Multi-Review Fusion-in-Context

Anonymous ACL submission

Abstract

Grounded text generation, encompassing tasks such as long-form question-answering and summarization, necessitates both content selection and content consolidation. Current end-to-end methods are difficult to control and interpret due to their opaqueness. Accordingly, recent works have proposed a modular approach, with separate components for each step. Specifically, we focus on the second subtask, of generating coherent text given pre-selected content in a multi-document setting. Concretely, we formalize *Fusion-in-Context* (`FiC`) as a standalone task, whose input consists of source texts with highlighted spans of targeted content. A model then needs to generate a coherent passage that includes all and only the target information. Our work includes the development of a curated dataset of 1000 instances in the reviews domain, alongside a novel evaluation framework for assessing the faithfulness and coverage of highlights, which strongly correlate to human judgment. Several baseline models exhibit promising outcomes and provide insightful analyses. This study lays the groundwork for further exploration of modular text generation in the multi-document setting, offering potential improvements in the quality and reliability of generated content.

1 Introduction

Grounded text generation focuses on producing a passage from source texts, where the output is anchored around specific, task-dependant spans within the grounding texts. It pertains to tasks such as long-form question-answering (LFQA; Fan et al., 2019; Stelmakh et al., 2023), summarization (Nallapati et al., 2016a,b; Kulkarni et al., 2020), and information-seeking dialogue (Thoppilan et al., 2022; Shuster et al., 2022). These tasks inherently require identifying the relevant spans and then fusing them into a coherent output.

Grounded text generation is commonly approached with end-to-end procedures that combine the two underlying subtasks of content selection and fusion, recently using Large Language Models (LLMs) (Su et al., 2022; Shuster et al., 2022; Zhang et al., 2023). While effective, this approach often lacks flexibility and control over the generation process, given its black-box-like nature.

Addressing this, Slobodkin et al. (2022) recently advocated splitting grounded generation tasks into their two subtasks, and particularly focused on the fusion step. They introduced *Controlled Text Reduction* (CTR), a task where pre-selected spans in a source document ('highlights') are fused into a coherent text that exclusively covers the spans. This approach enhances control and modularity in text generation, enabling a single CTR model to work with various content selection strategies and user preferences, applicable in different contexts like summarization or long-form question-answering. It could also support human-in-the-loop scenarios for tailored outputs based on user preferences, as explored in Slobodkin et al. (2023b). Further, the direct access to the highlights that contribute to the output facilitates attributed generation (Bohnet et al., 2023; Gao et al., 2023a,b), where models can cite source spans for generated text.

Despite its benefits, CTR's focus on single-input scenarios limits its applicability to the broader, and more complex, multi-document setting. In this paper, we bridge this gap and extend the task to the multi-document setting. For that, we introduce the task of *Fusion-in-Context* (`FiC`), a generalized version of the CTR task, which processes multiple documents with pre-selected highlights, and aims to fuse them into a coherent, non-redundant text covering all and only the highlighted content, as demonstrated in Figure 1. In addition to the challenges of the single-input CTR task, including coreference resolution and proper discourse for coherence, the multi-document setting also requires

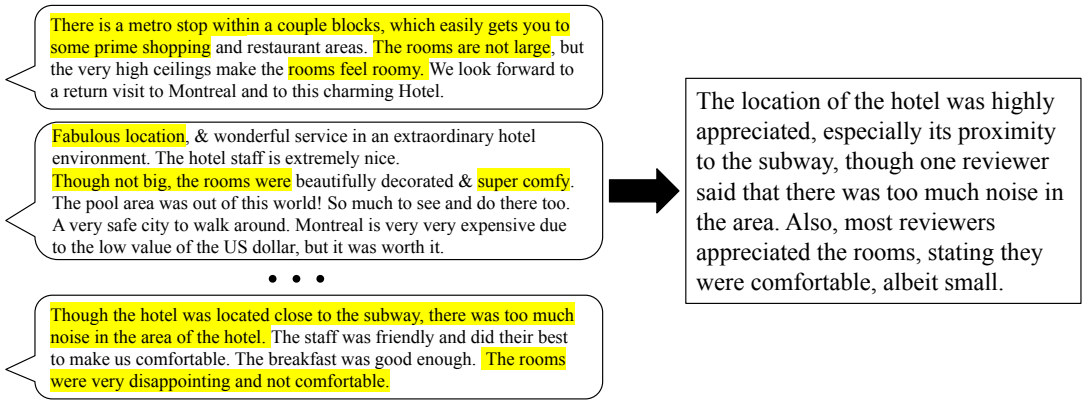


Figure 1: An example of an input, consisting of multiple reviews with highlights (left), and the generated text fusing the highlighted content while preserving coherence and non-redundancy (right). Such highlights in realistic use cases may be produced by different content-selection strategies.

handling repetitive, and sometimes conflicting information (Ma et al., 2020). Specifically, our work focuses on the business reviews domain, where contradicting opinions are more prevalent than in other more fact-oriented domains, such as news.

To promote research on FiC, we start by formally defining the task (§3). We then introduce a dataset (§4), carefully constructed via controlled crowdsourcing (Roit et al., 2020). Each of its 1000 instances comprises a set of inputs with highlights, and a corresponding fused text. The dataset is created through an efficient procedure, adapted from (Slobodkin et al., 2022), leveraging existing multi-document summarization datasets, specifically in the business reviews domain. We also develop an evaluation framework (§5) that assesses outputs' faithfulness and coverage of highlights, and explore various baseline models to benchmark performance in this setting (§6). Our findings reveal that while those models show promising results, there is still room for further improvement in future research.

Overall, the contributions of this paper are:

1. We introduce the *"Fusion-in-Context"* (FiC) task as a standalone module in modular multi-document grounded generation pipelines.
2. We develop the first high-quality FiC dataset.
3. We establish an evaluation framework for assessing the faithfulness to and coverage of highlights in a fused passage.
4. We present several supervised baseline models to set a foundation for future research.

2 Background

Grounded text generation, an area focusing on generating text from source documents, requires identifying relevant task-specific details within the inputs, such as salient content for summarization, as well as their coherent fusion. This field includes tasks like long-form question-answering (Fan et al., 2019; Stelmakh et al., 2023), summarization (Nallapati et al., 2016a,b; Shapira and Levy, 2020; Bražinskas et al., 2020b; Zhao et al., 2022), and dialogue systems (Yan et al., 2017; Xu et al., 2019; Thoppilan et al., 2022), with most related datasets aimed at end-to-end training (Fan et al., 2019; Bražinskas et al., 2020a; Liu et al., 2021; Iso et al., 2022a).

Despite the prevalence of end-to-end systems, there has been a growing trend towards decomposed pipeline approaches, particularly in summarization, with several recent studies focusing on content selection (Gehrmann et al., 2018; Lebanoff et al., 2020a; Ernst et al., 2021). Conversely, content fusion was largely explored at the full-sentence fusion level (Geva et al., 2019; Lebanoff et al., 2020b), with less emphasis on sub-sentence fusion.

Recently, Slobodkin et al. (2022, 2023a) have proposed a distinct separation of content selection from fusion, treating each as an independent task. They specifically concentrated on fusion, defining it as a standalone task termed *Controlled Text Reduction* (CTR). This task takes as input pre-selected spans, or 'highlights', within an input document, and requires a coherent merging of all the highlighted content, and nothing else. They also released a designated dataset and several CTR models showing strong adherence to these highlights.

While these studies acknowledged the benefits of decomposing grounded generation to subtasks, they mainly focused on single-document inputs. Our work builds on this decomposed approach, extending it to multi-document settings, which introduce new challenges such as managing longer

inputs, handling redundant highlights (Suzuki and Nagata, 2017; Calvo et al., 2018), and dealing with potentially conflicting facts or opinions (Kim and Zhai, 2009; Ma et al., 2022).

Additionally, previous CTR studies assessed highlight adherence by comparing outputs with the concatenated highlights, using lexical metrics like ROUGE (Lin, 2004) and METEOR (Banerjee and Lavie, 2005), and semantic metrics like BERTScore (Zhang et al., 2020). These methods, while suitable for single-input scenarios, are less effective for multi-document contexts where redundant and conflicting highlights are more prevalent. Additionally, these approaches did not distinctly evaluate faithfulness and coverage of highlights, typically assessing them jointly, with only manual evaluation for separate evaluation.

Addressing this, we explore more suitable metrics focusing separately on faithfulness and coverage of fused texts, inspired by recent progress in this area. Several recent studies have used Natural Language Inference (NLI) models for faithfulness evaluation (Laban et al., 2022; Schuster et al., 2022). There have also been advances in utilizing LLMs to evaluate faithfulness in a zero-shot setting with NLI-style prompts (Chen et al., 2023; Kocmi and Federmann, 2023; Liu et al., 2023), or after fine-tuning on synthetic data for faithfulness evaluation (Kryscinski et al., 2020; Yin et al., 2021; Gekhman et al., 2023). Yet, these works mainly targeted overall source text faithfulness rather than to specific segments. Moreover, they have not been widely applied to assess coverage, which has traditionally been evaluated using lexical metrics (Grusky et al., 2018) or manual evaluation (Syed et al., 2021). In our work, we adapt these methods to our highlights-focused setting, both for faithfulness and coverage, and assess their effectiveness.

3 Task Definition

The *Fusion in Context* (FiC) task is defined as the process of synthesizing a coherent text from a given set of documents, specifically focusing on pre-selected spans within these documents, referred to as *highlights*. Formally, given a document set D with marked spans $H = \{h_1, h_2, ..., h_n\}$ (such that h_i may be non-contiguous), a coherent and non-redundant passage f is generated, adhering to the following two criteria: (1) *highlight faithfulness* – f must be collectively entailed by the content in H, adding only minimal non-highlighted content required for coherence; (2) *highlight coverage* – each $h_i \in H$ must be represented in f, either explicitly, or via a generalized reference. For instance, if a highlight states *"the place serves great sushi"*, the output should either directly mention *"great sushi"* or refer to it in more general terms, such as *"great food"*. Moreover, the task permits the abstraction and aggregation of multiple highlights into a singule, synthesized statement. For example, separate highlights noting *"the beds were clean"*, *"the bathrooms were spotless"*, and *"the windows were clean"* could be collectively abstracted to a general statement like *"the rooms are clean"*. Overall, the goal is to produce a faithful, non-redundant and non-omissive, yet potentially abstractive and aggregated, fusion of the highlighted content.

4 Dataset for FiC

To comply with the task definition, an instance in a FiC dataset is expected to be a document set D with marked spans $H = \{h_1, h_2, ..., h_n\}$, and a corresponding fused text f. To compile such data, we leverage existing multi-document summarization datasets and extract high-quality FiC instances via controlled crowdsourcing (Roit et al., 2020), following Slobodkin et al. (2022), while adapting their method to the multi-text setting, and the business reviews domain.

4.1 Dataset Collection

Given a document set D and corresponding reference summary $\widehat{f}$ from an existing multi-document summarization dataset, the annotation process aims to identify the spans in the source texts $\{h_1, h_2, ..., h_n\}$ that cover all the information in $\widehat{f}$. This approach simplifies the annotation process compared to annotating from scratch, i.e., reading documents, marking highlights according to some specifications, and writing a coherently-fused text, which is reminiscent of standard formation of multi-document summarization datasets. Conversely, our approach requires locating and aligning spans between the source text and the already available reference summary, essentially "backwards engineering" the original human summarization process.

Source data. For our dataset, we turn to the business reviews domain, and sample review-sets and corresponding summaries from the CocoTrip (Iso et al., 2022b) and the FewSum (Bražinskas et al., 2020a) datasets. CocoTrip is a dataset of comparative opinion summaries of hotel review-sets,

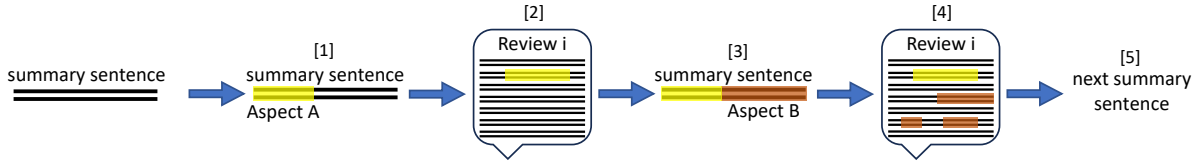


Figure 2: Illustration of the highlighting annotation process for a summary sentence, with reference to a specific review: [1] A *summary* aspect is identified and its statement highlighted; [2] Corresponding *review* spans are highlighted, and the alignment is saved; [3] Another *summary* aspect is identified and highlighted; [4] The matching *review* spans are highlighted, and the alignment is saved; [5] When all summary aspects that are alignable to the current review are highlighted, we proceed to the next sentence, and so on. In this example, the summary consists of two aspects, but steps 1 and 2 can be repeated as needed per sentence, until all alignable aspects are annotated. Borrowed and adapted from Slobodkin et al. (2022).

and FewSum consists of summaries of review-sets on businesses. Each review-set in these datasets comprises 8 reviews and upto 6 (average 3.13) corresponding reference summaries.

Annotation interface. To facilitate the annotation of alignments between reviews and their corresponding summary, we adapt a web-based annotation tool from (Slobodkin et al., 2022), and deploy it on Amazon Mechanical Turk[1] for crowdsourcing (§4.3 will explain the controlled crowdsourcing procedure). The application presents reviews and the respective summary side-by-side, and annotators are guided to highlight pairs of spans in the reviews and the summary that directly align. To reduce cognitive load, a summary is displayed alongside an individual review, and focus is placed on one summary sentence at a time. To further ease the process for annotators, lemmas in the review overlapping with lemmas in the currently focused summary sentence are emboldened.[2] This enables quick skimming through the review, however workers are trained not to rely solely on exact matches for highlighting (as reflected in §4.2 and §4.3).

Annotation procedure. Annotators are instructed to align statements dealing with a single aspect of a hotel or business (e.g., "room cleanliness") from the summary with the most relevant spans in the reviews, and to do so for all aspects in the summary in order to cover the whole summary text.[3] See §4.2 for the detailed annotation guidelines. This, in turn, creates instances of highlighted spans within reviews, with a corresponding coherent fusion of those highlights (the summary).

Each review-summary pair is annotated by a single trained annotator. To enhance quality, submissions are randomly sampled and reviewed, with feedback provided as necessary.

Resulting dataset. In total we sampled 1000 instances of review-set/summary pairs (700 instances from CocoTrip and 300 from FewSum). See Table 1 for full statistics.[4]

4.2 Annotation Guidelines and Data Traits

Figure 2 illustrates the annotation flow. When presented with a review of an entity (hotel or business) and a summary with a sentence in focus, an annotator first identifies aspects of the entity within the focused summary sentence. An aspect is not simply a facet of an entity, such as "rooms" or "staff", but more specifically it is a characteristic, such as "room cleanliness", "room style" or "staff helpfulness". (More on this in Appendix A.1.)

Upon identifying an aspect in the summary sentence, annotators are tasked with locating corresponding spans in the review. These are the minimal spans that adequately cover the information as in the summary regarding the aspect, where omitting any content would miss out on some detail of that aspect in the summary. For example, omitting any mention of the room being 'small' from the review highlights in Figure 1, would overlook this characteristic of the room, which is mentioned in the second summary sentence.

As outlined in §3, alignments on aspects need to consider two entailment-related traits. Firstly, a summary may express a generalized phrasing of an aspect that is stated in the reviews. For instance, a review may say *"great sushi"* while the summary might just say *"great food"*. Annotators are hence directed to also mark review excerpts that

[1] www.mturk.com

[2] Lemmatizing with spaCy (Honnibal and Montani, 2017).

[3] We observed that instructing annotators to focus on one aspect at a time enhances the efficiency in locating the relevant review spans, particularly when a summary sentence includes content that is scattered across different parts of the review.

[4] See Appendix F for more details.

	#unique sets of reviews	#summaries/ review-set (average)	#summary-review-set pairs	mean review/summary size (tokens)	max review/review-set/ summary (tokens)	mean review/summary size (sentences)	summary sentences aligning to multiple reviews	summary sentences aligning to multiple review sentences
Train	237	2.71	643	87.6/75.18	239/1118/231	5.89/5.08	82.51%	53.20%
Dev	23	4.30	99	77.97/69.05	197/829/174	5.47/4.71	87.34%	57.73%
Test	60	4.30	258	77.33/67.62	279/881/266	5.39/4.68	83.28%	51.82%
Overall	320	3.13	1000	83.99/72.62	279/1118/266	5.72/4.94	83.15%	53.29%

Table 1: Statistics of our dataset, including the number of unique review-sets, the average number of summaries per review-set, the number of summary/review-set pairs (a unique review-set creates a pair with each of its summaries), the mean review/summary size (in tokens and in sentences), the maximum review/review-set/summary size (in tokens), the percentage of summary sentences whose alignments span across more than one review, and the percentage of summary sentences whose alignments span across more than one review sentence within one of its reviews (namely, within a single review, the alignments come from more than one sentence).

are more specific than in the corresponding spans in the summary. Secondly, several spans in the reviews pertaining to the same aspect may yield an aggregated abstraction in the summary. Annotators must therefore also include review spans that exemplify the summary aspect. For example, aligning a review statement such as *"the beds were clean"* with a summary phrase *"the rooms were clean"*.

Additionally, reviews often express varying opinions about the same aspect, such as *"the service was great"* as opposed to *"the staff was unprofessional"*. When summarizing, all these varying opinions should be considered to reflect the overall sentiment. As a result, summary segments may range from statements like *"the staff was overall liked"* to *"some people liked the staff"*, depending on the spectrum of opinions. Hence, to properly capture this consolidation of differing viewpoints, annotators are also guided to align review mentions that either sentimentally entail *or* contradict the summary aspect. For example, the two above conflicting *review* spans should be aligned to the summary span *"the service was mostly good"*.

Finally, annotators may mark multiple spans in reviews that redundantly represent the same statement. The guidelines also address paraphrasing, non-consecutive highlights, and un-alignable summary spans. A detailed explanation of these guidelines is in Appendix A.2.

4.3 Annotator Training

The requirements of the aforementioned annotation process call for proficient-level annotations, which we achieved by means of controlled crowdsourcing (Roit et al., 2020). We identified qualified annotators through three open qualification rounds, followed by three closed rounds for selected annotators, focusing on further training and refinement. Each open round involves annotators reading a brief task description and accordingly aligning information between a single summary sentence and a short review, on a simplified interface. After each open round we reviewed the alignment and provided feedback. We then checked whether the annotator implemented our feedback in the following round (with a different sentence-review instance). If the annotator satsfyingly cooperated throughout the open rounds, they moved on to the closed rounds. The annotator started by watching a 25-minute tutorial on the full annotation tool and guidelines (§4.1 and §4.2). The closed rounds were conducted similarly to the open rounds, but with a whole summary and review, with all guidelines, and on the full interface. The qualification process was fully compensated with a customary wage, requiring up to 5 minutes per round. From this process, we were able to gather 8 trained annotators, who annotated the 1000 instances in our dataset.

4.4 Dataset Quality

To evaluate the quality of the compiled dataset, we compute the inter-annotator agreement. To this end, for every two annotators, we calculate intersection-over-union (*IoU*) of the tokens' indices (considering only content words) between the highlighted review spans that are aligned to the same summary sentence, similarly to Ernst et al. (2021). The *IoU* scores are gathered on the sentence level across three review-set/summary pairs, annotated by six crowdworkers. The resulting IoU score is 61.8.

To better understand the sources of disagreements, we analyzed all cases when $IoU < 90\%$. We found that the main cause of disagreement was related to our criteria for generalization and aggregation. Here, some annotators chose specific review spans they believed exemplified a summary characteristic, while others opted for different spans. This does not harm the quality of our data, as in all cases, the summary segment was indeed aligned with each of the corresponding re-

view spans, according to our criteria. Another common source of disagreement involved annotators including additional phrases that provided only insignificant extra details on top of the summary. For detailed examples, refer to Appendix G.

Finally, an interesting aspect of our dataset is that 80% of the summary sentences align to spans from multiple reviews, and over 50% of the summary sentences align with non-consecutive spans from different sentences within a single review (see Table 1). These properties reflect the real-world challenges faced by `FiC` models, expected to coherntly fuse disparate, and at times redundant, details.

5 Evaluation Framework

Consistent with the task definition in §3, a passage produced by a model as a fusion of highlights within source documents must uphold several criteria. (1) *Faithfulness*: it must only contain information from the highlights; (2) *Coverage*: it must cover all the information in the highlights, be it in an explicit, generalized, or aggregated form; (3) *Coherence and Redundancy*: it must convey the information in a well-structured and non-redundant form. In this section, we suggest several automatic metrics for faithfulness and coverage, and assess their effectiveness by correlating to human scores that we collected. Coherence and redundancy are measured using manual evaluation.

5.1 Limitations of Lexical and Semantic Matching

Output's adherence to highlights was previously measured in Slobodkin et al. (2022, 2023a) by comparing the output passage and the concatenated highlights, using lexical metrics like **ROUGE** (n-gram matching) and **METEOR** (word matching with synonyms), and semantic metrics like **BERTScore** (probability of generating the output text). Our work, however, extends beyond the single-document scenario explored in these previous works, to also include multi-document contexts. This shift introduces additional complexities, such as managing redundancy and contradictions among highlights drawn from diverse sources, which may not be fully captured by standard lexical and semantic matching techniques. Further, our setting also enables highlights aggregation and generalization, which these metrics may not adequately address. Additionally, these automated approaches primarily measured overall adherence to the highlights without making a distinction between faithfulness and coverage. These latter aspects were evaluated manually, but only on a limited number of instances.

5.2 NLI-based Faithfulness Metric

Highlight-faithfulness requires the output passage to be entailed by the collective highlighted content. We employ the `flan-t5-xxl` model (Chung et al., 2022), shown to exhibit high performance on NLI tasks, for evaluating faithfulness to highlights in a zero-shot setting with a standard **NLI** prompt (see Appendix B). Previous research that used NLI models for faithfulness evaluation in summarization (Maynez et al., 2020; Laban et al., 2022; Honovich et al., 2022) typically set the grounding text as the premise, and the generated text as the hypothesis. Accordingly, we set the highlights concatenation to serve as the premise, since the outputs are expected to be entailed by all the highlighted content collectively (see §3). For the hypothesis, we segment the output passage into sentences, with each sentence serving as a separate hypothesis. The average of the sentence-level entailment scores is used as the overall entailment probability of the corresponding passage. This approach, inspired by (Laban et al., 2022), was found to be more effective than using the entire output as a single hypothesis.[5]

5.3 Trained Coverage Metric

Inspired by recent work that evaluates faithfulness and factuality using a dedicated **trained** model (Yin et al., 2021; Utama et al., 2022; Gekhman et al., 2023; Soleimani et al., 2023), we finetune an LLM that is tasked to assess whether the generated passage fully covers the highlights. In our methodology, each highlight is individually input along with the entire output, and the model outputs a binary answer for whether the highlight is contained in the passage.[6]. We derive synthesized training data for this task from our `FiC` dataset, using highlights and their corresponding summaries. For negative samples, we remove the summary sentence that aligns with the highlight. For positive samples, a random non-aligning summary sentence is omitted (to avoid a potential bias caused by sentence exclusion in the negative samples). We finetune a `flan-t5-large` model (Chung et al., 2022) with

[5] We also experimented with other methods for evaluating faithfulness and coverage, which exhibited lower correlation to human judgment. See Appendix E for more details.

[6] We also tried concatenating all the highlights together, and found it to be inferior.

Metric	Faithfulness		Coverage	
	τ	CI	τ	CI
ROUGE-1 (R)	0.2319	0.23-0.24	0.3467	0.34-0.35
ROUGE-1 (P)	0.5468	0.54-0.55	-0.0533	-0.06–0.05
ROUGE-2 (R)	0.3555	0.35-0.36	0.2731	0.27-0.28
ROUGE-2 (P)	0.5253	0.52-0.53	0.0071	0.00-0.01
ROUGE-L (R)	0.0958	0.09-0.10	0.3835	0.38-0.39
ROUGE-L (P)	0.4898	0.48-0.49	-0.0367	-0.04–0.03
METEOR	0.4017	0.40-0.41	0.2736	0.27-0.28
BERTScore (R)	0.2380	0.23-0.24	0.4165	0.41-0.42
BERTScore (P)	0.6004	0.59-0.60	0.0529	0.05-0.06
NLI (Faithfulness)	**0.6745**	**0.67-0.68**	0.0929	0.09-0.10
Trained (Coverage)	0.1771	0.17-0.18	**0.4992**	**0.49-0.50**

Table 2: Average Kendall-Tau rank correlations (τ) and their 95% confidence intervals (CI) for tested evaluation metrics against human judgment. Recall-based metrics (R) are more effective for coverage, and precision-based metrics (P) for faithfulness. Best correlations for each axis are in bold.

the synthesized coverage data. The input to the model is the highlight and modified summary, and the output is 'yes' or 'no', for positive and negative samples, respectively. The final score is the average probability of the token 'yes' across all highlights.[7]

5.4 Meta-Evaluation

Setup. To assess our evaluation metrics we follow the common practice (Fabbri et al., 2021) of correlating scores to human judgment. To that end, we gather faithfulness and coverage ratings for generated outputs from three co-authors of this paper. The outputs were produced by two models (see Flan-T5$_{\text{H}}$ and Flan-T5$_{\text{no-H}}$ in §6.1). A total of 50 review sets were randomly selected from our test set, leading to 100 scores for each of coverage and faithfulness. A 1-to-7 Likert scale was used to rate faithfulness and coverage separately for an output.

To ensure agreement among annotators, the three authors first evaluated a separate set of 10 outputs, and inter-annotator agreement was computed with Cohen's Kappa coefficient (Cohen, 1960). The average Kappa coefficients were 0.49 and 0.42 for faithfulness and coverage, respectively, indicating a moderate level of agreement (Viera et al., 2005). For more details, see Appendix E.3.

After collecting scores for the 100 instances, we computed their correlation with human judgment using Kendall-Tau rank correlation, as suggested in (Deutsch et al., 2022).[8] We also apply bootstrapping (Efron, 1987) by performing 1000 samplings of 70 instances (with repetition) and calculating correlation scores for each such subset. We report the average correlation and 95% confidence intervals for each metric.

Results. Table 2 shows the average correlations with their 95% confidence intervals for faithfulness and coverage. We find that while certain lexical- and semantic-based metrics yield decent results, notably BERTScore-precision for faithfulness and BERTScore-recall for coverage, our proposed metrics demonstrate significantly higher correlations, with average values of 0.6745 and 0.4992 for faithfulness and coverage, respectively. In light of these findings, we employ our NLI-based and trained metrics for assessing model performance in terms of faithfulness and coverage, respectively (in §6.2).

5.5 Human Evaluation of Coherence and Redundancy

We adopt the coherence assessment methodology from (Slobodkin et al., 2022). Crowdworkers judge the coherence of 100 randomly selected instances from the test set, for each examined model. A score between 1 and 5 is specified, and each passage is reviewed by three workers and averaged. Similarly, the redundancy of information in a passage is appraised. This approach follows standard practice, where coherence and redundancy are best evaluated manually (Fabbri et al., 2021; Steen and Markert, 2021). For more details see Appendix D.

6 Experiments

6.1 Experimental Setup

We examine several baseline models for solving the `FiC` task. The input to a model is a document set with spans marked within the documents (highlights), and the model ouputs a fused passage around the highlights.

Models with full input. Using the training set of our dataset, we finetune a large language model, marking the highlights in the input via designated mark-ups, following Slobodkin et al. (2022). Specifically, we finetune a `flan-t5-large` model (Chung et al., 2022), that exhibited enhanced performance in tasks requiring constrained generation (Sanh et al., 2022; Wei et al., 2022). We will refer to

[7] We also explored an NLI-based coverage metric, where the passage serves as the premise and the highlights function as the hypothesis. We found it to achieve comparable results, however it requires substantially more computation time and memory. For more details, see Appendix E.

[8] Spearman correlations were also calculated, showing similar trends. See Appendix E.4.

Model	Faithfulness	Coverage	F-1	Coherence	Redundancy
Flan-T5$_{\text{H}}$	72.8	86.4	79.0	4.3	4.1
Flan-T5$_{\text{H}}$ (RL)	54.0	82.0	65.1	4.1	4.0
Flan-T5$_{\text{only-H}}$	**84.6**	**87.8**	**86.2**	3.6	3.8
Flan-T5$_{\text{no-H}}$	53.7	76.9	63.2	4.1	3.9
GPT-4	81.6	85.6	83.6	**4.7**	**4.5**

Table 3: Results for the proposed models on our FiC dataset. Faithfulness is measured with our NLI-based metric, and Coverage with our trained metric. The F-1 is a harmonic mean of the two latter scores. Coherence and Redundancy are measured through manual assessment. For each metric, the best score is in bold.

this model as **Flan-T5$_{\text{H}}$** ('H' for 'Highlights'). We develop an additional variant of Flan-T5$_{\text{H}}$, which we further finetune using Reinforcement Learning (**RL**), following the method in Slobodkin et al. (2023a). It applies the Quark algorithm (Lu et al., 2022) combined with a dual-reward policy (Pasunuru and Bansal, 2018), alternating between our NLI-based faithfulness and trained coverage metrics (§5) as rewards. We also examine the performance of a one-shot **GPT-4** model (OpenAI, 2023), guided with an example of the task.[9]

Models with highlights only. To reveal the importance of the surrounding context, we also train a `flan-t5-large` model only with a concatenation of the highlights as the input (excluding surrounding context). We denote this variant **Flan-T5$_{\text{only-H}}$**.

Models without highlights. Finally, we examine `flan-t5-large` in a standard summarization setting, where it is finetuned with the input review-set without the highlighted spans, denoting this variant **Flan-T5$_{\text{no-H}}$**. It offers insights into the model's ability to pick up on signals that point to highlights.

6.2 Results

We apply our evaluation metrics on the proposed systems, with results presented in Table 3. We first observe that the exclusion of context from the input (Flan-T5$_{\text{only-H}}$) yields the strongest faithfulness and coverage scores, yet the lowest coherence and redundancy scores. This shows the importance of incorporating context for more seamless outputs. Meanwhile, the removal of highlights (Flan-T5$_{\text{no-H}}$) leads to a substantial degradation in faithfulness and coverage. This indicates that the model indeed succeeds in learning to adjust the output according to the highlights, underlining the highlights' role in enhancing the model's performance.

[9]Preliminary experiments on a separate development set, with varying numbers of in-context examples, indicated that a single exemplar yields the best results. See Appendix C.

Interestingly, even though the RL reward functions used in the RL-enriched model are the faithfulness and coverage metrics themselves, the outputs are eventually negatively affected when evaluating with these metrics. This result calls for a more in-depth investigation of enhanced reward functions that can leverage the benefits of RL-enrichment, as was shown to be helpful in Slobodkin et al. (2023a) for the single-input setup. We also find that single-shot GPT-4 yields the most coherent and least redundant texts. While it ranks highly in faithfulness and coverage, it is still overtaken by the finetuned Flan-T5$_{\text{only-H}}$. Overall, our findings invite for further research on the FiC task, to develop fusion strategies that ensure comprehensive coverage and faithfulness to highlighted content, with coherent and low-redundancy outputs.

7 Conclusion

In this paper, we further promote the decomposition of grounded text generation as presented in (Slobodkin et al., 2022), extending it to the multi-document setting. To that end, we introduce the Fusion-in-Context (FiC) task, an extension of the task from (Slobodkin et al., 2022) which focuses on the content fusion step, to the multi-document setting. The FiC setting facilitates employing a single general-purpose fusion model for diverse content selection needs, capturing the challenges of repetitiveness and contradictions in source documents. To advance the task, we prepared a high-quality dataset, esablished an evaluation framework for faithfulness and coverage of selected spans, and provide several baseline models to stimulate further research and exploration.

Future work may include expanding the FiC task to other multi-input contexts, e.g., the news domain. We also plan to investigate ways to leverage the built-in traceability of the output text's origin, namely the highlights, for facilitating attributed generation.

8 Limitations

In this work, we construct the first FiC dataset, developed by instructing crowdworkers to identify relevant spans within reviews that align with the content of corresponding summaries. To reduce cognitive load, each summary was displayed alongside individual reviews. While this approach streamlined the annotation process, there are instances where viewing the complete set of input reviews is advantageous, particularly for aggregative summary segments. In such segments, multiple review spans are combined into a single summary span, necessitating a broader understanding of the entire input set for accurate highlighting.

Moreover, the focus of our dataset on the business reviews domain may constrain its generalizability to other contexts with distinct textual structures, like news articles. This limitation extends to our trained evaluation metrics, which were developed using a derivative of our crowdsourced dataset and, therefore, are tailored to the specific characteristics of business reviews.

9 Ethics Statement

The proposed Fusion-in-Context (FiC) task, despite offering enhanced control over the content generated, is not expected to achieve complete resolution. Therefore, integrating FiC modules in modular generative systems should be done so with caution, since there is a possibility that these modules may overlook certain highlighted content or inadvertently include content that was not highlighted. This concern is particularly relevant for future endeavors that aim to use FiC for attributed generation. In such cases, there is a risk that some portions of the generated content may not be directly traceable to the pre-defined highlighted segments, leading to potential inaccuracies, or incompleteness, in attribution.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2023. Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models. ArXiv:2212.08037 [cs].

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020a. Few-shot learning for opinion summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4119–4135.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020b. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Hiram Calvo, Pabel Carrillo-Mendoza, and Alexander Gelbukh. 2018. On redundancy in multi-document summarization. *Journal of Intelligent & Fuzzy Systems*, 34(5):3245–3255.

Shiqi Chen, Siyang Gao, and Junxian He. 2023. Evaluating factual consistency of summaries with large language models.

Hyung Won Chung, Le Hou, Shayne 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, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. Re-examining system-level correlations of automatic summarization evaluation metrics. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 6038–6052.

Bradley Efron. 1987. Better bootstrap confidence intervals. *Journal of the American statistical Association*, 82(397):171–185.

Ori Ernst, Ori Shapira, Ramakanth Pasunuru, Michael Lepioshkin, Jacob Goldberger, Mohit Bansal, and Ido Dagan. 2021. Summary-source proposition-level alignment: Task, datasets and supervised baseline. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 310–322, Online. Association for Computational Linguistics.

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

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023b. Enabling Large Language Models to Generate Text with Citations. ArXiv:2305.14627 [cs].

Sebastian Gehrmann, Yuntian Deng, and Alexander Rush. 2018. Bottom-up abstractive summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4098–4109, Brussels, Belgium. Association for Computational Linguistics.

Zorik Gekhman, Jonathan Herzig, Roee Aharoni, Chen Elkind, and Idan Szpektor. 2023. Trueteacher: Learning factual consistency evaluation with large language models.

Mor Geva, Eric Malmi, Idan Szpektor, and Jonathan Berant. 2019. DiscoFuse: A large-scale dataset for discourse-based sentence fusion. 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 3443–3455, Minneapolis, Minnesota. 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.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Hayate Iso, Xiaolan Wang, Stefanos Angelidis, and Yoshihiko Suhara. 2022a. Comparative opinion summarization via collaborative decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3307–3324.

Hayate Iso, Xiaolan Wang, Stefanos Angelidis, and Yoshihiko Suhara. 2022b. Comparative Opinion Summarization via Collaborative Decoding. In *Findings of the Association for Computational Linguistics (ACL)*.

Hyun Duk Kim and ChengXiang Zhai. 2009. Generating comparative summaries of contradictory opinions in text. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 385–394.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

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.

Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. 2020. AQuaMuSe: Automatically Generating Datasets for Query-Based Multi-Document Summarization. ArXiv:2010.12694 [cs].

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Logan Lebanoff, Franck Dernoncourt, Doo Soon Kim, Walter Chang, and Fei Liu. 2020a. A cascade approach to neural abstractive summarization with content selection and fusion. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 529–535, Suzhou, China. Association for Computational Linguistics.

Logan Lebanoff, Franck Dernoncourt, Doo Soon Kim, Lidan Wang, Walter Chang, and Fei Liu. 2020b. Learning to fuse sentences with transformers for summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4136–4142, Online. Association for Computational Linguistics.

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

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. Durecdial 2.0: A bilingual parallel corpus for conversational recommendation.

Ximing Lu, Sean Welleck, Jack Hessel, Liwei Jiang, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. 2022. QUARK: Controllable text generation with reinforced unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z Sheng. 2020. Multi-document summarization via deep learning techniques: A survey. *arXiv preprint arXiv:2011.04843*.

Congbo Ma, Wei Emma Zhang, Mingyu Guo, Hu Wang, and Quan Z Sheng. 2022. Multi-document summarization via deep learning techniques: A survey. *ACM Computing Surveys*, 55(5):1–37.

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.

Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016a. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.

Ramesh Nallapati, Bowen Zhou, and Mingbo Ma. 2016b. Classify or select: Neural architectures for extractive document summarization. *arXiv preprint arXiv:1611.04244*.

OpenAI. 2023. Gpt-4 technical report.

Ramakanth Pasunuru and Mohit Bansal. 2018. Multi-reward reinforced summarization with saliency and entailment. 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 646–653, New Orleans, Louisiana. Association for Computational Linguistics.

Paul Roit, Ayal Klein, Daniela Stepanov, Jonathan Mamou, Julian Michael, Gabriel Stanovsky, Luke Zettlemoyer, and Ido Dagan. 2020. Controlled crowdsourcing for high-quality QA-SRL annotation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7008–7013, Online. Association for Computational Linguistics.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. 2022. Multitask prompted training enables zero-shot task generalization.

Tal Schuster, Sihao Chen, Senaka Buthpitiya, Alex Fabrikant, and Donald Metzler. 2022. Stretching sentence-pair NLI models to reason over long documents and clusters. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 394–412, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ori Shapira and Ran Levy. 2020. Massive multi-document summarization of product reviews with weak supervision.

Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, Morteza Behrooz, William Ngan, Spencer Poff, Naman Goyal, Arthur Szlam, Y.-Lan Boureau, Melanie Kambadur, and Jason Weston. 2022. BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage. ArXiv:2208.03188 [cs].

Aviv Slobodkin, Avi Caciularu, Eran Hirsch, and Ido Dagan. 2023a. Dont add, dont miss: Effective content preserving generation from pre-selected text spans.

Aviv Slobodkin, Niv Nachum, Shmuel Amar, Ori Shapira, and Ido Dagan. 2023b. Summhelper: Collaborative human-computer summarization.

Aviv Slobodkin, Paul Roit, Eran Hirsch, Ori Ernst, and Ido Dagan. 2022. Controlled text reduction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5699–5715, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Amir Soleimani, Christof Monz, and Marcel Worring. 2023. NonFactS: NonFactual summary generation for factuality evaluation in document summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6405–6419, Toronto, Canada. Association for Computational Linguistics.

Julius Steen and Katja Markert. 2021. How to evaluate a summarizer: Study design and statistical analysis for manual linguistic quality evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1861–1875, Online. Association for Computational Linguistics.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2023. ASQA: Factoid Questions Meet Long-Form Answers. ArXiv:2204.06092 [cs].

Dan Su, Xiaoguang Li, Jindi Zhang, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Read before generate! faithful long form question answering with machine reading.

Jun Suzuki and Masaaki Nagata. 2017. Cutting-off redundant repeating generations for neural abstractive summarization. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 291–297, Valencia, Spain. Association for Computational Linguistics.

Shahbaz Syed, Tariq Yousef, Khalid Al Khatib, Stefan Jänicke, and Martin Potthast. 2021. Summary explorer: Visualizing the state of the art in text summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 185–194, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, YaGuang Li, Hongrae Lee, Huaixiu Steven Zheng, Amin Ghafouri, Marcelo Menegali, Yanping Huang, Maxim Krikun, Dmitry Lepikhin, James Qin, Dehao Chen, Yuanzhong Xu, Zhifeng Chen, Adam Roberts, Maarten Bosma, Vincent Zhao, Yanqi Zhou, Chung-Ching Chang, Igor Krivokon, Will Rusch, Marc Pickett, Pranesh Srinivasan, Laichee Man, Kathleen Meier-Hellstern, Meredith Ringel Morris, Tulsee Doshi, Renelito Delos Santos, Toju Duke, Johnny Soraker, Ben Zevenbergen, Vinodkumar Prabhakaran, Mark Diaz, Ben Hutchinson, Kristen Olson, Alejandra Molina, Erin Hoffman-John, Josh Lee, Lora Aroyo, Ravi Rajakumar, Alena Butryna, Matthew Lamm, Viktoriya Kuzmina, Joe Fenton, Aaron Cohen, Rachel Bernstein, Ray Kurzweil, Blaise Aguera-Arcas, Claire Cui, Marian Croak, Ed Chi, and Quoc Le. 2022. LaMDA: Language Models for Dialog Applications. ArXiv:2201.08239 [cs].

Prasetya Utama, Joshua Bambrick, Nafise Moosavi, and Iryna Gurevych. 2022. Falsesum: Generating document-level NLI examples for recognizing factual inconsistency in summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2763–2776, Seattle, United States. Association for Computational Linguistics.

Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end knowledge-routed relational dialogue system for automatic diagnosis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7346–7353.

Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building task-oriented dialogue systems for online shopping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A large-scale dataset for document-level natural language inference. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4913–4922, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2023. Benchmarking large language models for news summarization.

Qingjuan Zhao, Jianwei Niu, and Xuefeng Liu. 2022. Als-mrs: Incorporating aspect-level sentiment for abstractive multi-review summarization. *Knowledge-Based Systems*, page 109942.

A Annotation Full Guidelines

In this section, we provide the full annotation guidelines, presented to our workers.

A.1 Summary-related Guidelines

As mentioned in §4.2, we guide annotators to segment summary sentences into the different aspects of hotels or businesses. The annotation guidelines distinguish between two classifications of aspects:

• DIFFERENT ASPECTS: This refers to independent facets of the business, e.g., location and room quality.

• DIFFERENT CHARACTERISTICS OF THE SAME ASPECT: This pertains to addressing varied characteristics within the same aspect, for example, the cleanliness and size of a room.

A.2 Review-related Guidelines

This section provides a detailed overview of the review-related guidelines presented to our crowdworkers during their training:

• ANY MENTION OF THE ASPECT: Annotators are trained to align all review mentions of a summary aspect, encompassing both similar and contrasting

sentiments. For instance, if the summary aspect is *"The staff was friendly"*, both positive and negative mentions regarding staff friendliness are to be aligned.

• SPECIFICITY IN REVIEWS: Crowdworkers are advised to align review mentions that offer more specificity than the summary aspects. For example, a general summary statement like *"The staff was helpful"*, should be aligned with a more specific review comment, such as *"the concierge was very helpful"*. We also emphasize that the other way around, namely, that the summary is more specific than the reviews, should not be aligned.

• EXEMPLIFICATION IN REVIEWS: In line with the previous point, annotators are guided to focus on identifying review segments that provide examples of the summary statements. An example would be aligning the summary span *"The hotel is well-maintained"* with a review segment that exemplifies it, such as *"the pool area is very clean"*. As in the previous point, we discourage our crowdworkers from considering the reverse cases, when the summaries exemplify the reviews.

• PARAPHRASING: Annotators are instructed to align paraphrased mentions in reviews with the summary content, such as aligning *"the hotel is overpriced"* with *"you can stay at lovely B&B in the old town that is actually cheaper than this"*.

• CONSECUTIVENESS: We guide our workers to avoid highlighting unnecessary details, i.e., that did not appear in the summary span, and keep the highlights inconsecutive if needed.

• UNALIGNABLE SPANS: Recognizing that each summary is derived from multiple reviews, but reviewers assess only one review at a time, it is often the case that not all summary details will be present in the reviewed content. In such instances, annotators are instructed to leave such summary spans unhighlighted.

B NLI Zero-Shot Prompt

Figure 3 demonstrates the structure of the zero-shot prompt used for the nli-based evaluation frameworks of highlights coverage and faithfulness.

C GPT-4 Prompting

Table 4 presents the faithfulness, coverage and F-1 scores of the zero-shot GPT-4 model across 30 instances from the FiC development set, for varying numbers of in-context examples in the prompt. Based on these outcomes, we chose to proceed with a single in-context example.

Number of Exemplars	Faithfulness	Coverage	F-1
1	**80.1**	85.0	**82.5**
2	72.1	**86.1**	78.5
3	73.6	84.0	78.5
4	72.8	82.2	77.2

Table 4: Faithfulness, coverage and F-1 scores of the zero-shot GPT-4 model on 30 instances for the FiC development set, for varying numbers of in-context examples in the prompt. For each metric, the best scores are in bold.

D Fluency and Redundancy Human Annotation Protocol

We ask crowd-workers to assess the fluency and redundnacy of the texts produced by all models under examination. We employ annotators who have demonstrated proficiency in semantic tasks, including summarization, in previous experiments. For evaluation purposes, 100 instances are randomly selected from our test set, and the texts generated by each model for these instances are evaluated, resulting in 500 total samples. Each sample is reviewed by three different annotators, and their scores are averaged to obtain a final assessment. The evaluation is facilitated through two Amazon Mechanical Turk interfaces, specifically designed for this study. One interface focuses on evaluating coherence, while the other assesses redundancy, with each interface presenting the annotators with one of the 500 samples (as depicted in Figure 4). Consistent with the methodology of (Slobodkin et al., 2022), a 5-point Likert scale is employed to rate the fluency and redundancy of the generated summaries. To minimize ambiguity and promote consistent ratings, each score on the scale is accompanied by explicit criteria (also illustrated in Figure 4). Taking into account an average response time of 30 seconds for each evaluation, we set the compensation for each response at 10 ¢.

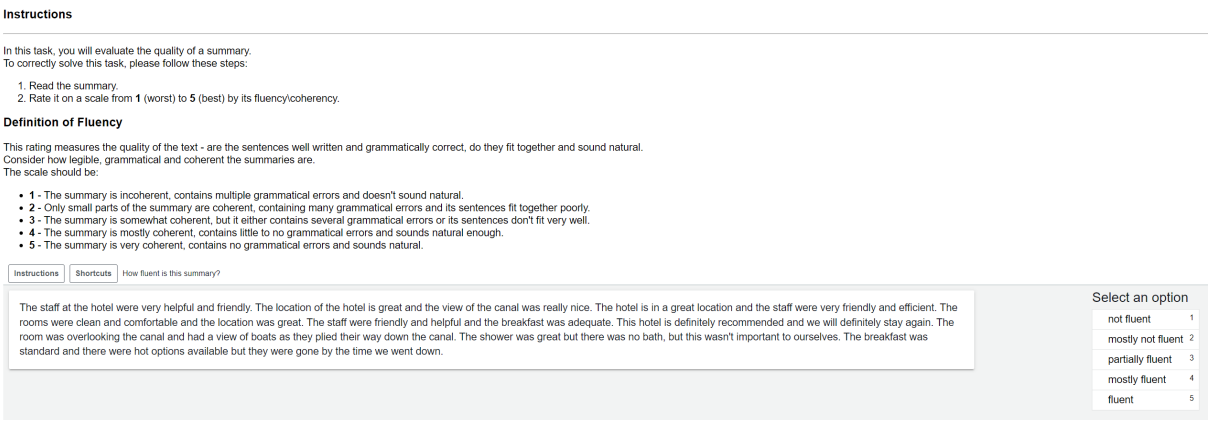
E Additional Evaluation Framework Details

E.1 Trained Faithfulness Metric

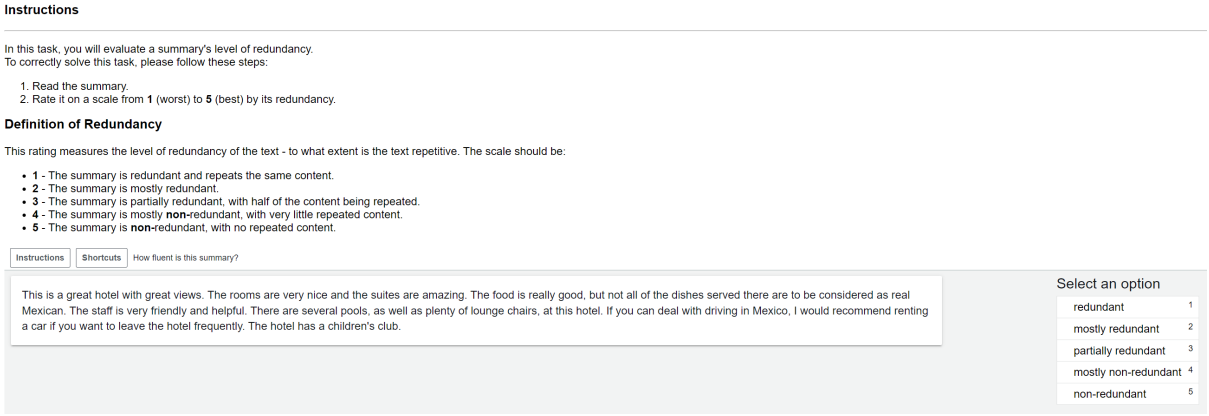
In a similar fashion to the trained coverage metric, we we use our crowdsourced dataset to generate training data for evaluating highlights faithfulness. This approach mirrors the NLI-based metric we proposed, wherein a model is trained to individually evaluate the faithfulness of each output sen-

```
### Instruction: Read the following and determine if the hypothesis can be inferred from
    the premise.
Options: Entailment, Contradiction, or Neutral

### Input:
Premise:  {Premise}
Hypothesis: {Hypothesis}

### Response (choose only one of the options from above):
```

Figure 3: The prompt structure employed in zero-shot configurations as a basis for evaluating the frameworks of faithfulness and coverage.



(a) Fluency Evaluation Interface



(b) Redundancy Evaluation Interface

Figure 4: Example of the data collection interfaces used by the crowd-workers to evaluate the fluency (4a) and redundancy (4b) of summaries.

tence, subsequently averaging the scores across all sentences.

For the positive training instances, we separate each summary from our crowdsourced dataset into sentences, and pair each sentence with all the instance's highlights. In contrast, for the creation of negative instances, we remove all highlights that were aligned with any segment of the corresponding summary sentence. The training process involves fine-tuning a `flan-t5-large` model (Chung et al., 2022). In this setup, the input comprises the highlights and the summary sentence, while the output is either the token 'yes' for positive instances, or 'no' for negative ones. The final score is calculated based on the probability assigned to the token 'yes' by the model.

Judges	Faithfulness	Coverage
1-2	0.37	0.31
2-3	0.71	0.67
1-3	0.39	0.27

Table 5: The individual Cohen's Kappa coefficients for each pair of judges, on the faithfulness and coverage axes.

E.2 NLI-based Coverage Metric

For the evaluation of highlight-coverage using Natural Language Inference (NLI), our approach mirrors the one implemented for assessing faithfulness using NLI (see §5.2), albeit with a role reversal, where the output serves as the premise and the highlights function as the hypothesis. Rather than treating all highlights collectively as the hypothesis, we calculate the coverage of each highlight separately and then average across all highlights.[10]

E.3 Additional Meta Evaluation Setup Details

Pairwise Cohen Kappa Coefficients Table 5 shows the pairwise Cohen's Kappa coefficients for each pair of judges.

Reconciliation Process To achieve further agreement between the three authors, a supplementary reconciliation procedure was undertaken for the ten instances annotated by all three authors. This procedure entailed discussions for each instance where the annotations diverged by more than one point, separately for the faithfulness and coverage scores. During these discussions, each author explained the rationale behind their assigned score. Subsequently, the authors endeavored to reach a unanimous agreement on each instance, thereby further aligned their scoring criteria.

E.4 Additional Meta Evaluation Results

Tables 6 and 7 present the full correlations with human judgments using the Kendall-Tau rank correlations and Spearman's rank correlations, respectively, including the additional evaluation frameworks we explored (see Appendices E.1 and E.2), and the F-1 scores for the ROUGE and BERTScore metrics.

[10]We consider each individual alignment in our crowdsourced dataset as a distinct highlight.

	Faithfulness		Coverage	
Metric	τ	**CI**	τ	**CI**
ROUGE-1 (R)	0.2319	0.23-0.24	0.3467	0.34-0.35
ROUGE-1 (P)	0.5468	0.54-0.55	-0.0533	-0.06–0.05
ROUGE-1 (F1)	0.5587	0.55-0.56	0.1497	0.14-0.16
ROUGE-2 (R)	0.3555	0.35-0.36	0.2731	0.27-0.28
ROUGE-2 (P)	0.5253	0.52-0.53	0.0071	0.00-0.01
ROUGE-2 (F1)	0.4964	0.49-0.50	0.1477	0.14-0.15
ROUGE-L (R)	0.0958	0.09-0.10	0.3835	0.38-0.39
ROUGE-L (P)	0.4898	0.48-0.49	-0.0367	-0.04–0.03
ROUGE-L (F1)	0.3880	0.38-0.39	0.1950	0.19-0.20
METEOR	0.4017	0.40-0.41	0.2736	0.27-0.28
BERTScore (R)	0.2380	0.23-0.24	0.4165	0.41-0.42
BERTScore (P)	0.6004	0.59-0.60	0.0529	0.05-0.06
BERTScore (F1)	0.4958	0.49-0.50	0.2555	0.25-0.26
NLI (Faithfulness)	**0.6745**	**0.67-0.68**	0.0929	0.09-0.10
NLI (Coverage)	0.2255	0.22-0.23	**0.5084**	**0.50-0.51**
Trained (Faithfulness)	0.5836	0.58-0.59	0.2495	0.24-0.25
Trained (Coverage)	0.1771	0.17-0.18	0.4992	0.49-0.50

Table 6: Average Kendall-Tau rank correlations (τ) and their 95% confidence intervals (CI) for tested evaluation metrics against human judgment. Recall-based metrics (R) are more effective for coverage, and precision-based metrics (P) for faithfulness. Best correlations for each axis are in bold.

	Faithfulness		Coverage	
Metric	τ	**CI**	τ	**CI**
ROUGE-1 (R)	0.3124	0.31-0.32	0.4440	0.44, 0.45
ROUGE-1 (P)	0.6892	0.68-0.69	-0.0861	-0.09–0.08
ROUGE-1 (F1)	0.7172	0.71-0.72	0.1654	0.16-0.17
ROUGE-2 (R)	0.4842	0.48-0.49	0.3537	0.35-0.36
ROUGE-2 (P)	0.6807	0.68-0.69	0.0005	-0.01-0.01
ROUGE-2 (F1)	0.6590	0.65-0.66	0.1885	0.18-0.20
ROUGE-L (R)	0.1237	0.12-0.13	0.4902	0.48-0.50
ROUGE-L (P)	0.6420	0.64-0.65	-0.0596	-0.07–0.05
ROUGE-L (F1)	0.5160	0.51-0.52	0.2531	0.25-0.26
METEOR	0.5412	0.54-0.55	0.3487	0.34-0.36
BERTScore (R)	0.3141	0.31-0.32	0.5237	0.52-0.53
BERTScore (P)	0.7450	0.74-0.75	0.0485	0.04-0.06
BERTScore (F1)	0.6516	0.65-0.66	0.3267	0.32-0.33
NLI (Faithfulness)	**0.8257**	**0.82-0.83**	0.1088	0.10-0.12
NLI (Coverage)	0.2831	0.28-0.29	**0.6355**	**0.63-0.64**
Trained (Faithfulness)	0.7268	0.72-0.73	0.3271	0.32-0.33
Trained (Coverage)	0.2315	0.22-0.24	0.6178	0.61-0.62

Table 7: Average Spearman's rank correlations (τ) and their 95% confidence intervals (CI) for tested evaluation metrics against human judgment. Recall-based metrics (R) are more effective for coverage, and precision-based metrics (P) for faithfulness. Best correlations for each axis are in bold.

F Additional Dataset Details

F.1 Full FiC Dataset Statistics

Table 8 presents the full FiC dataset statistics, including specific statistics for each of the dataset's splits and instances origin, i.e., CocoTrip or FewSum.

	#unique sets of reviews	#summaries/ review-set (average)	#summary-review-set pairs	mean review/summary size (tkns)	max review/review-set/ summary (tkns)	mean review/summary size (sents)	summary sents aligning to multiple reviews	summary sents aligning to multiple review sents
Train								
CocoTrip	184	2.63	484	97.56/80.15	239/1118/231	6.22/5.32	80.74%	50.02%
FewSum	53	3.00	159	57.27/60.06	75/497/104	4.87/4.34	89.13%	65.07%
Total	237	2.71	643	87.6/75.18	239/1118/231	5.89/5.08	82.51%	53.20%
Dev								
CocoTrip	10	6.00	60	91.24/75.93	197/829/174	5.64/5.07	85.53%	46.05%
FewSum	13	3.00	39	57.55/58.46	78/493/102	5.21/4.15	90.74%	79.63%
Total	23	4.30	99	77.97/69.05	197/829/174	5.47/4.71	87.34%	57.73%
Test								
CocoTrip	26	6.00	156	90.34/74.28	279/881/266	5.64/4.86	79.95%	42.74%
FewSum	34	3.00	102	57.43/57.44	74/509/105	5.0/4.41	88.89%	67.11%
Total	60	4.30	258	77.33/67.62	279/881/266	5.39/4.68	83.28%	51.82%
Overall								
CocoTrip	220	3.18	700	95.41/78.48	279/1118/266	6.04/5.2	80.97%	48.17%
FewSum	100	3.00	300	57.36/58.96	78/509/105	4.96/4.34	89.25%	67.59%
Total	320	3.13	1000	83.99/72.62	279/1118/266	5.72/4.94	83.15%	53.29%

Table 8: Statistics of our dataset. The two right-most columns indicate the percentage of summary sentences that align with spans in more than one review, and the percentage of summary sentences that align with a non-continuous span from across more than one review sentence in one of its reviews.

F.2 Annotation Cost

Each annotation instance, averaging 4 minutes, is priced at 70¢. Annotators also receive compensation for training activities, including a 5$ bonus for taking the 25-minute tutorial and an additional 2$ for reviewing feedback. The total cost for the dataset amounted to approximately 5700$.

F.3 Additional Details about the Annotators Recruitment

For our crowdsourcing project, we hired annotators from English-speaking countries who had over 5000 approved HITs as well as an approval rate higher than 98% on amazon Mechanical Turk. During the recruitment process, in addition to explaining the annotation guidelines, we also explained to the crowdworkers the purpose of the dataset, in order to rationalize different aspects of the annotation protocol.

G IAA disagreement Examples

Figure 5 demonstrates two instances of disagreements between our annotators.

H Additional Experimental Details

To incorporate the highlighting signal in the baseline Flan-T5$_{\text{H}}$, <*extra_token_1*> and <*extra_token_2*> tokens were added to the input, before and after each highlight. For all trained models, we set the maximum input length to 2048, to accommodate the input length of the language model.

We also set the maximum target length to 200, which we found works best, as well as setting the batch size to 1. The other parameters are similar to Slobodkin et al. (2022, 2023a). The model is trained for 10k steps. Training is performed on a two A100-SXM4-80GB GPUs, and costs about 12 GPU hours for the supervised models (Flan-T5$_{\text{H}}$, Flan-T5$_{\text{no-H}}$, and Flan-T5$_{\text{only-H}}$) and about 36 GPU hours for the RL-tuned variant of Flan-T5$_{\text{H}}$.

Additionally, to train the trained faithfulness and coverage evaluators, we concatenate the highlights concatenation and the output's sentence (for faithfulness) and the generated output with each of the highlights (for coverage), and use the special token <*extra_token_4*> as a delimiter. For both evaluators, we set the maximum input length to 1024, the maximum target length to 4 , and the batch size to 1. We train the models for 10 epochs. Training is performed on a single A100-SXM4-80GB GPU, and costs about 4 GPU hours.

Overall, our trained models, both for faithfulness and coverage evaluation and for the `FiC` task, use `flan-t5-large` as their backbone model, which consists of 780 million parameters, and our zero-shot NLI-based evaluation frameworks use `flan-t5-xxl` as the backbone model, consisting of 11 billion parameters.

I List of Data and Software Licenses Employed in this Paper

Our framework dependencies are:

1. CocoTrip dataset: `https://github.`

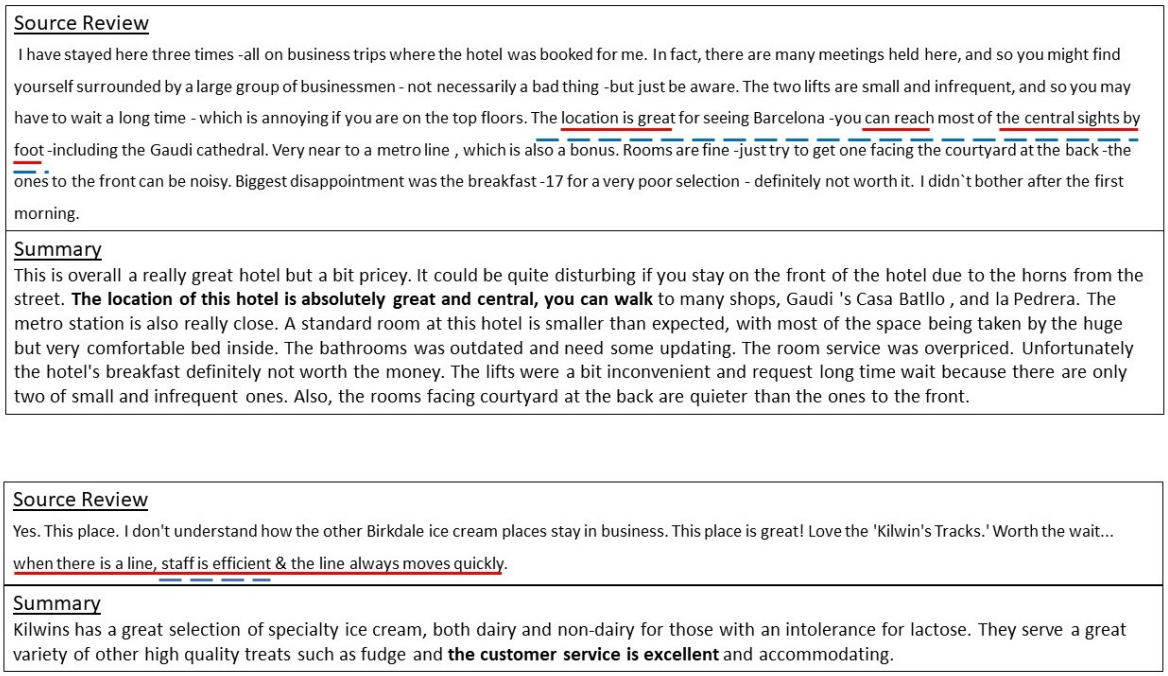


Figure 5: Two examples of disagreement between annotators. For each example, the bottom part is the summary (with the summary span over which there was disagreement in bold) and the top part is a review with both the annotators' highlights (marked with a red solid line and a blue dashed line to indicate each highlight).

com/megagonlabs/cocosum/blob/main/LICENSE, under an Apache License 2.0.

2. FewSum dataset: https://github.com/abrazinskas/FewSum/blob/master/LICENSE.txt, under the MIT License.

3. Quark: https://github.com/GXimingLu/Quark, Misc.

4. Baseline model for the zero-shot NLI-based evaluation frameworks: https://huggingface.co/google/flan-t5-xxl/tree/main, under an Apache License 2.0.

5. Baseline model for the trained evaluation frameworks and models: https://huggingface.co/google/flan-t5-large/tree/main, under an Apache License 2.0.