



WILLOWTREE®  
a TELUS Digital Company

# LLM-as-a-Judge for Evaluating Text Summarization Performance

GAIL SMITH

TOBIAS SARTORI DA SILVA

AGNER PITON

CHRISTOPHER FRENCHI



October 2024



ai@willowtreeapps.com  
1-888-329-9875



#### 4 Executive Summary

#### 6 Introduction

- 6 The Increasing Importance of Text Summarization
- 7 Why Evaluating Text Summarization Is Hard (And Becoming More Difficult)

#### 10 Background

- 10 Types of Text Summarization
- 11 Performance, Quality, and Metrics
- 12 Gold Standard Datasets
- 13 Traditional Evaluation Metrics
- 15 Why Traditional Metrics Fall Short

#### 16 LLM-as-a-Judge Methodology

- 16 Concept and Implementation
- 19 Metrics for LLM-as-a-Judge
- 21 Prompting for LLM-as-a-Judge

#### 22 Results

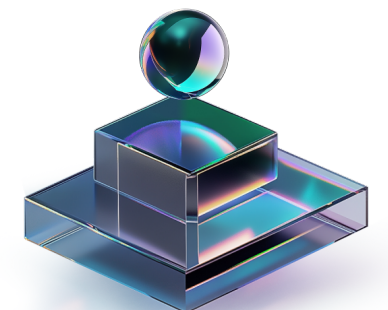
- 22 Comparing Gold Standard Dataset Outputs to LLM-Generated Summaries
- 25 Comparing Dataset Inputs to LLM-Generated Summaries
- 28 Human Feedback
- 29 Challenges to Evaluation

#### 30 Practical Considerations for Real-World Applications

- 30 Evaluation Tradeoffs
- 31 Challenges in Specific Domains

#### 32 Conclusion

#### 34 References



AUTHORS



Gail  
Smith



Agner  
Piton



Tobias  
Sartori Da Silva



Christopher  
Frenchi

PART I

# Executive Summary

---

This ebook explores using large language models (LLMs) to evaluate text summarization performance by applying LLM-as-a-Judge, a powerful technique for evaluating a wide range of subjectively-graded tasks. Due to the exponential growth of digital information, developers and engineers must find more effective methods for measuring text summarization performance. However, traditional evaluation metrics often fall short in capturing the subtleties and nuances that determine summary quality. LLM-as-a-Judge offers a promising solution here by leveraging the semantic understanding capabilities of advanced language models.

**Key points covered in this ebook:**

- Challenges of evaluating text summarization in an era of exponential information growth
- Limitations of traditional metrics like ROUGE and BLEU to measure performance
- Implementation and advantages of LLM-as-a-Judge methodology
- Results of applying LLM-as-a-Judge using a conversational dataset
- Best practices for integrating LLM-as-Judge in production environments



Text summarization – extracting the most essential information from longer texts – is no longer just a helpful tool, but now a necessity in many fields. Whether it's for use cases like news aggregation (where thousands of articles need to be turned into digestible snippets) or for analyzing lengthy legal documents (a time-consuming manual review process often demanding large teams), effective summarization is key.

## Why Evaluating Text Summarization Is Hard (And Becoming More Difficult)

Evaluating summarization systems is not a straightforward task. Summarization is inherently subjective: What one person deems as the most important information in a text might differ significantly from another person's perspective. This subjectivity makes it challenging to create universal metrics that can accurately assess the quality of a summary.

Consider the example to the right comparing two different summarizations about the effects of global warming.

```
python
```

```
1     # Example: Subjectivity in summarization
2     original_text = "The effects of climate change
are far-reaching and complex. Rising temperatures lead
to melting ice caps, which in turn cause sea levels
to rise. This can lead to flooding in coastal areas,
displacing millions of people. Additionally, changing
weather patterns can disrupt agriculture, leading to food
shortages."
3
4     summary_1 = "Climate change causes rising
temperatures and sea levels, leading to coastal flooding
and agricultural disruption."
5     summary_2 = "Climate change impacts include melting
ice caps, rising sea levels, potential displacement of
coastal populations, and agricultural challenges due to
changing weather patterns."
6
7     # Both summaries are valid but focus on different
aspects and levels of detail
```

This puts generative AI practitioners in a bind. Not only is the world's information increasing, but so is its nuance and complexity. The more volume, nuance, and complexity grow, the more traditional metrics like ROUGE and BLEU (more on these in the next section) fall short. Though they can quantify some aspects of summarization performance, these metrics may not fully capture the nuances of content relevance or the coherence of a summary.

PART III

# Background

---

## Types of Text Summarization

Before delving into evaluation methods, it's crucial to understand the two primary types of text summarization techniques, extractive and abstractive summarization. Note we're primarily interested in abstractive summarization because it gives us a better opportunity to use the advanced semantic understanding of LLMs for evaluating generated summaries.

1. **Extractive summarization** is akin to a highlighter tool. It identifies and extracts key phrases or sentences from the original text to construct the summary. The idea is to pull out (i.e., extract) the most informative parts of the text verbatim. The chief benefit of this approach is it doesn't involve generating new sentences, making it relatively simpler and more widely used than abstractive summarization.
2. **Abstractive summarization**, on the other hand, takes a more sophisticated approach. Instead of merely extracting critical portions of the text, this method aims to understand the context and "rephrase" (i.e., abstract) the information in a condensed form,

much like how a human would summarize a piece of text. Abstractive summarization can handle nuances and complex linguistic constructs, but it also brings challenges to maintaining accuracy and coherence.

## Performance, Quality, and Metrics

**Performance** – How we define performance when evaluating a text summarization system varies from use case to use case. However, the main ideas will be the same: How are we evaluating the speed, cost, and consistency of our application? By defining performance this way, we establish a baseline so we can see how any changes to our prompts, models, or system architectures impact the overall performance of the application in question.

**Quality** – Quality is determined by our key metrics. Do the metrics we've chosen offer the team insights into addressing issues concerning our summarization task? Do we have too many metrics? Not enough? These questions can be difficult to answer depending on the task, but they follow the same rules as testing software and data science projects. We need metrics that will give us confidence in the generated results.

**Metrics** – Choosing the right metrics to evaluate a text summarization system is critical for two reasons. First, the right metrics drive a more systematic evaluation process by providing a standardized, objective framework for assessing the performance and quality. Our metrics serve as quantitative measures that enable comparisons between different summarization approaches. Furthermore, they facilitate the identification of strengths and weaknesses within the system.

Second, metrics help address the subjectivity inherent in summarization tasks. While summarization is indeed subjective, metrics provide a way to quantify and measure various aspects of the

summaries, such as:

- Factual accuracy
- Coherence
- Relevance
- Fluency

These metrics act as proxies for human judgment, allowing us to expedite, automate, and scale evaluations that would otherwise take lots of man-hours.

## Gold Standard Datasets

A gold standard dataset is a prerequisite for many traditional evaluation methods (e.g., think of ROUGE requiring a set of reference summaries). By creating a gold standard dataset with known inputs and expected outputs, we have trusted benchmarks to compare against our test results, instilling confidence in:

1. our system's ability to complete its task
2. our ability to measure quality metrics

Of course, we don't always have the luxury of time or resources to procure or create these datasets. **But it's important to know that without a gold standard dataset, tracking the performance and quality of our system over time becomes much more difficult.** Whatever time we have to invest upfront in our gold standard dataset, the ROI will be more than justified in the speed and confidence we gain for our system tests. (Alternatively, we could look at solutions such as [ChainPoll](#), which would replace our gold standard datasets with chain-of-thought and averaged self-consistency scores.)

When choosing or creating a dataset for a specific summarization task, consider your inputs and your expected outputs, as these will guide your performance evaluation. By passing known inputs into the summarization tasks (e.g., comments, chunks, documents, etc.), we can

produce an LLM-generated summary and compare it against our gold standard dataset output or summary.

The inputs should match what we expect to put into our summarization task, and the outputs should be what compare with our LLM-generated responses. The dataset(s) we use forms the baseline for testing our summarization task against different metrics, helping us efficiently measure the quality of our summarization system.

## Traditional Evaluation Metrics

The following intrinsic evaluation metrics have traditionally been used to evaluate text summarization.

**ROUGE:** Recall-Oriented Understudy for Gisting Evaluation, or ROUGE, is a set of metrics used to evaluate automatic summarization and machine translation. It works by comparing the generated summary to a set of reference summaries:

- ROUGE-1 measures unigram overlap.
- ROUGE-2 measures bigram overlap.
- ROUGE-L considers the longest common subsequence between the generated and reference summaries.

Together, these metrics provide a quantitative measure of the quality of the generated summary.

**BLEU:** Bilingual Evaluation Understudy, or BLEU, is another metric commonly used in natural language processing (NLP) tasks. Originally designed for machine translation, BLEU measures the overlap of n-grams (sequences of n words) between machine-generated text and reference texts, rewarding exact matches of phrases and words. However, its application in text summarization has been controversial.

While BLEU is effective in translation tasks where fidelity to the source text is essential, its reliance on surface-level similarities makes it less suitable for evaluating text summarization. Summarization often involves paraphrasing, rephrasing, or reorganizing information, and BLEU tends to penalize valid summaries that convey the same meaning but use different wording.

**BERTScore:** BERTScore is a metric developed to overcome some of the limitations of BLEU and ROUGE. It leverages the powerful BERT language model to compute similarity scores between pairs of sentences. While promising, BERTScore can be misleading without proper interpretation and explanation.

**Precision, Recall, F1-Score:** These are fundamental metrics in machine learning and information retrieval:

- Precision measures how many selected items are relevant.
- Recall measures how many relevant items are selected.
- F-score provides a single metric that balances both precision and recall.

In the context of text summarization, the above metrics can be used to measure the overlap between a generated summary and its reference summaries.

**Cosine similarity:** Cosine similarity is another useful tool in assessing the performance of a summarization task. This metric helps identify the cosine distance of the angle between two vectors, providing an understanding of the similarity between the summary and the original text based on embeddings. While the semantic meaning is captured, changes in structure could affect the overall score. That makes cosine similarity well-suited to extractive summarization. But for abstractive tasks, it may not be the best metric to use.

**Length:** Another metric to consider is the summarization length. Word count, character count, or even compressibility can be used here to determine if the summarization task is falling within what is expected based on the gold standard length. By putting the comparison lengths into a ratio you can determine how close the summarization task compares with the gold standard.

## Why Traditional Metrics Fall Short

Though the aforementioned evaluation metrics tend to perform well in providing quantitative measures of text summarization performance, they often fall short when it comes to capturing semantic meaning and overall quality of summaries.

For instance, ROUGE focuses on keyword matching and may not capture the overall coherence and relevance of a summary. Similarly, BLEU's emphasis on exact word matches can overlook valid summaries that paraphrase the source content. As for BERTScore, its lack of explanation can mislead us since we only get a numeric value, not the reasoning behind the value.

In light of these shortcomings, we need to explore more methods for evaluating text summarization performance. Specifically, ones that expand our key metrics so we have more flexibility and greater accuracy in capturing the subtleties of semantic meaning in a given summarization task. That makes LLM-as-a-Judge a logical choice to broaden our tool belt.

## PART V

# LLM-as-a-Judge Methodology

## Concept and Implementation

The LLM-as-a-Judge approach uses a large language model to evaluate the quality of generated text summaries.<sup>1</sup> This method leverages the semantic understanding capabilities of advanced LLMs to provide more nuanced and context-aware evaluations. We can do this by prompting an LLM to perform a pairwise comparison, single-answer grading or pointwise scoring system, or even reference-guided grading.<sup>2,3</sup>

At its heart, the LLM-as-a-Judge method works by passing inputs into the system to be evaluated based on a defined metric, then having the LLM perform the evaluation of that metric,<sup>4</sup> as illustrated in Figure 1 on the next page.

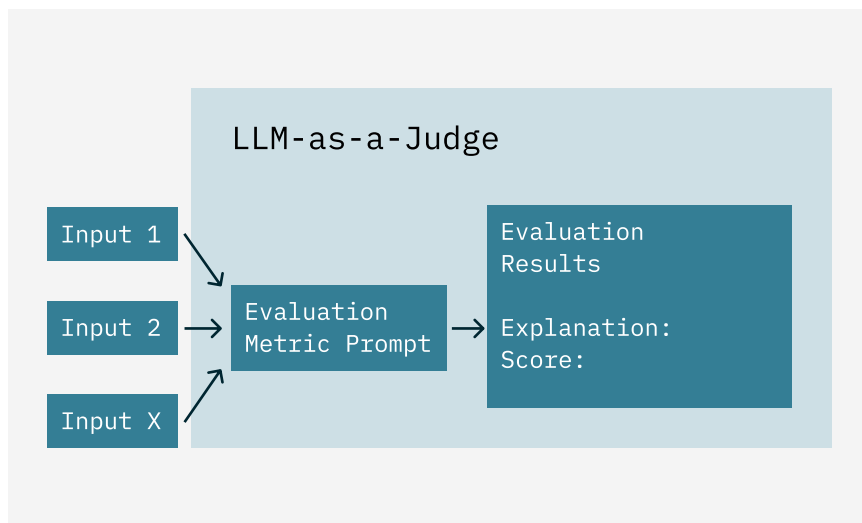


Figure 1. The idea behind using LLM-as-a-Judge: The system will take in various inputs and perform an evaluation based on a given metric prompt. The resulting evaluation and the output from the evaluation is known as the LLM-as-a-Judge process.

Here's another look at the LLM-as-a-Judge process written in Python.

```
python
```

```
1 def llm_as_judge(original_text, generated_summary,
2 evaluation_criterion):
3     prompt = f"""
4     Original Text: {original_text}
5     Generated Summary: {generated_summary}
6     Evaluate the summary based on the following
7     criterion:
8     {evaluation_criterion}
```

```
9     Provide a score from 1-5 and a detailed explanation
10    for each criterion.
11    """
12    response = llm_model.generate(prompt)
13    evaluation, score = parse_llm_response(response)
14    return evaluation, score
15
16    # Example usage
17    original = "...
18    summary = "...
19    criteria = ["Factual Accuracy", "Coherence",
20               "Relevance"]
21
22    results = {}
23    for criterion in criteria:
24        evaluation, score = llm_as_judge(original,
25                                        summary, criterion)
26        results[criterion] = {"evaluation": evaluation,
27                              "score": score}
```

In the context of summarization, LLM-as-a-Judge offers several benefits depending on the task at hand. By employing a binary or score metric (e.g., a score from 1-5 like in the code above), the LLM can assess the quality of the generated response against predefined criteria. Alternatively, we can enable the LLM to make a choice between its own output or our gold standard dataset, allowing for pairwise evaluation and ensuring a more objective assessment process.

The concept of a gold standard dataset plays an integral role in the evaluation process. The gold standard refers to a known output that serves as a benchmark for comparing the results of the LLM's output<sup>4</sup> –

in this case, a summarization task consisting of input comments and an output summary. Given an input and its corresponding gold standard output, the LLM can score different results based on predefined metrics. This approach enables us to not only understand the semantic meaning from inputs to outputs, but also to gauge key metrics to instill confidence in our LLM-generated results.

**By comparing the generated summary with the gold standard and providing an explanation along with a score, we gain valuable insights into the strengths and areas for improvement in our summarization system.** This helps us understand where our summarization tasks perform well and where they need improvement.

## Metrics for LLM-as-a-Judge

Evaluating the output of a system, especially in the context of summarization tasks, requires a variety of metrics to account for the many different dimensions that determine a generated summary's quality and relevance. Determining these quality metrics is often difficult since it depends on our task and our datasets. After passing in our inputs and getting our LLM-generated summary, we have to ask: Does the summary make sense? Is it relevant? Does it flow together well?

We answer these questions by choosing the appropriate metric(s) based on our use case. Balancing both performance and quality metrics is important when taking our summarization (or any LLM-powered application) into production.

Here are some key metrics often used when grading a summarization task:

- **Factual accuracy:** Assesses whether the individual inputs are accurately captured in an LLM's response or summary.

- **Omission:** Evaluates whether any of the individual inputs are left out of the summary.
- **Hallucinations:** Checks if there are additional misleading or incorrect details provided in the LLM response.
- **Relevancy:** Evaluates whether the summary includes only the most pertinent and important information from the source text, excluding irrelevant details.
- **Coherence:** Measures how logically connected or unified the summarized information is so the summary is easy to follow and understand, without disjointed ideas or abrupt transitions.
- **Informativeness:** Assesses how much new, useful, or insightful information the summary provides to the reader.
- **Fluency:** Considers whether the summary is written in proper, grammatically correct language that flows smoothly and is easy to read.
- **Completeness:** Measures whether the summary covers all the key points and arguments presented in the source text.
- **Conciseness:** Evaluates whether the summary conveys the information in a succinct manner, avoiding redundancy or unnecessary details.
- **Consistency:** Checks if the summary maintains the same perspective, tone, and level of detail as the source text.
- **Readability:** Measures how easy it is for a reader to comprehend the summary, considering factors like sentence complexity, word choice, and overall structure.
- **Structure and organization:** Examines how well the summary is organized, whether it follows a logical structure and sequence, and if it presents information in a way that makes sense to readers.
- **Politeness:** Assesses hard-to-measure qualities of conversation (e.g., politeness, empathy, compassion) whose multidimensional nature make their definition fluid based on context (e.g., industry, geography, culture). Learn more in "[How to Evaluate Conversational AI for Politeness.](#)"

- **Your metric here:** The flexibility of LLM-as-a-Judge allows you to define your own metrics for the given task (i.e., not just summarization). For instance, you could evaluate tone or style to improve consistency of brand voice.

Note this is not a complete list of summarization metrics. Moreover, these metrics can be used individually or in combination depending on the specific requirements of your evaluation task. Always return to guiding questions like, “What metrics are most important to our task? Which will help us optimize your system and give us confidence in our results?”

## Prompting for LLM-as-a-Judge

Always follow best practices when prompting LLM-as-a-Judge as an evaluator for text summarization. A good starting point is the established best practices for your specific LLM, such as:

- [OpenAI](#)
- [Anthropic](#)
- [Meta](#)

We can also start with a template or example prompt, then tweak the prompt to fit our metric(s) of interest, thereby creating our own evaluation metrics.

For more information and examples on prompting LLMs to behave as evaluators, check other WillowTree content such as “[LLM Evaluation Framework: How to Prevent Drift and System Degradation.](#)”

## PART VI

# Results

## Comparing Gold Standard Dataset Outputs to LLM-Generated Summaries

Using the [Samsun](#) conversational dataset\* from Kaggle, we can showcase an example summarization task where we distill the ideas of a conversation and run an evaluation. After building out our dataset to map each query to a gold standard summary, we can use our LLM-as-a-Judge technique with pointwise grading – plus a point system similar to a Likert scale – to tell us how well the LLM-generated summary compares to our gold standard in regards to factual accuracy. We can also capture the key metrics from traditional machine learning approaches and compare the metrics to LLM-as-a-Judge.

Figure 2 on the next page visualizes passing inputs into the system for evaluation, followed by Figure 3 where the LLM-as-a-Judge scores

---

\*Note some examples in this dataset were flagged as inappropriate due to sexual content by our AI platform, [Fuel iX](#), and therefore were not included.

and explains the discrepancy between our gold standard and LLM-generated summary using BERTScore, BLEU, and ROUGE.

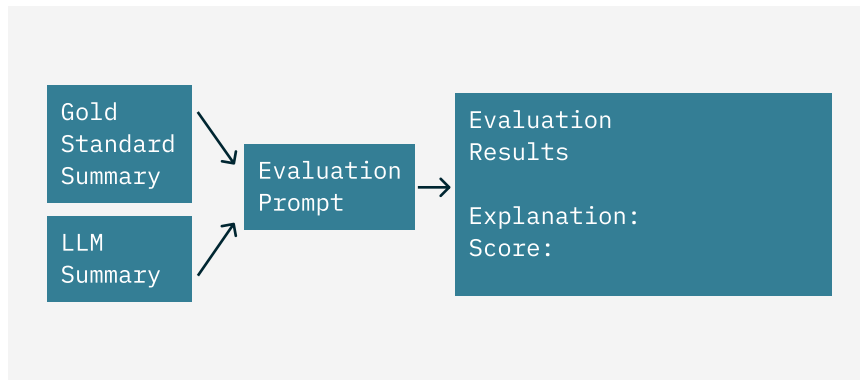


Figure 2. A gold standard summary and LLM-generated summary being passed in as inputs into the metric evaluation prompt.

Gold Standard Summary	AI-Generated Summary	LLM-as-a-Judge Explanation and Score				
Steve will order the second display. He needs to make sure he can connect his card to it.	Steve and Bill discussed two display models, with Bill advising against the first due to bad reviews and recommending the second one, which Steve decided to order.	Explanation:\n\nThe AI-generated summary captures the main points of the conversation between Steve and Bill regarding the two display models. It correctly notes that Bill advised against the first display due to bad reviews and recommended the second one, which Steve decided to order. However, the AI-generated summary misses an important detail from the source comment: Steve needs to ensure he can connect his card to the second display. This omission affects the completeness of the summary.\n\nScore: 3				
LLM-as-a-Judge Normalized Score	BERTScore	BLEU Score	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
0.75	0.879628	0	0.311111	0.046512	0.177778	0.177778

Figure 3. Results from an LLM-as-a-Judge factual accuracy evaluation compared to BERTScore, BLEU, and ROUGE.

Figure 3 shows how having our LLM-as-a-Judge not only score, but also explain the discrepancy between the gold standard and LLM-generated summary is a huge win. Comparing these results to BERTScore highlights

the benefits of having an explanation paired with the resulting score.

**BERTScore:**

0.879628

**LLM-as-a-Judge:**

Explanation: The AI-generated summary captures the main points of the conversation between Steve and Bill regarding the two display models. It correctly notes that Bill advised against the first display due to bad reviews and recommended the second one, which Steve decided to order. However, the AI-generated summary misses an important detail from the source comment: Steve needs to ensure he can connect his card to the second display. This omission affects the completeness of the summary.

Score: 3

Looking at just the BERTScore alone might lead us to believe the summarization results are fairly accurate, and that the score doesn't contain insights for improving the prompt or making changes to our summarization task approach. But with the LLM-as-a-Judge insights, we can incorporate the explanations into our summarization prompt, making our generated responses more inline with the gold standard. Likewise, we can also use these explanations to launch a line inquiry. Do the evaluation results raise an issue with our dataset and/or what we expect to be a gold standard? Is there an issue with our evaluation prompt, or were the generated results closer to what was expected for the task? What does our subject matter expert (SME) think about the two?

While all of these metrics have their place in NLP tasks, given our text summarization evaluation, we need to decide what will provide the best insights for prompt optimization and making our summary task better. Running the LLM-as-a-Judge evaluations over the dataset, we can see that our summarization prompt needs some work. **With actionable**

insights into where the LLM-as-a-Judge scores are low, we can now optimize and make our summarization prompt even better. Coupled with prompt versioning and tracking, we can see how our prompting affects our scores over time with the aim of delivering a more optimized summarization task fit to our use case.

## Comparing Dataset Inputs to LLM-Generated Summaries

When reference summaries are not available, such as in cases of novel document types or real-time summarization, the evaluation process becomes significantly more challenging. However, certain LLM-as-a-Judge strategies can help us assess the quality of summarization tasks in these scenarios.

Using the same dataset as above, we can compare the inputs of our dataset to our LLM-generated summary based on the metrics key to our task. Choosing the correct metrics will of course be dependent on the summarization task, but we will show below in Figures 4 and 5 an example for possible outputs when checking for 1) factual accuracy and 2) omissions while raising key questions to consider when looking at the results.

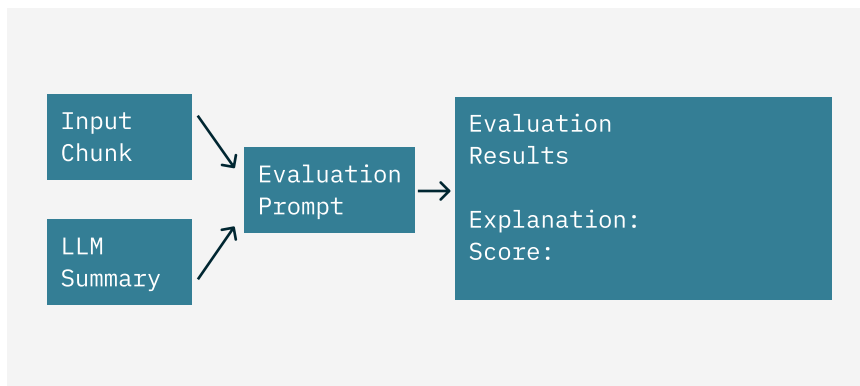


Figure 4 (bottom left of page). When gold standard reference summaries are not available, we can still compare the input chunks from our knowledge store to our LLM summary. Here, we can take in the chunks making up our summary and evaluate against specific criteria.

Inputs	LLM-Generated Summary	Factual Accuracy	Factual Accuracy Score
[Steve: hi, what do you think about this display?, Steve: <file_other>, Bill: hi! let me check, Bill: never had the chance to use it personally but the parameters looks fine, Bill: just remember that if you have a GeForce card it will not work with FreeSync, Steve: is that a problem?, Bill: it will work as a display, you just won't be able to activate this technology, Bill: I did found something more worrying , Bill: there's lots of reviews about frequent bad pixels in this model and issues with the power supply, Steve: ok, thanks for the info, Steve: I found another one, Steve: <file_other>, Bill: the reviews for this one look much better and the parameters are similar to the previous one, Bill: the only reported issue are faded colors but that's always a subjective matter, Bill: from those two I would definitely choose the second one, Steve: ok, great :), Steve: I'll order it, Bill: I don't remember what card you have at the moment, Bill: so just make sure you'll be able to connect it properly, Steve: I will, Steve: thanks for your help, Bill: no problem]	Steve and Bill discussed two display models, with Bill advising against the first due to bad reviews and recommending the second one, which Steve decided to order. \n\nUpon reviewing the Source Comments, the following points are noted:\n1. Steve asked Bill about the first display model.\n2. Bill checked and mentioned that while the parameters looked fine, there were bad reviews about frequent bad pixels and power supply issues.\n3. Steve acknowledged the information and found another display model.\n4. Bill reviewed the second model and found the reviews to be better, with only a subjective issue of faded colors.\n5. Bill recommended the second model over the first.\n6. Steve decided to order the second model.\n\nThe AI-generated summary accurately captures the essence of the conversation:\n- It mentions the discussion of two display models.\n- It notes Bill's advice against the first model due to bad reviews.\n- It includes Bill's recommendation of the second model.\n- It concludes with Steve's decision to order the second model.\n\nThe summary is factually accurate and includes the key points from the Source Comments.\n\nScore: 1	The AI-generated summary states that Steve and Bill discussed two display models, with Bill advising against the first due to bad reviews and recommending the second one, which Steve decided to order. \n\nUpon reviewing the Source Comments, the following points are noted:\n1. Steve asked Bill about the first display model.\n2. Bill checked and mentioned that while the parameters looked fine, there were bad reviews about frequent bad pixels and power supply issues.\n3. Steve acknowledged the information and found another display model.\n4. Bill reviewed the second model and found the reviews to be better, with only a subjective issue of faded colors.\n5. Bill recommended the second model over the first.\n6. Steve decided to order the second model.\n\nThe AI-generated summary accurately captures the essence of the conversation:\n- It mentions the discussion of two display models.\n- It notes Bill's advice against the first model due to bad reviews.\n- It includes Bill's recommendation of the second model.\n- It concludes with Steve's decision to order the second model.\n\nThe summary is factually accurate and includes the key points from the Source Comments.\n\nScore: 1	1

Figure 5. The inputs and the LLM-generated summary were passed into the factual accuracy evaluation prompt and returned the evaluation explanation along with a score of 1, indicating that the two summaries were factually accurate.

When faced with these results and their explanation, we can ask, "Is the summary what was expected? Do we need to update our evaluations or do we need to update our summarization task?" Pairing these metrics with our gold standard metrics offers a good range of evaluations to make actionable prompt optimizations.

Our insights deepen when we pair the results of our factual accuracy evaluation with those of our omission evaluation, shown in Figure 6 below.

Inputs	LLM-Generated Summary	Omission	Omission Score
[Steve: hi, what do you think about this display?, Steve: <file_other>, Bill: hi! let me check, Bill: never had the chance to use it personally but the parameters looks fine, Bill: just remember that if you have a GeForce card it will not work with FreeSync, Steve: is that a problem?, Bill: it will work as a display, you just won't be able to activate this technology, Bill: I did found something more worrying , Bill: there's lots of reviews about frequent bad pixels in this model and issues with the power supply, Steve: ok, thanks for the info, Steve: I found another one, Steve: <file_other>, Bill: the reviews for this one look much better and the parameters are similar to the previous one, Bill: the only reported issue are faded colors but that's always a subjective matter, Bill: from those two I would definitely choose the second one, Steve: ok, great :), Steve: I'll order it, Bill: I don't remember what card you have at the moment, Bill: so just make sure you'll be able to connect it properly, Steve: I will, Steve: thanks for your help, Bill: no problem]	Steve and Bill discussed two display models, with Bill advising against the first due to bad reviews and recommending the second one, which Steve decided to order.	Explanation:\nThe AI-generated summary captures the essence of the conversation between Steve and Bill regarding two display models. However, it omits several key details from the Source Comment:\n1. Bill's initial check and comments on the first display's parameters and compatibility with GeForce cards.\n2. The specific issues Bill found with the first display, such as bad pixels and power supply problems.\n3. The detailed feedback on the second display, including the better reviews and the issue of faded colors.\n4. Bill's reminder to Steve about ensuring compatibility with his current card.\n\nWhile the summary does capture the main decision points, it leaves out significant details that were part of the conversation.\n\nScore: 1	1

Figure 6. The inputs and the LLM-generated summary were passed into the omission evaluation prompt and returned the evaluation explanation along with a score of 1, indicating that omissions were found.

Now we have more context for optimizing the evaluation prompt. Is omission the correct metric to use for the summarization task? Would relevancy work better here? Does the summarization task need additional metrics? What insights would we hope to gain from the results, and how would those metrics lead to fixing problems and providing a better experience?

## Human Feedback

While automation and machine learning play pivotal roles in the evaluation process, the importance of human feedback cannot be overstated. Due to the inherent subjectivity of summarization, gaining insights from SMEs for key tasks is crucial. Therefore, having a human SME evaluate the results is extremely important in instilling trust into the system.

A/B testing is one effective way of incorporating human feedback into the evaluation process. This method involves presenting two versions of a summary to SMEs and asking them to choose the superior one, similar to having the LLM picking the best summary in a pairwise comparison.<sup>1</sup> Their opinion on how well the task is performing is invaluable in refining the summarization prompts used in pipelines.

**In essence, while technology can provide the tools to generate and evaluate summaries, human feedback remains vital in optimizing these tasks.**

Another approach would be using human feedback as the point of optimization (i.e., direct preference optimization). In this case, a human is given two choices of summaries and asked to choose the better one. Based on this feedback, we can optimize the system by using the chosen summary as a means to better the output.

## Challenges to Evaluation

It's important to reiterate that while using an LLM-as-a-Judge greatly improves our accuracy and efficiency compared to relying on traditional metrics, evaluating text summarization is still too difficult and complex to run as an automated, hands-free operation with little to no oversight. This is chiefly due to challenges such as:

- **Reference comparison:** Without a clear reference to compare the resulting summary against, it can be difficult to validate the summarization's accuracy and relevance.
- **Task dependency:** The effectiveness of a summarization system is highly dependent on the specific task it is designed for.
- **Metric determination:** Ensuring we are evaluating the summarization task correctly with the proper metrics that are actionable.
- **Human feedback:** While LLMs can generate metrics, coupling these results with feedback from SMEs provides the best gauge of our summarization task's success.

To keep these challenges under control, ask yourself these key questions when reviewing your LLM-as-a-Judge metric evaluations:

- Are the results pointing to the summarization task failing?
- Is our evaluation metric passing or failing for the correct reason?
- Does our dataset need to be updated?
- What metrics does our team care about and how will the evaluations impact the ability to diagnose and optimize the summarization process?

## PART VII

# Practical Considerations for Real-World Applications

## Evaluation Tradeoffs

When considering the real-world application of text summarization models, weigh the pros and cons between automated and human evaluations. Because LLMs are nondeterministic by nature (i.e., they don't always produce the same output from the same input) and because success changes based on use case (ditto for system vulnerabilities), it's important to ensure robust and responsible AI practices. In particular, proper guardrails and a governance framework are needed to guide [continuous benchmarking and evaluation](#).

For instance, there may be some summarization tasks where human evaluations surpass automated ones in terms of accuracy. After all, humans better understand context and semantics, which is crucial for summarizing complex or nuanced texts. Therefore, an optimal text summarization model might involve a combination of both automated and human evaluations.

Research by Chiang et al.<sup>3</sup> found there are four key benefits to using LLMs as evaluation tools:

- LLMs enhance reproducibility (as compared to humans).
- Each evaluation is independent of the other evaluation methods.
- LLMs are more efficient (i.e., cheaper and faster than human evaluation).
- Human exposure to objectionable datasets during testing is minimized.

## Challenges in Specific Domains

Text summarization applications in specialized fields like legal services and healthcare often face unique challenges. Legal text summarization, for instance, must accurately encapsulate intricate legal terminologies and concepts. Similarly, medical text summarization requires precise interpretation of complex medical terms.

In these scenarios, expert review becomes crucial to verify the accuracy and relevancy of the generated summaries. Moreover, the issue of [responsible AI](#) becomes critical – what if a summarization model makes an error in a critical domain? The implications could be severe, highlighting the need for robust evaluation and checks.

## PART VIII

# Conclusion

---

The evaluation of text summarization systems is a multifaceted challenge, one that requires a blend of traditional metrics, advanced techniques like LLM-as-a-Judge, and human feedback. While traditional metrics such as ROUGE and BLEU offer quantitative measures, they often fall short in capturing the nuanced quality of summaries. LLM-as-a-Judge provides a more comprehensive evaluation framework by leveraging the semantic understanding of large language models and offering detailed explanations alongside scores, thus enabling more actionable insights for prompt optimization and system improvement.

Human feedback remains an indispensable component of the evaluation process, especially in specialized fields like legal services and healthcare, where accuracy and context are paramount. Incorporating expert reviews and A/B testing ensures that summarization systems are not only accurate, but also contextually relevant and reliable. As the field evolves, the integration of both automated and human evaluations will be crucial for developing robust and trustworthy summarization models.

In conclusion, the complexities of text summarization evaluation necessitate a thoughtful approach that combines various metrics and methodologies. By leveraging the strengths of LLM-as-a-Judge and incorporating human expertise, we can achieve a more holistic and reliable evaluation process. This balanced approach will pave the way for more effective and contextually accurate summarization systems, ultimately enhancing their utility across diverse applications.

If you're ready to develop a proof of concept for your own summarization system, or want to refine the evaluation process for your current summaries, the Data & AI Research Team (DART) at WillowTree can help you do so in a matter of weeks. [Learn more about our approach to Generative AI.](#)

# References

---

- 1 [Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena](#)
- 2 [On the Limitations of Fine-Tuned Judge Models for LLM Evaluation](#)
- 3 [Can Large Language Models Be an Alternative to Human Evaluations](#)
- 4 [LLM Evaluation Framework: How to Prevent Drift and System Degradation](#)

## Connect with WillowTree's Data & AI Research Team

### Gagan Tandon

Chief Data and AI Officer  
WillowTree, a TELUS Digital Company



### Conner Brew

Director, Data & AI Delivery  
WillowTree, a TELUS Digital Company





WILLOWTREE®  
a TELUS Digital Company