0.153	−0.002
REFLECT^{MLE}	0.631	0.099	0.706	0.067	0.480	0.273	0.002	0.355	0.277	0.006
REFLECT^{RL}	0.603	0.098	0.753	0.070	0.483	0.188	−0.011	0.294	0.201	0.002
Pegasus	0.688	0.120	0.774	0.060	0.447	0.208	−0.004	0.258	0.216	0.004
LED	0.582	0.092	0.730	0.099	0.505	0.260	0.001	0.341	0.261	0.002
PRIMERA	0.625	0.099	0.704	0.060	0.436	0.259	0.006	0.313	0.250	−0.003
T5-Small	0.603	0.063	0.633	0.033	0.258	0.204	−0.001	0.233	0.201	−0.004
T5-Base	0.613	0.092	0.692	0.064	0.405	0.208	0.002	0.300	0.211	0.005
Flan-T5-S	0.603	0.055	0.632	0.049	0.361	0.081	0.000	0.333	0.080	−0.001
Flan-T5-B	0.637	0.099	0.761	0.078	0.500	0.195	0.001	0.300	0.198	0.004
Flan-T5-L	0.673	0.117	0.771	0.079	0.478	0.177	−0.041	0.281	0.174	−0.044
Flan-T5-XL	0.594	0.107	0.665	0.057	0.394	0.271	0.003	0.311	0.269	0.001
Reference	0.577		0.686							

Table 10: **Systematic reviews** results with multiple generate-then-select predictions, this time using the top-5 results from standard beam-search.

Summary	Sent.
You Don't Mess With the Zohan's handful of laughs are almost enough to compensate for its inconsistent tone and stale, obvious jokes.	0.243
You Don't Mess with the Zohan has a handful of crotch thrusts, but not enough of them land.	0.429
You Don't Mess With the Zohan's handful of laughs are almost enough to compensate for its aimless, crass script.	0.288
You Don't Mess with the Zohan has its moments, but not all of them – and the jokes are embarrassingly crass and often crude.	0.434
You Don't Mess with the Zohan has its moments, but not all of them – and the jokes are embarrassingly crass and often crude. The script	0.406

Table 11: Diverse meta-review generations and automatically inferred sentiment scores for ''You Don't Mess With The Zohan''. Target meta-review sentiment of 37%: We **bold** the closest generation in terms of (inferred) sentiment.

has been an active subfield within NLP for decades. We have focused our analysis on modern, neural abstractive models for conditional text generation (Bahdanau et al., 2015). In light of their empirical success, we have specifically evaluated a set of Transformer-based (Vaswani et al., 2017) models which have recently been used for multi-document summarization (Beltagy et al., 2020; Zhang et al., 2020; Xiao et al., 2022; Raffel et al., 2020). There has been some work on highlighting conflicting evidence in health literature specifically (Shah et al., 2021b,a), though this focused primarily on highlighting conflicting evidence and explicitly aggregating extracted content.

Multiple works have attempted gauge the difficulty of multi-document summarization. Wolhandler et al. (2022) measures the difficulty of abstractive multi-document news summarization as a function of inputs necessary to produce a final summary; they find that two to four well-chosen documents can cover a news topic sufficiently for the summarizer. They also find systematic reviews are particularly ill-suited to this minimal covering approach. Giorgi et al. (2022) study the impact of document retrieval behaviors on multi-document summarization performance, and find that models are sensitive to missing inputs.

Sentence Fusion. One view on synthesis might be that is a particular kind of *sentence fusion* (Barzilay and McKeown, 2005). However, past work on ''fusing'' sentences has assumed that the aim is to generate an output that contains the information common to similar sentences (Thadani and McKeown 2013). This is intuitive in the context of, e.g., summarizing multiple news articles covering the same event. But here we are interested in the more challenging setting in which the output should reflect an aggregate measure of potentially conflicting evidence or opinions.

Review and Opinion Summarization. considers a similar task to ours: Aggregating (usually product) reviews and opinions into a single coherent text. Oved and Levy (2021) developed a system with a similar generate-then-select approach, however this work was focused on generating *plausible summaries* rather than accurate *syntheses*, by selecting amongst candidates via a voting mechanism designed to mimic human preferences. Other related work has considered generating personalized and/or aspect-oriented summaries (He et al., 2017; Angelidis and Lapata, 2018; Amplayo and Lapata, 2020, 2021; Amplayo et al., 2021; Angelidis et al., 2021). Amplayo and Lapata (2021) propose a T5 variant for pooling instance representations, and also use Rotten Tomatoes as a dataset. This work (and Amplayo et al., 2021) includes a manual evaluation of how well system summaries are *supported* by input reviews, in contrast to how well a summary agrees with *all* inputs in the precise sense we have considered. We note that none of these prior works directly *probe* model responsiveness to changes in input composition.

Also related is the work of Chu and Liu (2019), which considered *unsupervised* approaches to multi-document summarization of Yelp! and Amazon reviews; they adopt an auto-encoder that ''decodes'' the mean of input representations to target summaries. They similarly note that output texts should convey mean input sentiment, and report ''sentiment accuracy'' as one of their metrics. But the synthesis aspect is not their main focus, and they consider only unsupervised settings (rather than the SOTA fine-tuned summarization models we have evaluated).

Interpretation and Analysis of Neural Models for NLP. This work is also related to the emerging body of work on analyzing neural NLP models, their behaviors, ''knowledge'', and ''abilities'' in general (e.g., Linzen et al., 2016; Tenney et al., 2019; Petroni et al., 2019; Niven and Kao, 2019; Meng et al., 2022). There has been some work specifically on analyzing neural summarization models. Xu et al. (2020a) investigated when a model is likely to copy rather than generate. Xu and Durrett (2021) assessed when models were relying on the local input to produce particular output tokens, and when they instead rely mostly on a background language distribution acquired in pre-training. In contrast to Giorgi et al. (2022) we explore beyond surface forms and explore the specific aspect of text *synthesis*.

Factuality of Neural Summarizers. Neural conditional generation models have proven adept at producing fluent outputs, but when summarizing they are prone to *hallucinating* content unsupported by input documents (Maynez et al., 2020; Kryscinski et al., 2019). Automated metrics such as ROUGE do not reliably capture such phenomena (Falke et al., 2019; Maynez et al., 2020). This has motivated the design of automated factuality metrics (e.g., Wang et al., 2020; Xu et al., 2020b); see Pagnoni et al. (2021) for an overview.

7 Conclusions

We have outlined and investigated the problem of *synthesis* as related to some summarization tasks. We showed that existing models are partially able to synthesize implicitly, but do so imperfectly: the aggregation they perform is sensitive to input ordering, and they are not as sensitive to perturbations in the composition of inputs as one would hope. Some models specifically designed for these tasks (AceSum, QT, REFLECT) are *less* sensitive to these perturbations, but offer worse overall performance than an equivalently sized transformer model (compare LED and REFLECT - REFLECT integrates a model with the same base LLM parameters as a portion of its synthesis model). Furthermore, increasing model size within an architecture can lead to fairly substantial improvements (LED to PRIMERA, T5 Small to Base, similarly for Flan-T5). Pretraining methods have some impact as well: T5 and Flan-T5 do not perform identically despite an identical model structure and comparable sizes, and GPT-4 clearly outperforms all models in this case, including the bespoke ones.

We proposed and validated a straightforward inference time method to improve model synthesis capabilities by preferentially outputting summary candidates that align with a predicted aggregate measure, and demonstrated empirically that this offers gains in performance. These gains are primarily limited by the underlying models' behaviors, but potentially bring performance on these single, task-specific metrics, on par to human performance, when the model is capable of providing a response that aligns with the proxy metrics.

We hope this work encourages additional research into summarization models that explicitly optimize to accurately synthesize potentially conflicting evidence. We are particularly interested in understanding *why* models fail to synthesize—they clearly learn to produce synthesis-like text, but fail to yield the best option, even among their top candidates. We use summary reranking as a means to surface these more-appropriate summaries, but this is solely post-hoc as opposed to controlling for a more suitable generation, or ideally improving base model performance.

Our methods focus solely on improving performance at single specific task measures, potentially at a cost to other review qualities. Users of such systems may have auxiliary goals, perhaps requiring multiple measures of synthesis quality, other measures of overall review quality, or a greater (or lesser) willingness to abstain. Abstinence can be a feature beyond the case of systematic reviews; systems may have other specific rules for when to abstain: e.g. toxic language, challenging to verify statements, or distance from an overall objective (i.e. abstaining in the movie reviews case).

This work has several limitations. We have made an effort to fine-tune several popular summarization models, but limited our analysis to models of relatively modest size (due to the GPU memory required to train long sequence summarization models). These behaviors appear to change with larger models (e.g., the small vs base-sized models, GPT-4 (OpenAI, 2023)), but building robustness to perturbations while maintaining sensitivity to input composition is a non-obvious challenge. We also have reported results on only English-language tasks. Finally, we focused on a relatively narrow behavior (synthesis

of a single aspect); models may succeed in this respect while failing in other ways.

Acknowledgments

This work was supported by the National Science Foundation (NSF) RI 2211954, and by the National Institutes of Health (NIH) under the National Library of Medicine (NLM), R01-LM012086-01A1. We thank the Northeastern University Research Computing Discovery Cluster. We thank the anonymous reviewers for their feedback.

References

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2020. Unsupervised opinion summarization with content planning. In *AAAI Conference on Artificial Intelligence*. https://doi.org/10.1609/aaai.v35i14.17481

Reinald Kim Amplayo, Stefanos Angelidis, and Mirella Lapata. 2021. Aspect-controllable opinion summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6578–6593, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.528

Reinald Kim Amplayo and Mirella Lapata. 2020. Unsupervised opinion summarization with noising and denoising. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1934–1945, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.175

Reinald Kim Amplayo and Mirella Lapata. 2021. Informative and controllable opinion summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2662–2672, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.229

Stefanos Angelidis, Reinald Kim Amplayo, Yoshihiko Suhara, Xiaolan Wang, and Mirella Lapata. 2021. Extractive opinion summarization in quantized transformer spaces. *Transactions of the Association for Computational Linguistics*, 9:277–293. https://doi.org/10.1162/tacl_a_00366

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1403

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR)*.

Brooks Barnes. 2017. Attacked by Rotten Tomatoes. *The New York Times*. Accessed 27 April 2023.

Regina Barzilay and Kathleen R. McKeown. 2005. Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328. https://doi.org/10.1162/089120105774321091

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

Eric Chu and Peter Liu. 2019. MeanSum: A neural model for unsupervised multi-document abstractive summarization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1223–1232. PMLR.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang,

Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Hoa Trang Dang. 2005. Overview of DUC 2005. In *Document Understand Conference*.

Hoa Trang Dang. 2006. Overview of DUC. In *Proceedings of HLT-NAACL*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu 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. https://doi.org/10.18653/v1/2021.emnlp-main.594

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.bionlp-1.13

David Kirk Evans, Judith L. Klavans, and Kathleen R. McKeown. 2004. Columbia newsblaster: Multilingual news summarization on the web. In *Demonstration Papers at HLT-NAACL 2004*, pages 1–4, Boston, Massachusetts, USA. Association for Computational Linguistics. https://doi.org/10.3115/1614025.1614026

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1102

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1213

César Ferri, Peter Flach, and José Hernández-Orallo. 2004. Delegating classifiers. In *Proceedings of the twenty-first international conference on Machine learning*, page 37. https://doi.org/10.1145/1015330.1015395

Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. 2020. A large-scale multi-document summarization dataset from the Wikipedia current events portal. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1302–1308, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.120

John Giorgi, Luca Soldaini, Bo Wang, Gary Bader, Kyle Lo, Lucy Lu Wang, and Arman Cohan. 2022. Exploring the challenges of open domain multi-document summarization. *ArXiv*, abs/2212.10526.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. An unsupervised neural attention model for aspect extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 388–397, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1036

Yotam Hechtlinger, Barnabás Póczos, and Larry Wasserman. 2018. Cautious deep learning. *arXiv preprint arXiv:1805.09460*.

Tom Hosking, Hao Tang, and Mirella Lapata. 2023. Attributable and scalable opinion summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8488–8505, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.473

Gregory Kell, Iain Marshall, Byron Wallace, and Andre Jaun. 2021. What would it take to get biomedical QA systems into practice? In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 28–41, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.mrqa-1.3, PubMed: 35663506

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 540–551, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1051

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. https://doi.org/10.18653/v1/2020.emnlp-main.750

Stefano Leone. 2020. Rotten Tomatoes movies and critic reviews dataset. kaggle.com

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535. https://doi.org/10.1162/tacl_a_00115

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by summarizing long sequences. In *International Conference on Learning Representations*.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5070–5081, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1500

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. https://doi.org/10.18653/v1/2020.emnlp-main.648

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Iain Marshall, Joël Kuiper, Edward Banner, and Byron C. Wallace. 2017. Automating biomedical evidence synthesis: RobotReviewer. In *Proceedings of ACL 2017, System Demonstrations*, pages 7–12, Vancouver, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-4002, PubMed: 29093610

Mani Maybury. 1999. *Advances in Automatic Text Summarization*. MIT Press.

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. https://doi.org/10.18653/v1/2020.acl-main.173

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. *Advances in Neural Information Processing Systems*, 36. ArXiv:2202.05262.

Diego Mollá and María Elena Santiago-Martínez. 2012. Creation of a corpus for evidence based medicine summarisation. *The Australasian Medical Journal*, 5(9):503.

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. https://doi.org/10.18653/v1/2021.eacl-main.235

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. https://doi.org/10.18653/v1/2021.acl-long.536

Ani Nenkova and Kathleen McKeown. 2011. *Automatic summarization*. Now Publishers Inc. https://doi.org/10.1561/9781601984715

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1459

OpenAI. 2023. Gpt-4 technical report.

Nadav Oved and Ran Levy. 2021. PASS: Perturb-and-select summarizer for product reviews. 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 351–365, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.30

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. https://doi.org/10.18653/v1/2021.naacl-main.383

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124, Ann Arbor, Michigan. Association for Computational Linguistics. https://doi.org/10.3115/1219840.1219855

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu,

and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1250

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1).

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1195. https://doi.org/10.1109/CVPR.2017.131

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.529

Darsh Shah, Lili Yu, Tao Lei, and Regina Barzilay. 2021a. Nutri-bullets hybrid: Consensual multi-document summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5213–5222, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.411

Darsh J. Shah, Lili Yu, Tao Lei, and Regina Barzilay. 2021b. Nutri-bullets: Summarizing health studies by composing segments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13780–13788. https://doi.org/10.1609/aaai.v35i15.17624

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4596–4604. PMLR.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Yun-Zhu Song, Yi-Syuan Chen, and Hong-Han Shuai. 2022. Improving multi-document summarization through referenced flexible extraction with credit-awareness. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1667–1681, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.120

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *International Conference on Learning Representations*.

Kapil Thadani and Kathleen McKeown. 2013. Supervised sentence fusion with single-stage inference. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1410–1418.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

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*.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Byron C. Wallace, Sayantan Saha, Frank Soboczenski, and Iain J. Marshall. 2021. Generating (Factual?) Narrative summaries of RCTs: Experiments with neural multi-document summarization. In *Proceedings of AMIA Informatics Summit*.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.450

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. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Ruben Wolhandler, Arie Cattan, Ori Ernst, and Ido Dagan. 2022. How ‘‘multi’’ is multi-document summarization? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5761–5769, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.389

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. https://doi.org/10.18653/v1/2022.acl-long.360

Jiacheng Xu, Shrey Desai, and Greg Durrett. 2020a. Understanding neural abstractive summarization models via uncertainty. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6275–6281, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.508

Jiacheng Xu and Greg Durrett. 2021. Dissecting generation modes for abstractive summarization models via ablation and attribution. 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 6925–6940, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.539

Xinnuo Xu, Ondřej Dušek, Jingyi Li, Verena Rieser, and Ioannis Konstas. 2020b. Fact-based content weighting for evaluating abstractive summarisation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5071–5081, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.455

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.