Rethinking Scientific Abstract Summary Evaluation: Grounding Metrics on Facet-aware Benchmark

Xiuying Chen[1], Tairan Wang[1], Qingqing Zhu[2], Taicheng Guo[3], Shen Gao[4], Zhiyong Lu[2], Xin Gao[1], Xiangliang Zhang[1,3]
[1]King Abdullah University of Science & Technology [2] National Institutes of Health
[3] University of Notre Dame [4] University of Electronic Science and Technology of China

A FD dataset release

A.1 Accessing FD

The dataset source files are stored in JSON format, and they are uploaded to GitHub and can be downloaded publicly. A case in MDS consists of article, summary, and annotations. All annotations, code, and datasets are available at `https://github.com/iriscxy/FD`.

A.2 FD Distribution and Maintenance

License. FD is distributed under the Creative Commons (CC) copyright licenses. It is important to note that the source documents used in the dataset are already in the public domain, thereby respecting copyright regulations. To address any concerns related to personal or copyrighted content within FD, we have implemented a contact form on our website. This form serves as a channel for users to submit requests for the removal or blacklisting of specific links or content that may infringe upon personal rights or copyrights. We are committed to promptly and diligently processing these requests to maintain the integrity and legality of the dataset. The authors bear all responsibility in case of violation of rights and confirm the dataset licenses.

Maintenance. The authors are committed to providing long-term support for the FD dataset. At present, FD files are hosted on GitHub, allowing for easy access and collaboration. Additionally, the authors are committed to actively monitoring the usage of the dataset and addressing any issues that may arise. This includes promptly addressing bug fixes, resolving technical concerns, and providing necessary updates to ensure the dataset remains reliable and useful to the research community.

B Facet-aware Metric

B.1 Prompts

In our facet-aware metric, we employ GPT-4 to initially extract information of different aspects within the abstract. The prompt we use is:

> What is the background/method/result/conclus-
> ion of this work? Extract the segment of the input as the answer. Return the answer in JSON format, where the key is background/method/result/conclusion.
> If any category is not represented in the input, its value should be left empty.

The evaluation prompt for different facets with in-context samples are shown in Fig. 1 and Fig. 2.

Submitted to the 38th Conference on Neural Information Processing Systems (NeurIPS 2024) Track on Datasets and Benchmarks. Do not distribute.

Using a less strict criterion, assess the alignment (1-3) between the two inputs.

- 3: Input2 is generally consistent with Input1.
- 2: Input1 is not mentioned in Input2.
- 1: Input2 contradicts Input1.

Only return the number.

Example 1:
Input1: the use of 2-[18f]fluoro-2-deoxy - d - glucose ([18f]fdg) may help to establish the antitumor activity of enzastaurin , a novel protein kinase c - beta ii (pkc-ii) inhibitor , in mouse xenografts .
Input2: Imaging techniques, such as positron emission tomography (PET), are important for diagnosing and monitoring cancer patients. The glucose analogue 2-[F]fluoro-2-deoxy-D-glucose (FDG) is commonly used as a tracer in PET imaging to assess tissue glucose utilization. FDG PET is widely used in diagnosing various types of cancer, and it is being evaluated as a tool to assess the effects of anticancer drugs. Enzastaurin is a novel compound that inhibits protein kinase C-beta (PKC-), which has been implicated in tumor growth.
Number: 3

Example 2:
Input1: nissen fundoplication is an effective treatment of gastroesophageal reflux in infants . laparoscopic procedures after previous laparotomy are technically more challenging . the role of laparoscopic nissen fundoplication after neonatal laparotomy for diseases unrelated to reflux is poorly described.
Input2: The article discusses the complex nature of gastroesophageal reflux in neonates and infants, which is often caused by a combination of developmental and anatomical factors.
Number: 2

Example 3:
Input1: [18f]fdg pet imaging technique does not correlate with standard caliper assessments in xenografts to assess the antitumor activity of enzastaurin .
Input2: These findings suggest that [18F]FDG PET imaging is a useful tool for assessing the antitumor effects of novel compounds, such as enzastaurin, in preclinical studies.
Number: 1

Figure 1: Few-shot prompt for background/conclusion evaluation.

Assess the alignment (1-4) between the two inputs.

- 4: Input2 generally covers the information present in Input1, or omits minor details from Input1.
- 3: Input2 omits important information from Input1.
- 2: Input1 is not mentioned in Input2.
- 1: Input2 contradicts Input1.

Only return the number.

Example 1:
Input1: We analyzed the methylation status of protocadherin8 in 162 prostate cancer tissues and 47 benign prostatic hyperplasia tissues using methylation-specific PCR (MSP). The patients with prostate cancer were followed up for 15-60 months, and biochemical recurrence was defined as the period between radical prostatectomy and the measurement of 2 successive values of serum PSA level 0.2 ng/ml.
Input2: the promoter methylation status of protocadherin8 in 162 prostate cancer tissues and 47 normal prostate tissues was examined using methylation - specific pcr (msp) . subsequently , the relationships between protocadherin8 methylation and clinicopathological features of prostate cancer patients and biochemical recurrence - free survival of patients were analyzed.
Number: 4

Example 2:
Input1: the present study included 515 patients admitted to the coronary care units or equivalent cardiology wards of the participating hospitals between 2011 and 2012 in north punjab , pakistan . the analysis was focused on identifying the socioeconomic status , lifestyle , family history of mi , and risk factors (i.e. hypertension , diabetes , smoking , and hyperlipidemia) . a structured questionnaire was designed to collect data . the lipid profile was recorded from the investigation chart of every patient . for statistical analysis , the kruskal wallis , mann - whitney u , wilcoxon , and chi - square tests were used.
Input2: a population - based cross - sectional study was conducted in six regions in north punjab (urban and rural patients). data were collected using trained trained staff from the patients admitted in coronary care units or equivalent cardiology hospitals in the participating hospitals.
Number: 3

Example 3:
Input1: hyperglycemia , commencing on the first dose of the steroid given , persisted even after the discontinuation of steroids and improvement of other signs . there were no signs of pancreatitis or type 1 diabetes clinically in laboratory tests . her blood glucose levels were regulated at first with insulin and later with metformin . within 1 year of follow - up , still regulated with oral antidiabetics , she has been diagnosed with type 2 diabetes .
Input2: The patient was treated with discontinuation of carbamazepine, antihistaminic and systemic steroids, and her hyperglycemia resolved with metformin treatment. The patient's lung, skin, liver, and renal findings regressed, and a patch test with carbamazepine was positive.
Number: 2

Example 4:
Input1: hyperglycemia , commencing on the first dose of the steroid given , persisted even after the discontinuation of steroids and improvement of other signs . there were no signs of pancreatitis or type 1 diabetes clinically in laboratory tests . within 1 year of follow - up , still regulated with oral antidiabetics , she has been diagnosed with type 2 diabetes .
Input2: the patient recovered without any sequelae.
Number: 1

Figure 2: Few-shot prompt for method/result evaluation.

B.2 Facet Information Extraction Evaluation

We conducted a human evaluation to assess GPT-4's performance in the facet information extraction task as shown Fig. 3. Generally, GPT-4 exhibits solid performance in extraction tasks, achieving 90% accuracy. However, it does make errors, such as mixing different aspects, omitting certain aspects, and making up information that isn't present in the input. This issue of generating non-existent information, often referred to as hallucination, is a common phenomenon in LLMs. We are optimistic about the development of more refined LLMs in the future to address these challenges.

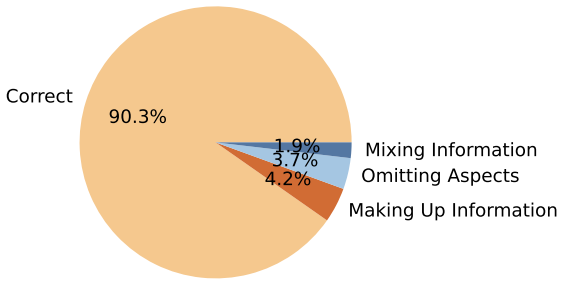

Figure 3: Human evaluation of GPT-4's facet information extraction.

C Comparison of Summarization Systems on arXiv

We show the performance of various summarization systems in different metrics on arXiv in Tab. 1, and the Spearman correlations among metrics within our FM paradigm and existing evaluation metrics on arXiv in Fig. 4.

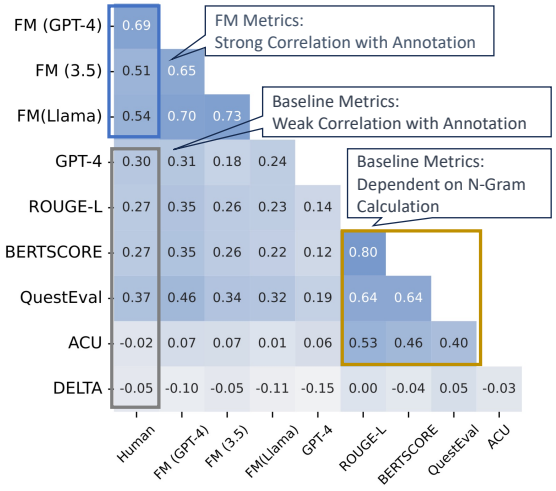

Figure 4: Spearman correlations among metrics within our FM paradigm, LLM-based baseline (GPT-4), and existing evaluation metrics (ROUGE-L, etc) on arXiv dataset.

Model	ROUGE-L	BERTScore	DELTA	QuestEval	ACU	FM(GPT-3.5)	FM(GPT-4)	Human
GPT-3.5	0.2023	0.8337	0.2553	0.1322	0.0552	0.6195	0.6092	0.6385
Llama2	**0.2338**	0.8367	**0.2593**	**0.1742**	0.0000	**0.6621**	**0.6915**	**0.7155**
FactSum	**0.3089**	**0.8664**	**0.3209**	**0.2129**	**0.1623**	**0.6536**	**0.6863**	**0.6843**
BART-Large	0.2270	**0.8495**	0.2597	0.1494	**0.0907**	0.5912	0.5785	0.6231

Table 1: Performance of various summarization systems in different metrics on arXiv dataset. **Bold** indicates the best result, while **bold** denote the second best. Generally, all metrics favor Llama2 and FactSum. Meanwhile, metrics adopting our facet-aware paradigm, including FM(GPT-3.5), FM(GPT-4), and Human, deviate from the existing baselines, often awarding higher scores to Llama2, providing a different perspective on model evaluation.

D Performance in Different Facets

Beyond the overall evaluation, we show the human evaluation of the models' performance in various aspects of abstract writing in Fig. 6. Firstly, all models show higher performance in the background

aspect, as it often involves just a broad, less detailed overview. In contrast, other aspects demand more precise alignment with the input, leading to generally lower model performance in these areas. Among these three aspects, Llama2 consistently exhibits relatively higher performance. In comparison, GPT-3.5 and other PLMs exhibit weaker performance, especially in formulating conclusions. Specifically, in scenarios where the work's conclusion deviates from conventional results mentioned in the background, *GPT-3.5 can adhere to the conventions instead of being faithful to the conclusion in the input.* This could be because GPT-3.5 relies more on its internal knowledge base, without thoroughly analyzing the input content. We additionally have a statistical analysis that reveals 34.7% of weak performance cases (where the conclusion score is below 3) in PLM models are due to fluency issues in longer text generation in the last conclusion part.

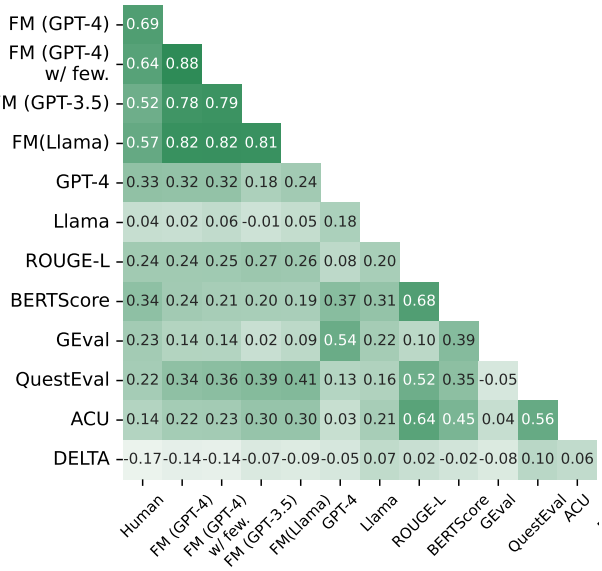

Figure 5: Spearman correlations among metrics within our FM paradigm, LLM-based baseline (GPT-4), and existing evaluation metrics (full version).

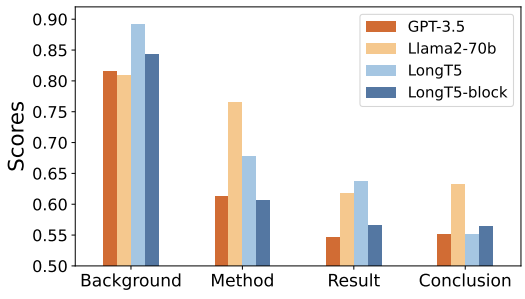

Figure 6: Model performance across four facets.

Background
Method
Result
Conclusion

Complete bone marrow infiltration with profound pancytopenia is very uncommon in breast cancer. Bone marrow metastasis can frequently occur following development of metastatic breast cancer. However, bone marrow failure as the herald of this disease is not typically seen. Very limited data exists as to the safest and most efficacious manner to treat patients with profound pancytopenia due to metastatic solid tumor involvement. In this case, the patient's thrombocytopenia was particularly worrisome, requiring daily platelet transfusions. There was also concern that cytotoxic chemotherapy would exacerbate the patient's thrombocytopenia and increase bleeding risk. The patient's dramatic response to chemotherapy with full platelet recovery is also highly unusual. For our patient, continuous doxorubicin successfully "unpacked" the bone marrow despite a low baseline platelet level, and without increasing the need for more frequent platelet transfusion or risk of catastrophic bleeding. Given the rarity of this presentation, it is currently unknown if the majority of similar patients experience near full recovery of hematopoietic function after initiation of appropriate systemic treatment for metastatic disease.

Evaluating multi-document summarization (MDS) quality is difficult. This is especially true in the case of MDS for biomedical literature reviews, where models must synthesize contradicting evidence reported across different documents. Prior work has shown that rather than performing the task, models may exploit shortcuts that are difficult to detect using stan- dard n-gram similarity metrics such as ROUGE. Better automated evaluation metrics are needed, but few resources exist to assess metrics when they are proposed. Therefore, we introduce a dataset of human-assessed summary quality facets and pairwise preferences to encourage and support the development of better auto- mated evaluation methods for literature review MDS. We take advantage of community sub- missions to the Multi-document Summarization for Literature Review (MSLR) shared task to compile a diverse and representative sam- ple of generated summaries. We analyze how automated summarization evaluation metrics correlate with lexical features of generated summaries, to other automated metrics including several we propose in this work, and to aspects of human-assessed summary quality. We find that not only do automated metrics fail to cap- ture aspects of quality as assessed by humans, in many cases the system rankings produced by these metrics are anti-correlated with rankings according to human annotators.

Figure 7: Highlight visualization for reading summaries during the question-answering task.

Generation	Reference	Score & Error Analysis
The results show that FDG uptake estimates can accurately characterize the antitumor activity of enzastaurin.	[18f] FDG pet imaging technique does not correlate with standard caliper assessments in xenografts to assess the antitumor activity of enzastaurin.	1 Contrary
33 giant pulses having peak flux densities between @xmath0 jy and @xmath1 jy were detected .	The results of the study, including pulse amplitude and broadening statistics, are summarized.	3 missing key information
We propose a new data-driven sparse-to-dense interpolation algorithm based on a fully convolutional network. We introduce lateral dependencies...	We propose, for the first time, a neural network based sparse - to - dense interpolation for optical flow.	3 missing key information

Figure 8: Case study across two datasets of our FM (GPT-4).