SummReason: Evaluating Large Language Model Based Metrics in Reasoning and Explaining Factual Inconsistent Summaries

Anonymous ACL submission

Abstract

Summarization is an important application of large language models (LLMs). One prominent aspect in summarization evaluation is factual consistency. The recent raise of using LLM to evaluate the factual consistency of machine-generated summaries has an advantage on interpretability by giving reasons to justify why a summary is inconsistent to its source. However, while LLM-based evaluation can generally provide a sounding explanation, the corresponding reasoning may not be faithful. The unfaithful reasoning may also lead to the wrong final assessment at the end when applying a Chain-of-Thought prompt. To evaluate the faithfulness of reasoning and explanations given by LLMs, we propose a benchmark called SummReason comprising LLM-generated post-hoc reasoning reviewed and corrected by human annotators regarding the inconsistency of machine-generated summaries. We then develop an automatic metric $CoTscore$ to measure the recall of an LLM-generated reasoning on summary factual consistency using human annotated explanations as references. Our benchmark reveals the gaps in reasoning between the open-source LLMs vs. GPT-4, CoT reasoning vs. post-hoc explanation, as well as human vs. LLMs. We hope our benchmark advances the evaluation of LLM in reasoning ability.

1 Introduction

Factual consistency checking in summarization gauges whether the information in a machine-generated summary is supported by its source document(s). It has rapidly emerged in the past couple of years because abstractive summarization systems may generate hallucinated content (Kryscinski et al., 2020), which reduces their reliability. To detect the inconsistency, traditional methods (Koh et al., 2022) use question answering systems or entailment classification models. Recent studies introduce a new method that leverages Large Language Models (LLMs) as natural language generation evaluators.

With a massive mount of knowledge gained from the pretraining stages, LLMs can outperform crowd-workers (Gilardi et al., 2023) on many text-annotation tasks. For LLM-based consistency evaluation (Gekhman et al., 2023; Luo et al., 2023; Gao et al., 2023; Liu et al., 2023; Fu et al., 2023), prompts are crafted to instruct the LLM acting as an annotator to judge the factuality of machine-generated summaries.

LLM-based factual consistency evaluation approaches can be categorized into two kinds based on the prompting strategy: 1. *Direct assessment* where an LLM directly gives the Likert-score or binary judgment without any rationale. 2. *Chain-of-Thoughts (CoT)* where an LLM is asked to generate step-by-step reasoning before yielding the final assessment. A CoT prompt allows an LLM to use extra space to expand complicated reasoning, thus providing extra context and guidance to help with the evaluation decision.

The CoT approach has been found to be superior than direct assessment with a higher correlation with human judgement or a better accuracy (Luo et al., 2023; Liu et al., 2023). Another advantage of using CoT-style prompts is that the generated reasoning can serve as explanations, therefore making the consistency metrics self-explainable. However, the generated reasoning may be unfaithful (Lanham et al., 2023) and may mislead an LLM to reach a wrong conclusion. Luo et al. (2023) show that ChatGPT may conduct false inferences, and ChatGPT using CoT prompts performs poorly on identifying the inconsistent summaries despite the high recall for consistency summaries.

In this work, we curated an evaluation dataset SummReason to benchmarking how faithful the LLMs are as factual consistency evaluators for summarization. In summary:

- We release the SummReason evaluation dataset with human annotations. We prompt an LLM to generate post-hoc explanations with factually inconsistent summaries from the SummaC (Laban et al., 2022) benchmark. Human annotators are guided to correct generated unfaithful explanations.
- We develop an automatic metric $CoTscore$ to evaluate LLM's reasoning trace using the human annotated explanations as references.
- We evaluate several well known LLMs using our data and metrics. Our benchmark reveals the gaps in reasoning between the open-source LLMs vs. GPT-4, CoT reasoning vs. post-hoc explanation, as well as human vs. LLMs.

Our code and data is publicly available at `https://anonymous.4open.science/r/SummReason-CB59/benchmark/readme.md`.

2 SummReason Dataset Creation

In this study, we regard the consistency detection of summarization as a binary classification, taking an article and a machine-generated summary as the input. The summary is considered consistent only if the evidence can be found in the article that supports the summary. This section details the steps in the curation of the **SummReason** evaluation dataset: 1. collecting LLM's reasoning with CoT prompt for consistency detection; 2. generating the post-hoc explanations from existing datasets with human annotated labels; 3. human review and correction of the explanations.

2.1 LLM Reasoning Collection

Following a previous work (Luo et al., 2023), we use a similar **CoT style prompt** to collect LLM's reasoning trace for summarization factual consistency checking:

> *Decide if the following summary is consistent with the corresponding article. Note that consistency means all information in the summary is supported by the article.*
>
> *<Article></Article>*
>
> *<Summary></Summary>*
>
> *Explain your reasoning step by step first, and then answer (yes or no) the question in the end:*

After collecting the LLM's output, the next step is to obtain the gold reference reasoning to compare LLM's reasoning with.

2.2 Referential Explanation Generation

Since asking humans to write the rationale from scratch for consistency checking is extremely laborious, we generate the gold reference in a semi-auto fashion where human annotators do post-editing on LLM's explanations. Instead of directly using CoT-style prompt where the reasoning trace could mismatch the ground truth label, we designed a label-first style prompt to elicit the post-hoc reasoning. Specifically, the consistency labels (in red below) of the input samples are explicitly told in the prompt so that the reasoning of LLM should match the ground truth. For the inconsistent summaries, the LLM is instructed to give a rationale for how some part of the summary is not supported by the article, the **post-hoc prompt** template is shown below:

> *Note that consistency means all information in the summary is supported by the article. It's known that the following summary is not consistent with the article. Find out why.*
>
> *<Article></Article>*
>
> *<Summary></Summary>*
>
> *Explain your reasoning step by step:*

For a consistent summary, it is difficult to tell an explicit reason why a summary is supported by its article. In addition, Luo et al. (2023) found that ChatGPT can successfully recognize more than 95% consistent summaries. Thus, we believe recognizing consistent summary is relatively an easy task, so we ignore the consistent case and **focus on reasoning for inconsistent summaries**.

Using the post-hoc style prompt template above and an existing summarization factual consistency dataset, we can generate the explanations using an LLM (`Llama-2-13b-chat-hf` (Touvron et al., 2023) in this work). In this study, we sampled 1,323 inconsistent summary and document pairs from the SummaC summarization benchmark's validation split to form our evaluation data.

2.3 Human Annotation

Containing the ground truth label in the prompt does not guarantee the faithfulness of the explanation generated by the LLM. Therefore, we incorporate human annotators to examine the LLM-generated explanation and post-edit if the explanation is not faithful.

The human annotators are given the typology of factual errors from the FRANK (Pagnoni et al., 2021) dataset to understand what kind of issues in the summary are considered as consistency error. Frank's typology provides 8 categories of factual consistency errors: Predicate Error, Entity Error, Circumstance Error, Coreference Error, Discourse Link Error, Out of Article Error, Grammatical Error and Other Error.

Then, the human annotators are given examples and guidelines on how to edit the inaccurate explanations. Following the guidelines, annotators are asked to organize each explanation in a numbered list. The annotations are done by the authors of this work, who are computer science undergraduate or graduate students with natural language processing background. Details of the annotation guidelines and examples can be found in Appendix A.

3 Evaluation Metric: CoTscore

Collecting human ratings for LLM's reasoning can be laborious, we aim to create an automatic metric to evaluate the faithfulness of reasoning steps. Reference-based metrics like ROUGE (Lin, 2004) and BertScore (Zhang et al., 2019) rely on lexical or embedding similarity. They may not be suitable to evaluate the logical reasoning between texts because the reasoning trace could contain many citation texts from the summary and article, making two logically different reasoning traces lexically or semantically similar.

Inspired by MT-bench (Zheng et al., 2023) and AlpacaEval (Dubois et al., 2023), two famous LLM benchmarks, we also introduce LLM-as-a-judge in our work. Unlike the two LLM judges of the two benchmarks above, we do not ask the judge LLM to give a rating or preference. Instead, we develop a fine-grained rating process. We set the rating policy as such: each bullet point in the gold reference is worth 1 point and LLM's reasoning gets 1 point if it covers or mentions one bullet point, and the final score for one sample is $CoTscore = \frac{\#covered_points}{\#total_points}$. The $CoTscore$ is essentially a recall score for measuring how LLM's reasoning covers human explanation. The prompt used for the LLM judge is:

> *You are an assessor to give judgment on a reasoning problem. Here is the text to be assessed:*
>
> *<text></text>*
>
> *Does the above text mention or contain the following reference reasoning step:*
>
> *<reference></reference>*
>
> *Answer (yes or no):*

# of Sample	Total	1323
	-XSUM	600
	-CNNDM	723
# of Article Tokens	Min	66
	Max	1978
	Avg.	707
# of Summary Tokens	Min	8
	Max	189
	Avg.	46
# of Referential Explanation Tokens	Min	42
	Max	327
	Avg.	141

Table 1: Dataset statistics of the SummReason. Article sources are from the CNNDM (See et al., 2017) and XSum (Narayan et al., 2018) datasets.

To achieve the rating process mentioned above, we iteratively feed each bullet point of the referential explanation to the LLM-judger to obtain the $CoTscore$. Note that only the LLM's output reasoning trace is fed to the <text>, and no summary or article is involved during the assessment.

4 Experiments

4.1 Evaluation Settings

The following LLMs are used for benchmarking:

- **GPT-3.5/4 Turbo** (OpenAI, 2023), `gpt-3.5-turbo-0301` and `gpt-4-1106-preview` specifically.
- **Llama-2-7B/13B/70B** (Touvron et al., 2023), the `chat-hf` version.
- **Mistral-7B-OpenOrca** (Lian et al., 2023), Mistral-7B (Jiang et al., 2023) model fine-tuned using OpenOrca (Mukherjee et al., 2023) dataset.

For the reproducibility of experiment results, we set $temperature = 0$ for the generation process while the number of generated tokens is limited in $max_tokens = 500$. We evaluate the above models on the SummReason dataset using the CoT style prompt and post-hoc prompt discussed in Sections 2.1 & 2.2. After taking cost and quality trade-off, the `gpt-3.5-turbo-0301` model is used as the LLM judge. The data statistics for our SummReason is shown in Table 1.

LLM	Specificity (%)	Average CoTscore (%)					
		True Negative		False Positive		Overall	
		CoT	Post-hoc	CoT	Post-hoc	CoT	Post-hoc
GPT-4-Turbo	74.98	75.72	83.97	47.50	70.93	68.66	80.71
GPT-3-Turbo	72.26	63.34	75.98	40.14	64.68	56.90	72.12
Llama-2-70B	53.36	61.28	77.87	34.41	70.66	48.75	74.51
Llama-2-13B	86.92	48.50	69.19	34.16	61.58	46.62	68.19
Llama-2-7B	6.27	64.98	76.43	25.34	70.86	27.83	71.20
Mistral-7B-OpenOrca	64.40	60.99	69.50	42.83	62.47	54.52	67.00

Table 2: Benchmark results. $CoTscore$ is calculated within each of LLM's classification categories. "CoT" and "Post-hoc" represents the two types of prompt template discussed in Section 2.2.

4.2 Results

The benchmarking results can be found in Table 2. The **specificity** is calculated based on the classification results of LLMs using the CoT-style prompt, it measures the percentage of inconsistent summaries retrieved by the model. To measure the LLM's reasoning and explanation ability, we calculated the $CoTscore$ for the LLM using CoT-style prompt and post-hoc prompt respectively. The $CoTscore$ is also presented in each model's classification category based on CoT classification results, where **True Negative (TN)** represents the case when inconsistent summaries are caught by an LLM and **False Positive (FP)** represents the case when inconsistent summaries are missed by an LLM.

4.3 Discussion

Based on the specificity results, the three most proficient LLMs in identifying inconsistent summaries are Llama-2-13B, GPT-4 Turbo, and GPT-3.5-Turbo. It is essential to recognize that a high specificity alone does not necessarily qualify a model as superior. This observation stems from the fact that our benchmark exclusively addresses inconsistent cases while overlooking consistent instances. Despite Llama-2-13B exhibiting the highest specificity, its $CoTscore$ in the TN classification category is comparatively low when subjected to the CoT prompt. This discrepancy signals that the model possesses a relatively significant proportion of unfaithful reasoning. Consequently, the presence of CoT reasoning with a correct final answer does not automatically denote the faithfulness of the entire reasoning process.

While GPT-4-Turbo and GPT-3.5-Turbo may lag behind Llama-2-13B in terms of specificity, their overall $CoTscores$, particularly when utilizing a CoT-style prompt, emerge as the most commendable. Notably, the GPT-4-Turbo model attains the highest overall CoTscore at 68.66%, underscoring that the LLM still exhibits a discernible gap in achieving human-level reasoning. Furthermore, our experiment reveals that the CoT reasoning prowess of OpenAI's GPT models surpasses that of the tested open-source models. Among the open-source models scrutinized, Mistral-7B, fine-tuned with Orca-style datasets, demonstrates the closest $CoTscore$ and, consequently, the most comparable CoT reasoning ability to the GPT models. This finding reinforces the notion that applying Orca-style fine-tuning enables smaller LLMs to benefit from learning the reasoning traces of GPT-4.

Regarding the different reasoning generated through the CoT-style prompt and the post-hoc prompt, it is not surprising that the $CoTscore$ is higher when utilizing the post-hoc prompt compared to the CoT prompt. This discrepancy can be attributed to the predetermined final answer provided in the post-hoc prompt. In contrast to CoT reasoning, employing the post-hoc prompt results in a substantial increase in $CoTscores$ across all LLMs, particularly in the FP category and to a lesser extent in the TN category. We hypothesize that this divergence stems from the post-hoc prompt motivating LLMs to seek additional evidence for the thoroughness of their responses, while CoT reasoning may conclude prematurely upon identifying a single compelling piece of evidence.

5 Conclusion

In this work, we introduce a challenging benchmark **SummReason** for reasoning or explaining factual inconsistency in summarization. The benchmark contains a dataset with human edited referential explanations and an automatic metric $CoTscore$ to judge a LLM's reasoning trace. We hope our work can be a valuable resource for testing LLM's ability in text reasoning.

Limitation

For evaluating natural generation tasks with references, the quality of the reference texts can affect the evaluation quality. The SummReason benchmark is created with human annotators doing post-editing on LLM's reasoning. In our task, there may be multiple reasons why a summary is not consistent, it is possible our referential explanation does not cover all the reasons. The problem can be mitigated by creating multiple references with different annotators writing reason for the same sample. We would consider the option in the next version of the benchmark if time and budget allow.

Our evaluation has an underlying assumption that if the LLM's reasoning covers human's explanation, then the reasoning entails that the evaluating summary is inconsistent. However, the assumption does not guarantee that the reasoning trace is precise. The LLM's reasoning can contain irrelevant or unnecessary reasoning steps but the $CoTscore$ ignores them. In other words, the $CoTscore$ only measures the recall but ignores precision. Despite the limit, we believe one could evaluate the precision easily by altering the LLM judger's prompt and following a similar process.

The automatic evaluation metric relies on a LLM as the judge to give a binary judgment on the coverage between the output text and the reference. Thus, the quality of the assessment is highly depended on the backbone LLM. We are also interested in evaluating how different backbone LLMs perform in judging the reasoning trace for future work.

References

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback.

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

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt.

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

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30).

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.

Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2022. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM Comput. Surv.*, 55(8).

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

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.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. Measuring faithfulness in chain-of-thought reasoning.

Wing Lian, Bleys Goodson, Guan Wang, Eugene Pentland, Austin Cook, Chanvichet Vong, and "Teknium". 2023. Mistralorca: Mistral-7b model instruct-tuned on filtered openorcav1 gpt-4 dataset. `https://huggingface.co/Open-Orca/Mistral-7B-OpenOrca`,.

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.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization.

Subhabrata Mukherjee, Arindam Mitra, Ganesh Jawahar, Sahaj Agarwal, Hamid Palangi, and Ahmed

Awadallah. 2023. Orca: Progressive learning from complex explanation traces of gpt-4.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829, Online. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. 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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, 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.

A Appendix

A.1 Guidelines of annotations

We gave the following guidelines to the annotators on how to edit the unfaithful explanations:

1. Put your annotation after "Corrected:". Even the reasoning is correct, still copy the steps under "Corrected:".

2. Delete incorrect reasoning and keep the **correct** reasoning.

3. Delete irrelevant sentences and **only** keep the reasoning, e.g. "Sure! Here's why the summary is not consistent with the article:" should be deleted.

4. Unify the format as an ordered list, following the format provided as Figure 1.

5. Add your own reasoning if necessary. Keep the reasoning **short** and **simple**.

In addition to the guide, we also provide example annotations to the annotators as in Figure 2 3 4 and 5.

1.<what the summary states>
2.<what the article actually states>
3.<point out where is the inconsistency>
4.<optionally more steps>
5.<.....>

Figure 1: Annotation Format

Note that consistency means all information in the summary is supported by the article.
It's known that the following summary is not consistent with the article.
Find out why.

<Article>
Mr Katter said the Government believes Mr Gordon would quit after he was recently accused of domestic violence.
</Article>

<Summary>
Mr Katter said he would quit after he was accused of domestic violence.
</Summary>

Explain your reasoning step by step:
1. The summary states that Mr Katter would quit after being accused of domestic violence.
2. However, the article states that the Government believes Mr Gordon would quit after being recently accused of domestic violence.
3. The inconsistency is due to the use of the word'recently'. The article implies that the Government believes Mr Gordon would quit after being recently accused, while the summary states that Mr Katter would quit after being accused.
4. To maintain consistency, the summary should use the same word as the article,'recently'.

Corrected:
1. The summary states that Mr Katter would quit after being accused of domestic violence.
2. However, the article states that the Government believes Mr Gordon would quit after being recently accused of domestic violence.
3. The inconsistency is due to the error that the article implies that the Government believes Mr Gordon would quit, while the summary states that Mr Katter would quit.

Figure 2: Annotation Example 1

Note that consistency means all information in the summary is supported by the article.
It's known that the following summary is not consistent with the article.
Find out why.

<Article>
Barcelona club president Josep Maria Bartomeu has insisted that the La Liga leaders have no plans to replace Luis Enrique and they're 'very happy' with him.
</Article>

<Summary>
Josep Maria Bartomeu says the La Liga leaders are very happy with him.
</Summary>

Explain your reasoning step by step:
1. The summary claims that Bartomeu is happy with himself, which contradicts the article's claim that the La Liga leaders are very happy with him.
2. The article provides evidence for the claim that the La Liga leaders are very happy with him, while the summary does not.
3. Therefore, the summary is not consistent with the article.

Corrected:
1. The summary claims that Bartomeu says the La Liga leaders are very happy with him.
2. The article provides evidence for the claim that the La Liga leaders have no plans to replace Luis Enrique and they're 'very happy' with him.
3. Therefore, the summary is not consistent with article on that the summary implies the La Liga leaders are very happy with Bartomeu, but the article implies the La Liga leaders are very happy with **Luis Enrique**.

Figure 3: Annotation Example 2

Note that consistency means all information in the summary is supported by the article.
It's known that the following summary is not consistent with the article.
Find out why.

<Article>
Goldfish are being caught weighing up to 2kg and koi carp up to 8kg and one metre in length.
</Article>

<Summary>
Goldfish are being caught weighing up to 8kg and one metre in length.
</Summary>

Explain your reasoning step by step:
1. The article states that goldfish can weigh up to 8kg, while the summary states that koi carp can weigh up to 8kg.
2. The article also states that koi carp can reach a length of 1 metre, while the summary states that goldfish can reach a length of up to 8cm.
3. Therefore, the summary is not consistent with the article.

Corrected:
1. The summary states that **goldfish** can weigh up to 8kg, while the article states that **koi carp** can weigh up to 8kg.

Figure 4: Annotation Example 3

Note that consistency means all information in the summary is supported by the article.
It's known that the following summary is not consistent with the article.
Find out why.

<Article>
Paul Merson had another dig at Andros Townsend after his appearance for Tottenham against Burnley ...Townsend hit back at Merson on Twitter after scoring for England against Italy.
</Article>

<Summary>
Paul Merson had another dig at andros townsend after scoring for England against Italy.
</Summary>

Explain your reasoning step by step:
1. The summary states that Paul Merson had another dig at Andros Townsend after scoring for England against Italy.
2. However, the article mentions that Andros Townsend was not even in the match against Italy.
3. Therefore, the summary is not consistent with the article.

Corrected:
1. The summary states that Paul Merson had another dig at Andros Townsend after scoring for England against Italy.
2. However, the article mentions that Paul Merson had another dig at Andros Townsend after his appearance for Tottenham against Burnley.
3. Therefore, the summary is not consistent with the article on the time of event where the summary states **after scoring for England against Italy**, while the article states **after his appearance for Tottenham against Burnley**.

Figure 5: Annotation Example 4