# One Prompt To Rule Them All: LLMs for Opinion Summary Evaluation

**Anonymous ACL submission**

## Abstract

Evaluation of opinion summaries using conventional reference-based metrics rarely provides a holistic evaluation and has been shown to have a relatively low correlation with human judgments. Recent studies suggest using Large Language Models (LLMs) as reference-free metrics for NLG evaluation, however, they remain unexplored for opinion summary evaluation. Moreover, limited opinion summary evaluation datasets inhibit progress. To address this, we release the SUMMEVAL-OP dataset covering 7 dimensions related to the evaluation of opinion summaries: `fluency`, `coherence`, `relevance`, `faithfulness`, `aspect coverage`, `sentiment consistency`, and `specificity`. We investigate OP-I-PROMPT, a dimension-independent prompt, and OP-PROMPTS, a dimension-dependent set of prompts for opinion summary evaluation. Experiments indicate that OP-I-PROMPT emerges as a good alternative for evaluating opinion summaries achieving an average Spearman correlation of **0.70** with humans, outperforming all previous approaches. *To the best of our knowledge,* we are the first to investigate LLMs as evaluators on both closed-source and open-source models in the opinion summarization domain.

## 1 Introduction

Opinion summarization systems predominantly use traditional metrics such as ROUGE (Lin, 2004) and BERTSCORE (Zhang et al., 2019) for automatic evaluation, however, they have been shown to have poor correlations with human judgments (Shen and Wan, 2023). Moreover, these metrics fall short of comprehensively evaluating opinion summaries. Additionally, obtaining reference-based datasets at a large scale is an expensive process.

Recently, Large Language Models (LLMs) have been utilized as reference-free evaluators for Natural Language Generation (NLG) outputs (Fu et al., 2023; Chiang and Lee, 2023a,b; Wang et al., 2023; Liu et al., 2023). The idea is to prompt a powerful LLM such as `ChatGPT-3.5`/`GPT-4` to evaluate an output on certain criteria. However, their suitability has not been explored at all for evaluating opinion summaries. Moreover, these approaches have been tested only on closed-source models (`ChatGPT-3.5`/`GPT-4`) primarily because of the limitations of the open-source models in following instructions and producing the desired output (Chiang and Lee, 2023b).

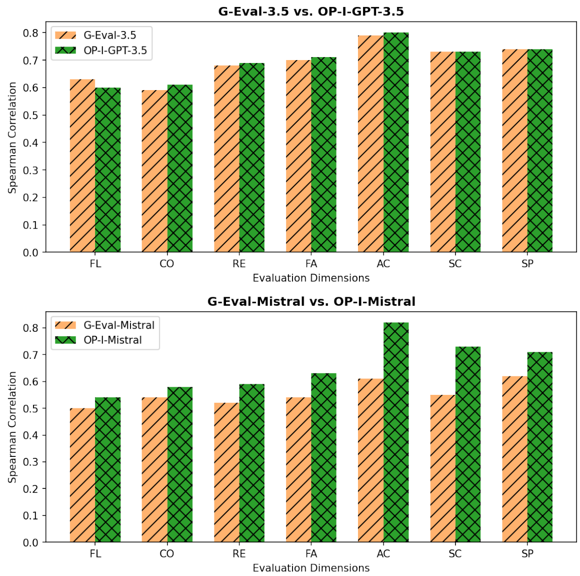


**Figure 1: G-EVAL vs. OP-I-PROMPT.** On closed-source model (`ChatGPT-3.5`) our OP-I-PROMPT shows comparable performance whereas on open-source model (`Mistral-7B`) our approach outperforms G-EVAL on 7 dimensions: `fluency` (FA), `coherence` (CO), `relevance` (RE), `faithfulness` (FA), `aspect coverage` (AC), `sentiment consistency` (SC), and `specificity` (SP). Check Figure 4 for more details.

To this end, we first create SUMMEVAL-OP, a reference-free opinion summarization dataset covering 7 dimensions, for the e-commerce domain. Next, we present OP-I-PROMPT and OP-PROMPTS tailored for opinion summary evaluation. We investigate their suitability to both closed-source and open-source models. Experiments reveal that Op-I-Prompt emerges as a good alternative for evaluating opinion summaries across all 7 dimensions.

Our contributions are:

1. **SummEval-Op**[1], an opinion summarization benchmark dataset, consisting of a total of 2,912 summary annotations, assessing 13 opinion summaries for 32 products from the Amazon test set. The evaluation covers **7 dimensions**- `fluency`, `coherence`, `relevance`, `faithfulness`, `aspect coverage`, `sentiment consistency`, and `specificity` related to the evaluation of opinion summaries (Section 4).

2. **Op-I-Prompt**, a dimension-independent prompt and **Op-Prompts**, a dimension-dependent set of prompts, enabling opinion summary evaluation for all the 7 dimensions. Experiments indicate that the Op-I-Prompt generally outperforms existing approaches on both closed-source and open-source models by **9**% on average in correlation with human judgments (Figure 1, Section 3). To the *best of our knowledge* we are the first to test the applicability of different prompt approaches on open-source LLMs.

3. Benchmarking of recent LLMs (closed and open-source) on the aforementioned 7 dimensions for the task of opinion summarization, which to the *best of our knowledge* is first of its kind (Section 6).

4. Detailed analysis, comparing an open-source LLM against a closed-source LLM acting as evaluators for automatic evaluation of opinion summaries on 7 dimensions. Analysis indicates that Op-I-Prompt emerges as a good alternative for evaluating opinion summaries showing a high correlation with humans when compared with alternatives (Section 6).

## 2 Related Work

**LLM-based Evaluators** Fu et al. (2023) introduced GPTScore that operates on the premise that a generative pre-training model (e.g. GPT-3) is likely to assign a higher probability to the generation of high-quality text in line with provided instructions and context. Chiang and Lee (2023a) were the first to explore LLMs for evaluation. Chiang and Lee (2023b) provide concrete guidelines that improve ChatGPT's correlation with humans. Wang et al. (2023) conducted an initial survey exploring the utilization of ChatGPT as an NLG evaluator. Kocmi and Federmann (2023) used GPT models for evaluating machine learning tasks. Liu et al. (2023) introduced G-Eval, a framework for evaluation of NLG outputs using *Chain of Thought* (CoT) (Wei et al., 2023) and assigning weights to a predetermined set of integer scores based on their generation probabilities from GPT-3/4. Chen et al. (2023) were the first to investigate approaches to reference-free NLG evaluation using LLMs, finding that an explicit score generated by ChatGPT is the most effective and stable approach. Zheng et al. (2023) show that strong LLMs such as GPT-4 achieve a similar level of agreement to that of humans and hence can be used to approximate human preferences. Our work investigates two prompt strategies and tests the applicability of different prompt approaches on closed-source and open-source LLMs for opinion summary evaluation for 7 dimensions.

**Opinion Summary Evaluation Benchmark** (Shen and Wan, 2023) created the OpinSumEval dataset, utilizing the Yelp test set (Chu and Liu, 2019), annotating for 4 dimensions relevant to opinion summary evaluation. Our work enhances this effort by introducing SummEval-OP, which focuses on the e-commerce domain, constructed using the Amazon test set (Bražinskas et al., 2020). Additionally, we collect annotations for 7 dimensions on the recent LLM summaries, subsequently establishing benchmarks for comparison.

## 3 Methodology

We describe our dimension independent and dependent prompts and the model scoring function.

### 3.1 Prompt Approaches

Figure 2 shows the different prompt approaches for evaluating opinion summaries. In general, the prompts include the following 3 components-

**`Task Description`**: Defines the task that the LLM will be performing. In our case, the task is to evaluate a summary corresponding to a set of reviews on a given metric/dimension.

**`Evaluation Criteria`**: Defines the criteria that will be used to perform the task. In our case, the

[1]Dataset, code, and prompts will be released publicly

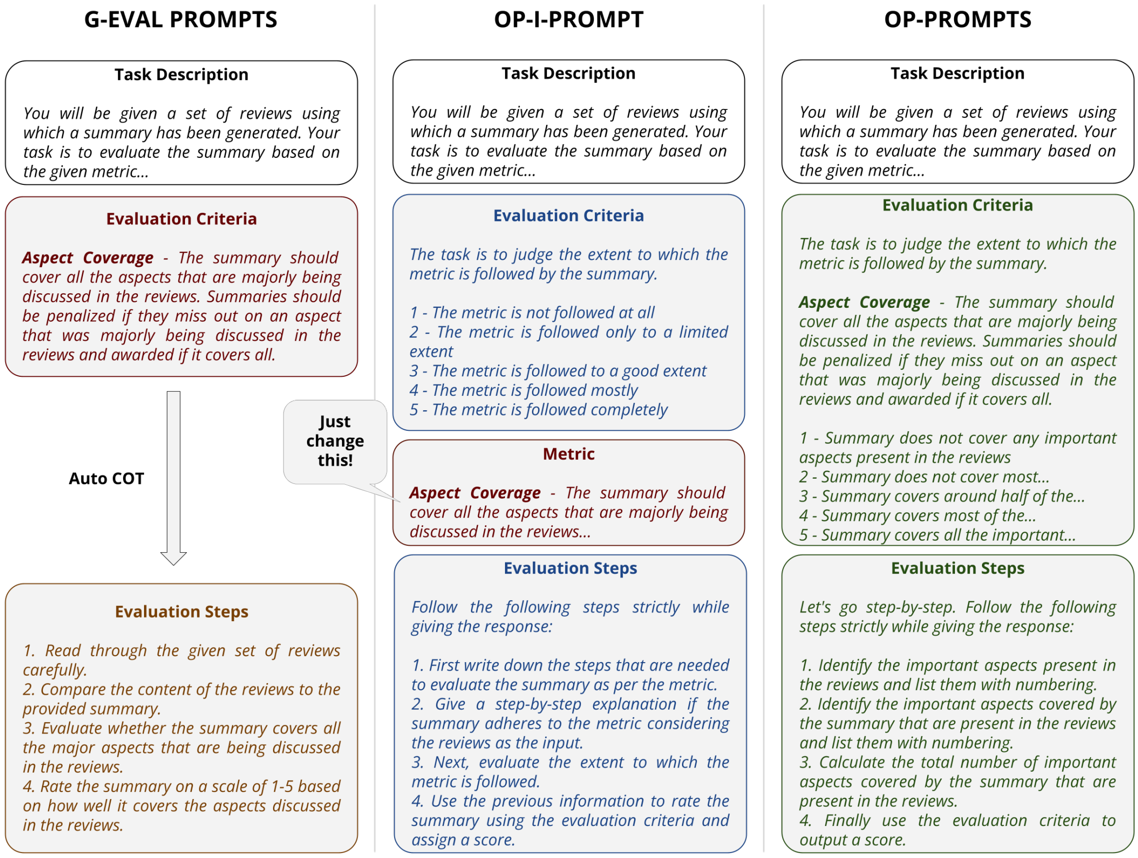


**Figure 2: Comparison of Prompt Approaches.** G-EVAL PROMPTS first generates the `Evaluation Steps` using `Task Description` and `Evaluation Criteria` in Chain-of-Thought fashion. Finally the full prompt is used to evaluate the opinion summaries. In contrast, our **OP-I-PROMPT** is simpler and has `Task Description`, `Evaluation Criteria`, and `Evaluation Steps` fixed for a dimension/metric independent evaluation. Here, only the `Metric` part needs to be changed for evaluating any dimension/metric. Finally **OP-PROMPTS** are dimension/metric dependent prompts that needs to be specifically crafted for each dimension/metric.

task being opinion summary evaluation, the criteria is to assign a score $(1-5)$ for a certain metric/dimension depending on the extent to which the summary adheres to it.

**`Evaluation Steps`**: This comprises the steps that the LLM must take to correctly perform the described task. In our case, it contains the steps that the LLM should follow to evaluate a certain metric/dimension.

We propose two prompt approaches for evaluating opinion summaries:

**OP-I-PROMPT** is a metric-independent opinion summary evaluation prompt. Here we split the `Evaluation Criteria` to create a new component `Metric` consisting only the evaluation dimension. All the remaining components i.e. `Task Description`, `Evaluation Criteria`, and `Evaluation Steps` are crafted in such a way that they are applicable in general to any opinion summary evaluation dimension. This benefits us in the following way: (*a*) we have a metric independent prompt that can now evaluate any metric/dimension just by replacing with the desired definition of the dimension within the `Metric` block (*b*) the remaining components, crafted specifically keeping the task in mind, ensures that the evaluation by LLM takes place as defined by us.

**OP-PROMPTS** is a set of metric-dependent prompts. We specifically handcrafted these prompts for each of the 7 evaluation dimensions. Although this ensures that the evaluation happens exactly in the way we define, this requires a certain level of expertise in the evaluation domain and prompting. This could be seen as a much stricter version of the prompt compared to OP-I-PROMPT where the prompt is suited to any evaluation dimension which is not the case here. A prompt defined for a certain dimension could not be utilized for any other dimension.

In contrast, G-EVAL (Liu et al., 2023) used auto chain-of-thoughts (Wei et al., 2022) by using `Task Description` and `Evaluation Criteria` to automatically generate the `Evaluation Steps`. Finally, all the components together constitute the G-EVAL prompt that is used by an LLM to evaluate summaries. Our work investigates the applicability of all these prompts to both closed-source and open-source models for evaluating opinion summaries.

### 3.2 Scoring Function

Liu et al. (2023) pointed out the limitation of LLM outputting an integer score and proposed using a weighted average of the scores as the LLMs output, where the weights are the probabilities of the corresponding score. Formally, say, the scoring is scheme is from $\{s_1, ..., s_j\}$, the probability of each score $p(s_k)$ is calculated by an LLM and the final score $o$ is computed as:

$$o = \sum_{k=1}^{j} p(s_k) \times s_k \tag{1}$$

$p(s_k)$ for an input $k$ is estimated through an LLM by sampling $\mathsf{n}$ outputs. In which case, the scoring function just translates to taking a mean over the $\mathsf{n}$ outputs. We ensure that $\mathsf{n}$ is large ($\sim 100$) to get a reliable estimate of the probabilities.

## 4 SUMMEVAL-OP Benchmark Dataset

We created the **SUMMEVAL-OP** benchmark dataset for evaluating the opinion summaries on 7 dimensions. In this section, we discuss the dataset used, opinion summary evaluation metrics, annotation details, and its analysis.

### 4.1 Dataset

We utilized the Amazon test set (He and McAuley, 2016; Bražinskas et al., 2020), comprising of reviews from 4 domains: *electronics, home & kitchen, personal care,* and *clothing, shoes & jewelry*. The test set contained a total of 32 products, each with 3 human-annotated reference summaries and 8 reviews per product. For our use, we needed one human reference summary per product which we obtained by randomly selecting one of the summaries out of the 3 for each product. We do not directly consider only one of the human summaries as this would bias the summaries to a single person.

### 4.2 Opinion Summarization Metrics

The evaluation of opinion summaries focused on the following 7 dimensions:

1. **`fluency` (FL)**- The quality of summary in terms of grammar, spelling, punctuation, capitalization, word choice, and sentence structure and should contain no errors. The summary should be easy to read, follow, comprehend and should contain no errors. Annotators received specific guidelines on how to penalize summaries based on fluency levels.

2. **`coherence` (CO)**- The collective quality of all sentences. The summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to a coherent body of information.

3. **`relevance` (RE)**- The summary should not contain opinions that are either not consensus or important. The summary should include only important opinions from the reviews. Annotators were instructed to penalize summaries if they contained redundancies and excess/unimportant information.

4. **`faithfulness` (FA)**- Every piece of information mentioned in the summary should be verifiable/supported/inferred from the reviews only. Summaries should be penalized if any piece of information is not verifiable/supported/inferred from the reviews or if the summary overgeneralizes something.

5. **`aspect coverage` (AC)**- The summary should cover all the aspects that are majorly being discussed in the reviews. Summaries should be penalized if they miss out on an aspect that was majorly being discussed in the reviews and awarded if it covers all.

6. **`sentiment consistency` (SC)**- All the aspects being discussed in the summary should accurately reflect the consensus sentiment of the corresponding aspects from the reviews. Summaries should be penalized if they do not cover accurately the sentiment regarding any aspect within the summary.

7. **`specificity` (SP)**- The summary should avoid containing generic opinions. All the opinions within the summary should contain detailed and specific information about the consensus opinions. Summaries should be penalized for missing out details and should be awarded if they are specific.

### 4.3 Annotation Details

For creating the **SUMMEVAL-OP** dataset, annotations were collected for a total of 13 summaries per product across 7 dimensions for 32 products from the Amazon test set. The 13 summaries comprised of 1 human-annotated reference summary (as mentioned in Section 4.1) and 12 different model-generated summaries (listed in Section 5.1). To ensure high quality of annotations, each summary was annotated by 3 raters for 7 dimensions, amounting to $2,912$ summary-level ratings. Raters were asked to rate the summaries on a Likert scale from 1 to 5 (higher is better) along the 7 dimensions. Each summary for each dimension was rated by 3 raters. The overall quantity of annotations is: 3 (*# of raters*) x 32 (*# of instances*) x 13 (*# of summaries*) x 7 (*# of dimensions*) = $8,736$ ratings.

We chose to hire 3 Masters' students with experience in opinion summarization as opposed to crowd workers for the following reasons: *(a)* Gillick and Liu (2010) demonstrated that summary evaluations from non-experts can significantly diverge from expert annotations and may display inferior inter-annotator agreement, and *(b)* Fabbri et al. (2021) emphasized the significance of employing expert annotators to mitigate quality concerns in human ratings. Similar to Fabbri et al. (2021), we conducted the process in two rounds, to ensure high-quality ratings. In Round II, ratings where the scores of any rater differed from any other rater by 2 or more points were re-evaluated. The re-evaluation was done through a discussion between the annotators such that ratings from all 3 differ by at most 1. We asked the raters to be critical and discuss the ratings during re-evaluation. Check **Appendix** B

### 4.4 Annotation Analysis

We evaluated the inter-rater agreement for the 3 raters using Krippendorff's alpha coefficient ($\alpha$) (Krippendorff, 2011). For Round-I, we found the coefficient to be 0.50 indicating *moderate agreement* ($0.41 \leq \alpha \leq 0.60$). For Round-II, the coefficient increased to 0.80, indicating *substantial agreement* ($0.61 \leq \alpha \leq 0.80$). We report the dimension-wise agreement scores for both rounds in Table 1. We observe that for both Round-I and Round-II, `faithfulness` and `aspect coverage` score higher than others. This is mostly because `faithfulness` and `aspect coverage` could be identified by cross-examining with the reviews. After Round-II, `coherence` and `specificity` are the most disagreed upon between raters. This could be attributed to their subjective nature (Kryściński et al., 2018).

| | **Round-I ↑** | **Round-II ↑** |
|---|---|---|
| `fluency` | 0.55 | 0.84 |
| `coherence` | 0.43 | 0.73 |
| `relevance` | 0.50 | 0.79 |
| `faithfulness` | 0.63 | 0.86 |
| `aspect coverage` | 0.64 | 0.82 |
| `sentiment consistency` | 0.41 | 0.78 |
| `specificity` | 0.34 | 0.76 |
| **AVG** | 0.50 | 0.80 |

**Table 1: Krippendorff's alpha coefficient** ($\alpha$) for Round-I and Round-II on 7 dimensions. As expected, we see an improvement in Round-II coefficient scores.

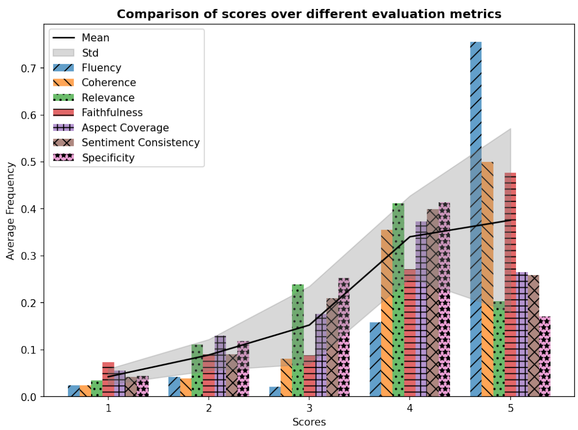


**Figure 3: Ratings Distribution.** We plot the average frequency of scores obtained by human raters across 7 dimensions. A score of 4 or 5 is mostly preferred.

Figure 3 shows the average frequency of assigning a particular score by human raters for 7 dimensions. We make some key observations: (*a*) a score of 4 or 5 is mostly preferred. This could be attributed to the fact that most of the models are LLMs which are doing pretty well for summary generation tasks. (*b*) for `fluency`, `coherence`, and `faithfulness` a score of 5 dominates. This indicates that the LLMs are doing good in terms of these dimensions. (*c*) for `relevance`, `aspect coverage`, `sentiment consistency`, and `specificity` raters majorly prefer a score of 4.

## 5 Experiments

We discuss the available benchmark dataset for opinion summary evaluation, the summarization models used for opinion summary generation, baseline metrics, and the implementation details.

### 5.1 Summarization Models

**Pre-LLMs**: For the Pre-LLMs, we obtain the publicly available summaries for the Amazon test set of these models. These models were trained in a self-supervised manner using only reviews data. (*1*) `PlanSum` (Amplayo and Lapata, 2020) uses content plans to create relevant review-summary pairs. The content plans take the form of aspect and sentiment distributions which are used along with input reviews for generating summaries. (*2*) `MultimodalSum` (Im et al., 2021) uses non-text data such as image and metadata along with reviews to generate opinion summaries. It uses a separate encoder for each modality and uses synthetic datasets to train the model in an end-to-end fashion. (*3*) Siledar et al. (2023) uses lexical and semantic similarities to create a highly relevant synthetic dataset of review-summary pairs. This is then used to fine-tune any pre-trained language model for generating opinion summaries. (hereby referred to as `LS-Sum-G`).

**LLMs**: For the LLMs, we use simple prompts [2] to generate opinion summaries. These models were not specifically fine-tuned for opinion summarization. We use the HuggingFace library (Wolf et al., 2020) to access these models. (*1*) `ChatGPT-3.5` and `GPT-4` (OpenAI, 2023) are closed-source models from OpenAI optimized for dialog. We use the `gpt-3.5-turbo-0125` and `gpt-4-0125-preview` versions. (*2*) `LLaMA2-7B` and `LLaMA2-13B` (Touvron et al., 2023) are open-source fine-tuned model from Meta with 7B and 13B parameters respectively. They were trained autoregressively using around 2T tokens. We use the chat version: `meta-llama/Llama-2-7b-chat-hf` model and `meta-llama/Llama-2-13b-chat-hf` from the HuggingFace library. (*3*) `Mistral-7B` (Jiang et al., 2023) is a 7B instruction-tuned LLM created by MistralAI. We use the instruct version: `mistralai/Mistral-7B-Instruct-v0.2` model. (*4*) `Vicuna-7B` and `Vicuna-13B` (Chiang et al., 2023) are open-source 7B and 13B parameter chat models trained by fine-tuning `LLaMA2` on 125K user-shared conversations collected from ShareGPT (ShareGPT). We use the: `lmsys/vicuna-7b-v1.5` model and `lmsys/vicuna-13b-v1.5` model. (*5*) `Solar-10.7B` (Kim et al., 2023) is an LLM with 10.7B parameters, showing remarkable performance for models with parameters under 30B. We use the version: `upstage/SOLAR-10.7B-Instruct-v1.0` model. (*6*) `Zephyr-7B` (Tunstall et al., 2023) is an open-sourced fine-tuned version of `mistralai/Mistral-7B-v0.1` that was trained on a mix of publicly available, synthetic datasets using Direct Preference Optimization (DPO) (Rafailov et al., 2023). We use the beta version: `HuggingFaceH4/zephyr-7b-beta` model.

### 5.2 Baselines

Following baseline metrics are used: ROUGE-{1,2,L} score (Lin, 2004), BERTSCORE (Zhang et al., 2019), BARTSCORE (Yuan et al., 2021), SUMMAC (Laban et al., 2022), UNIEVAL (Zhong et al., 2022). We include G-EVAL (Liu et al., 2023) as our prompt-based baseline. G-EVAL-3.5 and G-EVAL-MISTRAL use `ChatGPT-3.5` and `Mistral-7B` as their LLMs.

### 5.3 Implementation Details

For evaluation, we used `Mistral-7B` (`mistralai/Mistral-7B-Instruct-v0.2`) as our evaluator model. We chose `Mistral-7B` for these reasons: (*a*) it ranked best amongst the open-source models on the `lmsys/chatbot-arena-leaderboard`, (*b*) we found its instruction following-ness to be better than alternatives, and (*c*) its `7B` size ensures easy replication. We set the hyperparameters to `n=100, temperature=0.7` to sample multiple generations. Example prompts are in **Appendix D**.

## 6 Results and Analysis

**G-EVAL vs. OP-I-PROMPT vs. OP-PROMPTS.** Table 2 and Table 5 report the summary-level [3] correlation scores on the SUMMEVAL-OP and OPINSUMMEVAL dataset. In the case of closed-source models, we observe that our OP-I-GPT-3.5 outperforms or performs comparably to G-EVAL-3.5 across all dimensions on both datasets. Specifically, our OP-I-GPT-3.5 outperforms G-EVAL-3.5 on all 4 dimensions for the OPINSUMMEVAL dataset, whereas for the SUMMEVAL-OP dataset, outperforms on `coherence`, `faithfulness`, and `aspect coverage`, performs comparably on `relevance` and `sentiment consistency`, underperforms slightly on `fluency` and `specificity`.

[2]Check **Appendix D.4** for the prompt

[3]Check **Appendix C** for definition.

| | | FL ↑ | | CO ↑ | | RE ↑ | | FA ↑ | | AC ↑ | | SC ↑ | | SP ↑ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $\rho$ | $\tau$ | $\rho$ | $\tau$ | $\rho$ | $\tau$ | $\rho$ | $\tau$ | $\rho$ | $\tau$ | $\rho$ | $\tau$ | $\rho$ | $\tau$ |
| **SUMMEVAL-OP (Ours)** | HUMANS | 0.80 | 0.77 | 0.81 | 0.76 | 0.91 | 0.86 | 0.89 | 0.85 | 0.93 | 0.87 | 0.91 | 0.85 | 0.92 | 0.87 |
| | ROUGE-1 | -0.36 | -0.28 | -0.30 | -0.24 | -0.31 | -0.23 | -0.35 | -0.26 | -0.44 | -0.32 | -0.38 | -0.29 | -0.30 | -0.23 |
| | ROUGE-2 | -0.23 | -0.18 | -0.14 | -0.10 | -0.17 | -0.12 | -0.21 | -0.16 | -0.26 | -0.19 | -0.24 | -0.18 | -0.14 | -0.09 |
| | ROUGE-L | -0.39 | -0.32 | -0.30 | -0.23 | -0.34 | -0.25 | -0.40 | -0.30 | -0.51 | -0.37 | -0.45 | -0.33 | -0.38 | -0.27 |
| | BERTSCORE | -0.32 | -0.27 | -0.28 | -0.22 | -0.29 | -0.22 | -0.34 | -0.26 | -0.51 | -0.43 | -0.41 | -0.33 | -0.37 | -0.28 |
| | BARTSCORE | -0.19 | -0.15 | -0.19 | -0.14 | -0.29 | -0.22 | -0.33 | -0.25 | -0.45 | -0.35 | -0.37 | -0.28 | -0.36 | -0.27 |
| | SUMMAC | 0.23 | 0.20 | 0.18 | 0.14 | 0.30 | 0.25 | 0.25 | 0.21 | 0.24 | 0.19 | 0.25 | 0.20 | 0.26 | 0.21 |
| | UNIEVAL | 0.36 | 0.28 | 0.52 | 0.42 | 0.33 | 0.25 | 0.17 | 0.14 | - | - | - | - | - | - |
| | G-EVAL-3.5 | **0.63** | **0.55** | 0.59 | 0.49 | 0.68 | **0.56** | 0.70 | 0.58 | 0.79 | 0.67 | **0.73** | **0.61** | **0.75** | **0.63** |
| | **OP-I-GPT-3.5** | 0.60 | 0.51 | **0.61** | **0.51** | **0.69** | **0.56** | **0.71** | **0.59** | 0.80 | 0.68 | **0.73** | **0.61** | 0.74 | 0.61 |
| | G-EVAL-MISTRAL | 0.50 | 0.43 | 0.54 | 0.45 | 0.52 | 0.42 | 0.54 | 0.44 | 0.61 | 0.49 | 0.55 | 0.46 | 0.62 | 0.50 |
| | **OP-MISTRAL** | 0.38 | 0.32 | 0.58 | 0.47 | 0.56 | 0.45 | 0.57 | 0.46 | 0.80 | 0.67 | 0.60 | 0.49 | 0.75 | 0.62 |
| | **OP-I-MISTRAL** | 0.54 | 0.45 | 0.58 | 0.47 | 0.59 | 0.47 | 0.63* | 0.51* | **0.82*** | **0.70*** | **0.73*** | **0.61*** | 0.71* | 0.58* |

**Table 2:** *Spearman* ($\rho$) and *Kendall Tau* ($\tau$) correlations at summary-level on 7 dimensions for the SUMMEVAL-OP dataset. For closed-source, OP-I-PROMPT performs comparably to G-EVAL, whereas for open-source it outperforms alternatives. $*$ represents significant performance (p-value $< 0.05$) to G-EVAL-MISTRAL computed using Mann-Whitney U Test. HUMANS- averaged correlation of each annotator with the overall averaged ratings.

| Method | FL ↑ | CO ↑ | RE ↑ | FA ↑ | AC ↑ | SC ↑ | SP ↑ |
|---|---|---|---|---|---|---|---|
| Human Summaries | 4.39 | 4.41 | 3.78 | 3.98 | 3.54 | 3.71 | 3.66 |
| *Pre-LLMs* | | | | | | | |
| PlanSum | 1.86 | 1.94 | 1.60 | 1.38 | 1.52 | 1.59 | 1.56 |
| MultimodalSum | 4.62 | 4.09 | 2.63 | 2.27 | 2.18 | 2.76 | 2.43 |
| LS-Sum-G | 4.76 | 4.40 | 2.87 | 2.74 | 2.32 | 3.03 | 2.69 |
| *LLMs* | | | | | | | |
| ChatGPT-3.5 | 4.89 | 4.58 | **4.25** | 4.71 | 4.22 | 4.16 | 3.96 |
| GPT-4 | **5.00** | **4.91** | 3.52 | **4.96** | **4.93** | **4.83** | **4.57** |
| LLaMA2-7B | 4.79 | 4.34 | 3.77 | 4.49 | 3.67 | 3.79 | 3.46 |
| LLaMA2-13B | 4.87 | 4.49 | **4.25** | 4.62 | 4.02 | 4.00 | 3.94 |
| Mistral-7B | 4.86 | 4.60 | 4.33 | 4.66 | 4.56 | 4.35 | 4.25 |
| Vicuna-7B | 4.83 | 4.23 | 3.92 | 4.35 | 3.96 | 3.92 | 3.67 |
| Vicuna-13B | 4.87 | 4.41 | 4.09 | 4.43 | 4.03 | 4.00 | 3.77 |
| Solar-10.7B | 4.89 | 4.73 | 4.20 | 4.72 | 4.50 | 4.56 | 4.35 |
| Zephyr-7B | 4.89 | 4.36 | 4.08 | 4.54 | 4.18 | 3.95 | 3.83 |

**Table 3:** Model-wise averaged annotator ratings of opinion summaries along 7 dimensions for the Amazon test set. Best scores are in **bold**, second-best are underlined.

| | | AVG-S | MW ↓ | TT ↓ |
|---|---|---|---|---|
| FL | **G-EVAL-MISTRAL***<br>OP-I-MISTRAL | **0.48**<br>0.38 | $\mathbf{2.9 \times 10^{-4}}$ | $\mathbf{2.6 \times 10^{-4}}$ |
| CO | **G-EVAL-MISTRAL***<br>OP-I-MISTRAL | **0.52**<br>0.47 | $\mathbf{2.1 \times 10^{-4}}$ | $\mathbf{3.2 \times 10^{-8}}$ |
| RE | G-EVAL-MISTRAL<br>OP-I-MISTRAL | **0.51**<br>0.49 | $6.7 \times 10^{-2}$ | $\mathbf{4.6 \times 10^{-2}}$ |
| FA | G-EVAL-MISTRAL<br>OP-I-MISTRAL | 0.53<br>**0.54** | $\mathbf{1.9 \times 10^{-1}}$ | $4.7 \times 10^{-1}$ |
| AC | G-EVAL-MISTRAL<br>**OP-I-MISTRAL*** | 0.58<br>**0.74** | $\mathbf{2.1 \times 10^{-4}}$ | $\mathbf{1.9 \times 10^{-14}}$ |
| SC | G-EVAL-MISTRAL<br>**OP-I-MISTRAL*** | 0.54<br>**0.63** | $\mathbf{2.1 \times 10^{-4}}$ | $\mathbf{1.4 \times 10^{-7}}$ |
| SP | G-EVAL-MISTRAL<br>**OP-I-MISTRAL*** | 0.59<br>**0.63** | $\mathbf{7.4 \times 10^{-4}}$ | $\mathbf{3.0 \times 10^{-4}}$ |

**Table 4: Significance Test.** P-values computed using Mann-Whitney U Test (MW) and T-Test (TT) between the average Spearman correlation scores (AVG-S) taken over 10 independent generations from G-EVAL-MISTRAL and OP-I-MISTRAL. **Bold** for AVG-S indicates better performance, and for MW and TT indicates p-value $< 0.05$. $*$ represents significant performance.

For open-source models, overall, we observe that OP-I-MISTRAL performs the best, followed by OP-MISTRAL and then G-EVAL-MISTRAL. Figure 4 shows the performance of different prompt approaches over `n=100` generations for 7 dimensions. As we increase the number of generations we generally observe an improvement in the correlation. OP-I-MISTRAL shows an improvement against G-EVAL-MISTRAL across all 7 dimensions and by a large margin specifically for `aspect coverage`, `sentiment consistency`, and `specificity`.

**Significance Testing.** We perform significance testing using the Mann-Whitney U Test (McKnight and Najab, 2010) for comparison between OP-I-MISTRAL and G-EVAL-MISTRAL. Table 2 report results for Spearman and Kendall Tau scores computed by using the scoring function with `n=100`. OP-I-MISTRAL significantly (p-value $< 0.05$) outperforms G-EVAL-MISTRAL on `faithfulness`, `aspect coverage`, `sentiment consistency`, and `specificity`. Additionally, we group the 100 generations into 10 independent groups and compute Spearman correlations for each group. Table 4 reports the Mann-Whitney U Test and T-Test p-values and arrives at a similar observation of OP-I-MISTRAL significantly outperforming on aforementioned dimensions except `faithfulness`.

**Models for Opinion Summarization.** Table 3

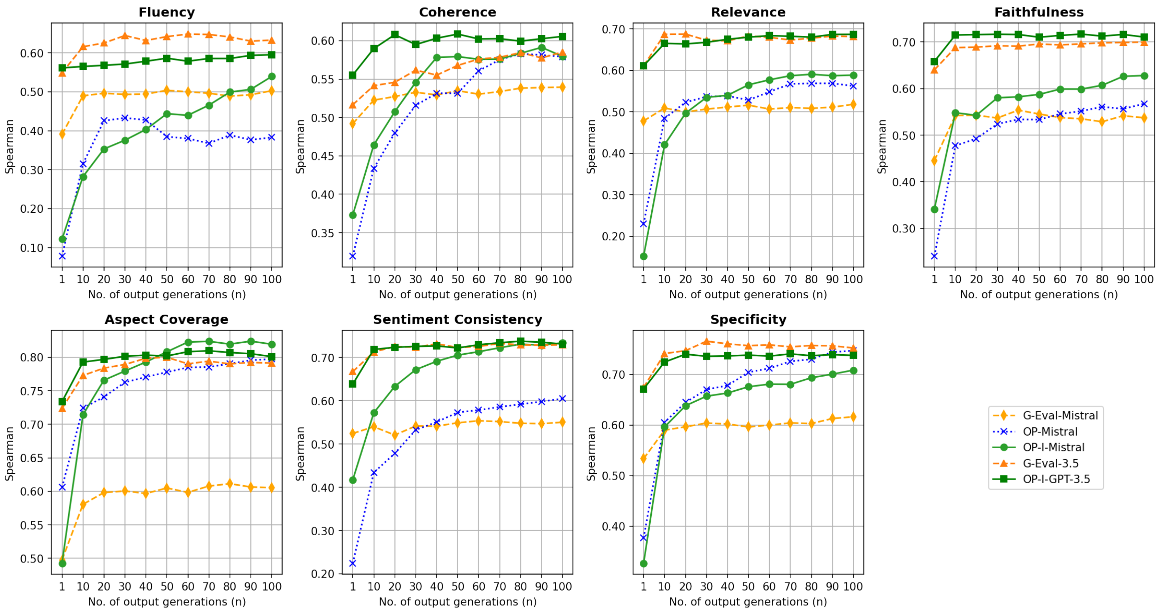


**Figure 4: *Spearman* correlation scores at different number of output generations** ($\mathtt{n}$) **for the** 7 **dimensions.** G-EVAL-3.5 and OP-I-GPT-3.5 use the G-EVAL and OP-I-PROMPT respectively, with closed-source `ChatGPT-3.5` as their LLM. G-EVAL-MISTRAL, OP-I-MISTRAL, and OP-MISTRAL use the G-EVAL, OP-I-PROMPT, and OP-PROMPTS respectively, with open-source `Mistral-7B` as their LLM. Generally, OP-I-PROMPT shows better relative performance on both closed-source and open-source models.

reports averaged annotator ratings for the 7 dimensions for each model (Refer to Figure 5 for a graphical view). Overall, `GPT-4` ranks the best, followed by `Solar-10.7B` and `Mistral-7B` ranking second-best, followed by `ChatGPT-3.5`. As expected, Pre-LLM models are rated the worst. These are self-supervised models and do not enjoy the liberty of being trained on trillions of tokens. All the LLMs outperform human summaries. However, because these summaries were written in the first person and as a review itself to cater to the needs when the test set was created, it is inconclusive if the LLMs outperform humans in general. We observe that `GPT-4` model scores poorly in `relevance` dimension. This we figured was due to the tendency of `GPT-4` model to try to cover every detail in the summary. Finally, `Solar-10.7B` and `Mistral-7B` with just 10.7B and 7B parameters respectively, outperformed `ChatGPT-3.5` and comes close in performance to `GPT-4`.

**Metric Evaluation.** Reference-based metrics (ROUGE 1,2,L, BERTSCORE) as expected show weak correlation with human judgments. Reference-free metrics such as BARTSCORE does very poorly, however, SUMMAC performs moderately well. UNIEVAL does well in `coherence` but still trails behind prompt-based approaches. *To summarize, reference-based metrics are inadequate for assessing model performances in the LLMs era.*

**How sensitive is OP-I-PROMPT?** We test OP-I-PROMPT for 3 definition variations of the `aspect coverage` dimension. We paraphrase the original definition (Section 4.2) to create 2 additional versions, ensuring the meaning is preserved (**Appendix E**). We let the OP-I-MISTRAL generate `n=100` responses to estimate the score using the scoring function (Section 3.2). The Spearman correlations for the 3 variations are 0.82 (Table 2), 0.82, and 0.81, indicating that OP-I-PROMPT *is indifferent to the variations of dimensions' definition.*

## 7 Conclusion

In this work, we present the SUMMEVAL-OP dataset, OP-I-PROMPT and OP-PROMPTS for opinion summary evaluation on 7 dimensions. Experimentally, we observe OP-I-PROMPT outperforms alternatives on open-source models and performs comparably better on closed-source models showing good correlations with human judgements. Some key takeaways are: (*a*) Prompts that do well for powerful closed-source LLMs may not work well for open-source LLMs; (*b*) Opinion summaries by LLMs are preferred by humans compared to reference and previous model summaries; (*c*) Reference-based summaries and metrics are inadequate in assessing LLM-based outputs.

## Limitations

1. We do not use GPT-4 for evaluation purpose due to cost constraints. The primary aim of our work was to design prompts and test their applicability to both open-source and closed-source. We use ChatGPT-3.5 as the closed-source model to perform our experiments.

2. Our OP-I-PROMPT was specifically designed to evaluate any dimension of the opinion summaries where OP-PROMPTS are dimension-dependent. However, their applicability to other tasks needs further investigation and appropriate changes to the prompt.

3. Due to the nature of the available test sets and for benchmarking the already available models, SUMMEVAL-OP considers only 8 input reviews following the literature. This we believe is a major limitation in the opinion summarization field. Datasets with a larger number of reviews need to be considered for the creation of future benchmark datasets.

4. The assessment quality of all the prompt approaches needs to be investigated for a larger amount of reviews as well.

## Ethical Considerations

The SUMMEVAL-OP dataset was created using the already available Amazon test set. We hired 3 raters who have written papers on opinion summarization (1) or are working in the opinion summarization domain (2). These were male Masters' students aged 21-30. All the raters received stipends suitable for the tasks.

The OP-I-PROMPT and OP-PROMPTS are designed to offer automatic evaluation of opinion summaries for multiple dimensions. Its primary aim is to assist researchers, developers, and other stakeholders in accurately assessing summaries generated by NLG systems. However, there are potential risks associated with these prompts if they fail to accurately evaluate the quality of opinion summaries or exhibit a bias towards LLM-created content. We urge the research community to use these prompts with caution and check their reliability for their use cases.

## References


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.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. 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.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the use of large language models for reference-free text quality evaluation: An empirical study.

Cheng-Han Chiang and Hung-yi Lee. 2023a. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023b. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

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.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire.

Dan Gillick and Yang Liu. 2010. Non-expert evaluation of summarization systems is risky. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 148–151, Los Angeles. Association for Computational Linguistics.

Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. *Proceedings of the 25th International Conference on World Wide Web*.

Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2021. Self-supervised multimodal opinion summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 388–403, Online. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Dahyun Kim, Chanjun Park, Sanghoon Kim, Wonsung Lee, Wonho Song, Yunsu Kim, Hyeonwoo Kim, Yungi Kim, Hyeonju Lee, Jihoo Kim, Changbae Ahn, Seonghoon Yang, Sukyung Lee, Hyunbyung Park, Gyoungjin Gim, Mikyoung Cha, Hwalsuk Lee, and Sunghun Kim. 2023. Solar 10.7b: Scaling large language models with simple yet effective depth up-scaling.

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.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Wojciech Kryściński, Romain Paulus, Caiming Xiong, and Richard Socher. 2018. Improving abstraction in text summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1808–1817, Brussels, Belgium. Association for Computational Linguistics.

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.

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. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Patrick E McKnight and Julius Najab. 2010. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1.

OpenAI. 2023. ChatGPT (August 3 Version). `https://chat.openai.com`.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

ShareGPT. Sharegpt. `https://sharegpt.com/`. Accessed on February 15, 2024.

Yuchen Shen and Xiaojun Wan. 2023. Opinsummeval: Revisiting automated evaluation for opinion summarization. *arXiv preprint arXiv:2310.18122*.

Tejpalsingh Siledar, Suman Banerjee, Amey Patil, Sudhanshu Singh, Muthusamy Chelliah, Nikesh Garera, and Pushpak Bhattacharyya. 2023. Synthesize, if you do not have: Effective synthetic dataset creation strategies for self-supervised opinion summarization in E-commerce. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13480–13491, Singapore. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages

1–11, Singapore. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi 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 M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

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

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

## A Available Benchmark Dataset

**OPINSUMMEVAL**: (Shen and Wan, 2023) used the Yelp test set (Chu and Liu, 2019) to annotate for 4 dimensions: `readability`, `self-coherence`, `aspect relevance`, and `sentiment consistency`. The dataset contains a total of 100 products with 8 reviews and 14 different model summaries per product. Each summary was rated by 2 annotators on 4 dimensions. For consistency, we hereby refer to the above-mentioned dimensions as `fluency`, `coherence`, `aspect coverage`, and `sentiment consistency` respectively, in line with our definitions.

| | | FL ↑ | | CO ↑ | | AC ↑ | | SC ↑ | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\rho$ | $\tau$ | $\rho$ | $\tau$ | $\rho$ | $\tau$ | $\rho$ | $\tau$ |
| OPINSUMMEVAL | ROUGE-1† | 0.08 | 0.06 | 0.11 | 0.09 | 0.14 | 0.11 | 0.00 | 0.00 |
| | ROUGE-2† | 0.13 | 0.10 | 0.13 | 0.11 | 0.15 | 0.11 | 0.04 | 0.04 |
| | ROUGE-L† | 0.13 | 0.10 | 0.18 | 0.15 | 0.18 | 0.14 | 0.07 | 0.05 |
| | BERTSCORE† | 0.38 | 0.30 | 0.20 | 0.17 | 0.20 | 0.16 | 0.08 | 0.06 |
| | BARTSCORE† | 0.42 | 0.33 | 0.35 | 0.29 | 0.28 | 0.22 | 0.41 | 0.34 |
| | SUMMAC† | 0.06 | 0.05 | 0.02 | 0.01 | 0.07 | 0.06 | 0.20 | 0.17 |
| | CHATGPT-3.5† | 0.47 | 0.42 | 0.28 | 0.25 | 0.34 | 0.30 | 0.37 | 0.33 |
| | G-EVAL-3.5† | 0.41 | 0.36 | 0.29 | 0.26 | 0.27 | 0.23 | 0.38 | 0.34 |
| | **OP-I-GPT-3.5** | **0.51** | **0.46** | 0.36 | 0.32 | 0.33 | 0.29 | 0.43 | 0.39 |
| | G-EVAL-MISTRAL | 0.46 | 0.41 | 0.41 | 0.38 | 0.36 | 0.32 | **0.49** | **0.45** |
| | **OP-MISTRAL** | 0.35 | 0.33 | **0.45** | **0.41** | 0.34 | 0.30 | 0.45 | 0.41 |
| | **OP-I-MISTRAL** | 0.46 | 0.41 | 0.37 | 0.35 | **0.38** | **0.34** | **0.49** | **0.45** |

**Table 5:** *Spearman* ($\rho$) and *Kendall Tau* ($\tau$) correlations at summary-level on 7 dimensions for the SUMMEVAL-OP dataset. For closed-source, OP-I-PROMPT performs comparably to G-EVAL, whereas for open-source it outperforms alternatives. † represents results as reported in Shen and Wan (2023)

## B Rater Agreement

We hired 3 raters who have written papers on opinion summarization (1) or are working in the opinion summarization domain (2). These were male Masters' students aged 21-30. They were provided with detailed guidelines for evaluating summaries on the 7 dimensions. All 3 raters received stipends suitable for the tasks. The annotation interface provided raters with the reviews and associated summaries product-wise. Models associated with summaries were not revealed to the raters to remove any bias.

Table 6 reports pairwise root mean squared error scores for the 3 raters. For Round-I, we observe a difference of more than 1 on average. For Round-II, as expected the average difference between any two ratings come down to below 1. Table 7 reports the pairwise correlations between raters as well as the correlation between each rater and average ratings for both Round-I and Round-II.

## C Opinion Summary Evaluation

For each product $d_i$ in dataset $\mathcal{D}$, $i \in \{1, ..., \mathcal{Z}\}$ we have $\mathcal{N}$ opinion summaries from different models. Let $s_{ij}$ denote the $j^{th}$ summary of the product $d_i$, $\mathcal{M}_m$ denote the $m^{th}$ evaluation metric, and $\mathcal{K}$ denote the correlation measure. Bhandari et al. (2020) defines the summary-level correlation as:

$$\mathcal{R}(a,b) = \frac{1}{\mathcal{Z}} \sum_i \mathcal{K}([\mathcal{M}_a(s_{i1}), ..., \mathcal{M}_a(s_{i\mathcal{N}})], [\mathcal{M}_b(s_{i1}), ..., \mathcal{M}_b(s_{i\mathcal{N}})]) \quad (2)$$

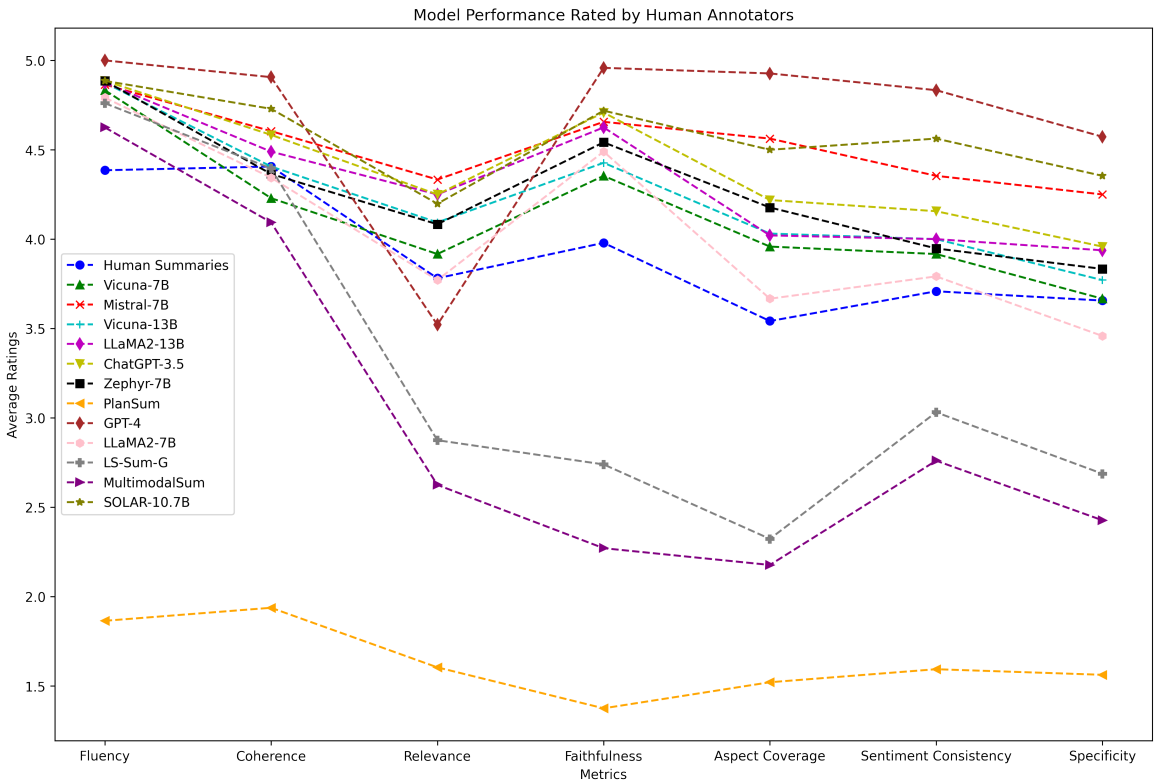


**Figure 5: Performance of different models as rated by human annotators.** We observe that GPT-4 performs the best followed by Solar-10.7B and Mistral-7B. Self-supervised models perform worse. In general, all the LLMs perform better than human annotated summaries.

| | FL ↓ | CO ↓ | RE ↓ | FA ↓ | AC | SC ↓ | SP ↓ |
|---|---|---|---|---|---|---|---|
| | | | *Round-I* | | | | |
| **A1-A2** | 0.95 | 1.06 | 1.01 | 1.09 | 0.91 | 1.08 | 0.95 |
| **A2-A3** | 0.44 | 0.86 | 1.09 | 1.05 | 0.84 | 1.19 | 1.42 |
| **A1-A3** | 1.00 | 1.23 | 1.16 | 1.24 | 1.15 | 1.47 | 1.55 |
| **AVG-I** | 0.80 | 1.05 | 1.09 | 1.13 | 0.97 | 1.25 | 1.31 |
| | | | *Round-II* | | | | |
| **A1-A2** | 0.55 | 0.66 | 0.65 | 0.60 | 0.64 | 0.64 | 0.60 |
| **A2-A3** | 0.31 | 0.62 | 0.67 | 0.67 | 0.68 | 0.71 | 0.79 |
| **A1-A3** | 0.53 | 0.73 | 0.67 | 0.68 | 0.76 | 0.76 | 0.73 |
| **AVG-II** | 0.47 | 0.67 | 0.66 | 0.65 | 0.69 | 0.70 | 0.71 |

**Table 6: Round-I and Round-II Ratings:** Pairwise *Root Mean Squared Error* scores for 3 raters A1, A2, and A3.

## D Prompts

For brevity, we provide different prompts for only a single dimension- Aspect Coverage. We will release prompts for all the dimensions across different approaches publicly.

### D.1 Op-I-Prompt for Aspect Coverage

**Task Description**:
You will be given a set of reviews using which a summary has been generated. Your task is to evaluate the summary based on the given metric. Evaluate to which extent does the summary follows the given metric considering the reviews as the input. Use the following evaluation criteria to judge the extent to which the metric is followed. Make sure you understand the task and the following evaluation metric very clearly.

**Evaluation Criteria**:
The task is to judge the extent to which the metric is followed by the summary. Following are the scores and the evaluation criteria according to which scores must be assigned.
<score>1</score> - The metric is not followed at all while generating the summary from the reviews.
<score>2</score> - The metric is followed only to a limited extent while generating the summary from the reviews.
<score>3</score> - The metric is followed to a good extent while generating the

| Round | | FL ↑ ρ | FL ↑ τ | CO ↑ ρ | CO ↑ τ | RE ↑ ρ | RE ↑ τ | FA ↑ ρ | FA ↑ τ | AC ↑ ρ | AC ↑ τ | SC ↑ ρ | SC ↑ τ | SP ↑ ρ | SP ↑ τ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Round-I* | *Pairwise correlation among raters* | | | | | | | | | | | | | | |
| | **A1-A2** | 0.58 | 0.56 | 0.54 | 0.50 | 0.65 | 0.60 | 0.73 | 0.68 | 0.78 | 0.72 | 0.60 | 0.53 | 0.65 | 0.59 |
| | **A2-A3** | 0.79 | 0.78 | 0.40 | 0.38 | 0.52 | 0.47 | 0.63 | 0.58 | 0.77 | 0.71 | 0.56 | 0.51 | 0.58 | 0.53 |
| | **A1-A3** | 0.55 | 0.53 | 0.34 | 0.31 | 0.40 | 0.36 | 0.60 | 0.54 | 0.74 | 0.68 | 0.57 | 0.51 | 0.57 | 0.51 |
| | **AVG-I** | 0.64 | 0.62 | 0.43 | 0.40 | 0.52 | 0.48 | 0.65 | 0.60 | 0.76 | 0.70 | 0.58 | 0.52 | 0.60 | 0.54 |
| | *Pairwise correlation between raters and the overall average ratings* | | | | | | | | | | | | | | |
| | **A-A1** | 0.95 | 0.93 | 0.86 | 0.82 | 0.81 | 0.74 | 0.86 | 0.80 | 0.91 | 0.85 | 0.85 | 0.77 | 0.87 | 0.80 |
| | **A-A2** | 0.70 | 0.67 | 0.72 | 0.67 | 0.82 | 0.75 | 0.81 | 0.75 | 0.91 | 0.85 | 0.85 | 0.78 | 0.87 | 0.80 |
| | **A-A3** | 0.57 | 0.54 | 0.56 | 0.51 | 0.74 | 0.67 | 0.81 | 0.76 | 0.88 | 0.81 | 0.76 | 0.69 | 0.76 | 0.69 |
| | **AVG-II** | 0.74 | 0.71 | 0.71 | 0.67 | 0.79 | 0.72 | 0.83 | 0.77 | 0.90 | 0.84 | 0.82 | 0.75 | 0.83 | 0.76 |
| *Round-II* | *Pairwise correlation among raters* | | | | | | | | | | | | | | |
| | **A1-A2** | 0.63 | 0.61 | 0.64 | 0.61 | 0.80 | 0.75 | 0.83 | 0.79 | 0.85 | 0.80 | 0.77 | 0.72 | 0.81 | 0.76 |
| | **A2-A3** | 0.85 | 0.84 | 0.59 | 0.56 | 0.78 | 0.73 | 0.77 | 0.73 | 0.83 | 0.78 | 0.77 | 0.73 | 0.81 | 0.77 |
| | **A1-A3** | 0.66 | 0.65 | 0.59 | 0.56 | 0.77 | 0.72 | 0.78 | 0.73 | 0.84 | 0.79 | 0.79 | 0.74 | 0.82 | 0.78 |
| | **AVG-I** | 0.71 | 0.70 | 0.61 | 0.58 | 0.78 | 0.73 | 0.79 | 0.75 | 0.84 | 0.79 | 0.78 | 0.73 | 0.81 | 0.77 |
| | *Pairwise correlation between raters and the overall average ratings* | | | | | | | | | | | | | | |
| | **A-A1** | 0.94 | 0.92 | 0.87 | 0.83 | 0.91 | 0.85 | 0.89 | 0.84 | 0.94 | 0.88 | 0.92 | 0.86 | 0.92 | 0.87 |
| | **A-A2** | 0.76 | 0.74 | 0.80 | 0.75 | 0.91 | 0.86 | 0.87 | 0.82 | 0.93 | 0.88 | 0.91 | 0.85 | 0.92 | 0.87 |
| | **A-A2** | 0.69 | 0.67 | 0.76 | 0.71 | 0.91 | 0.85 | 0.92 | 0.88 | 0.93 | 0.86 | 0.91 | 0.85 | 0.92 | 0.87 |
| | **AVG-II** | 0.80 | 0.77 | 0.81 | 0.76 | 0.91 | 0.86 | 0.89 | 0.85 | 0.93 | 0.87 | 0.91 | 0.85 | 0.92 | 0.87 |

**Table 7: Rater Correlations:** Pairwise *Spearman* ($\rho$) and *Kendall Tau* ($\tau$) correlations at summary-level for 3 raters A1, A2, and A3 along with the average of their ratings A.

summary from the reviews.
<score>4</score> - The metric is followed mostly while generating the summary from the reviews.
<score>5</score> - The metric is followed completely while generating the summary from the reviews.

**Metric**:
Aspect Coverage - The summary should cover all the aspects that are majorly being discussed in the reviews. Summaries should be penalized if they miss out on an aspect that was majorly being discussed in the reviews and awarded if it covers all.

**Reviews**:
{}

**Summary**:
{}

**Evaluation Steps**:
Follow the following steps strictly while giving the response:
1.First write down the steps that are needed to evaluate the summary as per the metric. Reiterate what metric you will be using to evaluate the summary.
2.Give a step-by-step explanation if the summary adheres to the metric considering the reviews as the input. Stick to the metric only for evaluation.
3.Next, evaluate the extent to which the metric is followed.
4.Use the previous information to rate the summary using the evaluation criteria and assign a score within the <score></score> tags.

Note: Strictly give the score within <score></score> tags only e.g Score- <score>5</score>.

First give a detailed explanation and then finally give a single score following the format: Score- <score>5</score>

THE EVALUATION AND SCORE MUST BE ASSIGNED STRICTLY ACCORDING TO THE METRIC ONLY AND NOTHING ELSE!

**Response**:

## D.2 Op-Prompts for Aspect Coverage

**Task Description**:
You will be given a set of reviews. You will then be given one summary written for the set of reviews. Your task is to rate the summary on one metric. Make sure you understand the following evaluation metric very clearly. Your task is to rate the summary corresponding to the given reviews on the evaluation criteria.

**Evaluation Criteria**:
Aspect Coverage - The summary should cover all the aspects that are majorly being discussed in the reviews. Summaries should be penalized if they miss out on an aspect that was majorly being discussed in the reviews and awarded if it covers all.
<score>1</score> - Summary does not cover any important aspects present in the reviews.
<score>2</score> - Summary does not cover most of the important aspects present in the reviews.
<score>3</score> - Summary covers around half of the important aspects present in the reviews.
<score>4</score> - Summary covers most of the important aspects present in reviews.
<score>5</score> - Summary covers all the important aspects discussed in reviews.

**Metric**:
Aspect Coverage - The summary should cover all the aspects that are majorly being discussed in the reviews. Summaries should be penalized if they miss out on an aspect that was majorly being discussed in the reviews and awarded if it covers all.

**Reviews**:
{}

**Summary**:
{}

**Evaluation Steps**:
Let's go step-by-step. Follow the following steps strictly while giving the response:
1.Identify the important aspects present in the reviews and list them with numbering
2.Identify the important aspects present in the summary and list them with numbering
3.Identify the important aspects covered by the summary that are present in the reviews and list them with numbering
4.Calculate the total number of important aspects covered by the summary that are present in the reviews
5.Calculate the total number of important aspects present in the reviews
6.Finally use the evaluation criteria to output only a single score within <score></score> tags.

Note: Strictly give the score within <score></score> tags only e.g Score- <score>5</score>.

First give a detailed explanation of how much is the coverage and then finally give a single score following the format: Score- <score>5</score>

**Response**:

## D.3 G-Eval for Aspect Coverage

**Task Description**:
You will be given a set of reviews and a corresponding summary. Make sure you understand the following evaluation metric very clearly. Your task is to rate the summary corresponding to the given reviews on the evaluation criteria.

**Evaluation Criteria**: Aspect Coverage (1-5) - The summary should cover all the aspects that are majorly being discussed in the reviews. Summaries should be penalized if they miss out on an aspect that was majorly being discussed in the reviews and awarded if it covers all.

**Reviews**:
{}

**Summary**:
{}

**Evaluation Steps**:
1.Read through the given set of reviews carefully.
2.Compare the content of the reviews to the provided summary.
3.Evaluate whether the summary covers all the major aspects that are being discussed in the reviews.
4.Rate the summary on a scale of 1-5 based on how well it covers the aspects discussed in the reviews.
5.Provide a brief explanation for your rating, citing specific examples from the reviews and summary.

Note: Strictly give the score within <score></score> tags only e.g Score: <score>5</score>.

**Response**:

### D.4 Summarization Prompt

Generate a summary for the following set of reviews. Generate the summary in a paragraph format. No bulletpoints or explanations needed. Just output the summary text.

**Reviews**:
{}

**Summary**:

## E Dimension Definitions

For ablation, we try out three different definition variations of `aspect coverage`.

**Definition 1:** The summary should cover all the aspects that are majorly being discussed in the reviews. Summaries should be penalized if they miss out on an aspect that was majorly being discussed in the reviews and awarded if it covers all.

**Definition 2:** This refers to the comprehensiveness of a summary in capturing all significant aspects discussed in reviews. A summary is evaluated based on its ability to include major topics of discussion; it is deemed deficient if it overlooks any crucial aspect and commendable if it encompasses them all.

**Definition 3:** Aspect coverage pertains to the extent to which a summary encapsulates the key facets discussed in reviews. Summaries are evaluated based on their ability to incorporate major discussion points. They are considered deficient if they omit any critical aspect and commendable if they address them all comprehensively.