# Measuring and mitigating factual hallucinations for text summarization

**Erencan Tatar** 12681598 | **Luka Wong Chung** 12676926 | **Paulius Skaisgiris** 14279576 | **Myrthe Moring** 11319119

## Abstract

Advancements in NLG have improved text generation quality but still suffer from *hallucinations*, which lead to irrelevant information and factual inconsistencies. This paper provides an ensemble of metrics that measure whether the generated text is factually correct. Using these metrics we find that fine-tuning is a fruitful hallucination mitigation approach whilst prompt engineering is not.[1]

## 1 Introduction

Recent NLG advances enhance fluency and coherence in tasks like summarization and dialogue generation. However, models can produce nonsensical, grammatically incorrect, or off-topic content known as "hallucinations" (Koehn and Knowles, 2017). Some NLG tasks prioritize creativity over strict factual alignment (Augenstein et al., 2023). Researchers are exploring measurement and mitigation methods (Ji et al., 2023), but the combination of measurements and the impact of mitigation on hallucination metrics remain unresolved.

In this paper, we explore abstractive text summarization, specifically addressing factual hallucinations - *text conflicting with or absent from the reference*. We tackle the challenge of combining measurement and mitigation methods and answer the following research questions:

1. Is an ensemble of metrics more suitable for measuring factual hallucinations as opposed to a single metric?
2. Do smaller models tend to hallucinate more than larger ones?
3. Do the following methods help alleviate hallucinations?
   (a) Specifying in the prompt to the model to only refer to the source text
   (b) Chain-of-verification approach
   (c) Finetuning the model on more text summarization data

[1]Project code: https://github.com/p-skaisgiris/dl4nlp-text-summarization

## 2 Methodology

Our investigation involves presenting, comparing, and combining hallucination metrics. We'll then assess language models using these metrics and conduct experiments to mitigate hallucinations.

### 2.1 Dataset

We chose to use the XSum (Narayan et al., 2018) dataset because of its suitable size and the summaries calling for more abstractive, rather than extractive, text summarization techniques. We believe this demonstrates the models' semantic qualities better.

XSum includes about 226,000 news articles spanning diverse topics, such as politics, sports, entertainment, and technology. Professional editors crafted the high-quality, single-sentence summaries, which are approximately 20 tokens in length on average, while the articles can extend up to 800 tokens.

### 2.2 Models

For most of the experiments, we used the T5 language model introduced by Raffel et al. (2023). T5 is an encoder-decoder type model which was pre-trained on a multi-task mixture of unsupervised and supervised tasks. We opted for the T5 language model due to its accessibility via the HuggingFace Python library, its manageable size for conducting experiments, and its strong baseline language modeling capabilities. Here is a list of the specific model variants used:

1. **t5-small** is a checkpoint with 60 million parameters and the smallest version of the t5 model. We use this to compare the results with the larger version and also to investigate the effect of fine-tuning a smaller model on our hallucination metrics.
2. **t5-small-xsum** is the small version of the t5 model, fine-tuned on the XSUM dataset for text summarization. We use this to investigate the level of factual hallucinations on a smaller model after fine-tuning.
3. **t5-large** is the t5 checkpoint with 770 million

parameters. We use this to compare the results with the smaller version and compare it with its fine-tuned variants.

4. **t5-large-xsum** is the large version of t5 fine-tuned on XSum.
5. **t5-large-xsum-cnn** is based on the t5-large model, fine-tuned on the XSUM and CNN Daily Mail summarization (See et al., 2017) datasets. We use this to investigate whether fine-tuning on more data leads to better results.

*See appendix C for specific model links.*

### 2.3 Factual hallucination metrics

Our methodology takes a multifaceted approach by employing diverse metrics to ensure addressing the complex issues associated with hallucinatory content effectively. Besides ROUGE (Lin, 2004), we employ the following metrics to assess and apply them to machine-generated summaries. Please note that these metrics are calculated between the source document and the predicted summary.

- **QAGS** (Durmus et al., 2020) is a factual consistency metric using question answering (QA). It encompasses three key steps: 1) Question generation (QG) creates questions about the *generated summary*, with standard answers (named entities) from the summary's content. 2) A QA model responds to these questions using the *source document*. 3) The metric calculates factual consistency by comparing the generated answers to the expected ones. See Figure 2 for a visual explanation.
- **BLEURT** (Sellam et al., 2020) is a context-aware metric surpassing traditional ones like BLEU and ROUGE. It employs pre-trained transformers to gauge similarity between generated and reference text, capturing nuances in quality such as fluency and coherence.
- **FACT** (Heo, 2021) is a triple relation-based metric that leverages pre-trained models to extract factual triples from both the source document and the summary. Its output is a ratio of how many triples extracted from the summary are also found in the source document. See Figure 1 for a visual explanation.
- **SUMMAC** (Laban et al., 2022) breaks down source documents and summaries into sentences. Using Natural Language Inference (NLI), it computes entailment probabilities between document and summary sentences. We used the scoring method *SCConv* which uses a convolutional neural network for NLI aggregation.

### 2.4 Detecting factual hallucinations

All scores range between 0 and 1. The metrics provide these advantages in spotting hallucinations: ROUGE identifies instances where the reference content is missing, BLEURT measures semantic similarity of the summary and reference text, SummaC infers hallucination based on low entailment and consistency scores, and QAGS evaluates inaccuracies, gaps in context, and unsupported claims through question and answering, while FACT does the same using named entity recognition. In our experiments, we use a linear ensemble, with each metric equally weighted, to benefit from the various different properties of these metrics. Consequently, higher ensemble scores indicate the presence of both content *and* language similar to the reference text, whereas lower scores imply content that lacks support and utilizes language that significantly differs from the reference text.

## 3 Results

**Q1: Is an ensemble of metrics more suitable for measuring factual hallucinations as opposed to a single metric?**

We generated summaries for the initial 1,000 XSum test split articles using five language models and assessed their factual quality with our proposed metrics. We also employed classic ROUGE scores to evaluate summarization quality, comparing predicted summaries to gold summaries. The results can be found in Tables 1 and 4. Furthermore, our metrics were applied to human-written (Gold) summaries, revealing the lowest metric scores compared to other models. This finding led us to scrutinize the dataset more closely. We arrived at the conclusion that, by our criteria for factual hallucinations, even human-written XSum summaries occasionally exhibit hallucinations.

We provide a detailed illustration of our metrics' evaluation by creating various summaries for a single XSum article, as shown in Table 5. Notably, the gold summary includes a hallucination by mentioning 'Dundee,' which is absent in the article, although Caird Hall is a concert venue in Dundee. To further examine this, we generated three summaries: one with a fixed gold summary using information from the article (replacing 'Dundee' with 'High Street'), a non-hallucinated summary based solely on article content, and a hallucinated sum-

| Metrics | Gold | Models | | | | |
|---|---|---|---|---|---|---|
| | | t5-small | | t5-large | | |
| | | Base | XSum | Base | XSum | XSum + CNN |
| QAGS | 0.1632 | **0.7666** | 0.2356 | 0.7621 | 0.1875 | 0.3550 |
| Rouge-L | 0.0763 | 0.1565 | 0.0998 | **0.1602** | 0.0928 | 0.111 |
| Fact | 0.0412 | 0.2614 | 0.0617 | **0.2648** | 0.0686 | 0.0998 |
| BLEURT | 0.3490 | 0.3788 | 0.3556 | **0.3898** | 0.3383 | 0.3606 |
| Summac | 0.2364 | **0.7311** | 0.2380 | 0.7139 | 0.2301 | 0.2472 |
| Ensemble | 0.1732 | **0.4589** | 0.1982 | 0.4582 | 0.1834 | 0.2347 |

Table 1: Hallucination metrics for two t5-style models: small and large. The column names below these models refer to the type of datasets they have been fine-tuned on. The gold column refers to the hallucination scores of the human-written summaries of XSum.

mary with additional (incorrect) factual details.

Table 6 demonstrates that QAGS, Rouge, and Fact metrics react to factual content with lower scores for less factual summaries. QAGS scores depend on Table 7 question-answer pairs, highlighting the importance of factual content. A non-hallucinated summary without factual focus yields more incorrect answers, while a hallucinated summary, despite a contradiction, provides accurate responses due to substantial factual content. BLEURT favors summaries with language closely resembling the article. Surprisingly, SummaC shows no significant variance in factual information detection, suggesting it's less sensitive to factual content overlap between the source article and the generated summary.

These findings highlight the linear metric ensemble's preference for verbatim source text. When dealing with abstractive summarization, future work should consider adjusting the ensemble weights to prioritize BLEURT and SummaC. This is essential to prevent semantically and grammatically sound summaries from receiving low factual scores. Another approach is modifying QAGS and Fact scores to include synonyms as correct answers, not just exact entities. However, the current linear ensemble is more suitable for extractive summarization, given its reliance on the source document's exact wording. Due to time constraints, we didn't explore reweighting the ensemble for abstractive summarization tasks.

Given these conclusions, it is not surprising that fine-tuning of our models on XSum have significantly worsened their scores. This is because the models were essentially pushed to be more hallucinatory, use abstractive language that did not appear in the source text in order to be as concise as the concise golden summaries in XSUM.

**Q2: Do smaller models tend to hallucinate more than larger ones?**

According to the results in Table 1, no definitive conclusion can be drawn regarding the disparities in hallucination during summarization between the t5-small and t5-large base models. We can observe a small decrease across most metrics for t5-large compared to t5-small models. A possible explanation as to why the larger t5 model fine-tuned on XSum seems to perform worse is that the XSum dataset is not complex enough to fine-tune such a large model. However, further research is needed, especially on datasets with more detailed summaries.

**Q3: Do the following methods help alleviate hallucinations?**

#### 3.0.1 Increasing prompt specificity

In the first experiment, we prompted the language model to rely solely on the source text to boost ROUGE and FACT scores. This test assessed the impact of prompt comprehension on hallucination metrics, potentially offering an efficient way to reduce hallucinations through specific instructions. We tested t5 model variants on the initial 1,000 XSum test articles with the following prompts:

- (The prompt used in all other of our evaluations) **Prompt 1**: Summarize the following text: <article>
- **Prompt 2**: Provide an abstractive summary of the following text while preserving key quotes and phrases: <article>
- **Prompt 3**: Do not paraphrase or deviate from the following text's exact wording. Provide

| Metrics | Prompt 1 | Prompt 2 | Prompt 3 |
|---|---|---|---|
| QAGS | 0.2356 | **0.2448** | 0.2376 |
| Rouge-L | **0.0998** | 0.0991 | 0.0978 |
| Fact | 0.0617 | 0.0565 | **0.0701** |
| BLEURT | **0.3556** | 0.3551 | 0.3530 |
| Summac | 0.2380 | 0.2388 | **0.2410** |
| Ensemble | 0.1982 | 0.1989 | **0.1999** |

Table 2: Results of experiment 1 - prompt specificity. Prompts with higher numbers were more assertive about asking the model to refer to the reference text.

> an abstractive summary of the following text, rephrasing it in your own words while strictly adhering to the original wording and retaining the essential meaning: <article>

Table 2 shows the results. Prompts 2 and 3 exhibit a slight increase in ensemble metric, with a trade-off in performance; Prompt 3 slightly outperforms in Fact and SummaC metrics. However, these results do not allow us to conclude that whether more specific prompts decrease hallucinations. Further research for different models and more sophisticated prompt engineering is necessary, leading to the next experiment.

#### 3.0.2 Chain-of-verification

| Metrics | small | small-x | large | large-x |
|---|---|---|---|---|
| QAGS | **0.3589** | 0.2116 | 0.3574 | 0.1343 |
| Rouge-L | 0.0848 | **0.0895** | 0.0847 | 0.0750 |
| Fact | 0.0463 | **0.0675** | 0.0459 | 0.0416 |
| BLEURT | 0.3010 | **0.3430** | 0.3005 | 0.3193 |
| Summac | **0.4218** | 0.2335 | 0.4200 | 0.2299 |
| Ensemble | **0.2425** | 0.1890 | 0.2417 | 0.1600 |

Table 3: Results of the chain-of-verification experiment. All models are t5 and the '-x' refers to whether the model has been fine-tuned on XSum or not.

This experiment is based on (Dhuliawala et al., 2023), which describes an automatic prompt engineering technique. It seeks to elicit the model's correct knowledge about the initial prompt and fact-check itself leading to the model hallucinating less. A question-generating model creates questions based on the initial summary, the summarizer answers the questions, and the final prompt with questions, answers, and the original prompt is fed into the summarizer again (See Figure 3).

In Table 3, we observe the results of this experiment, which should be compared to Table 1 for context. Notably, the CoVe approach has negatively impacted all models. This could be due to several factors. Firstly, the models lacked fine-tuning for question-answering, potentially causing confusion with the appended questions and their sometimes nonsensical answers. Secondly, incorrect model responses to the questions may have reinforced an erroneous understanding of the source text, affecting the summary (see Table 9). Lastly, it's important to note that some gold summaries in the XSum dataset might contain factual hallucinations as per our definition, and in this context, the generated summaries appear more suitable in terms of quality and hallucination (see Table 8).

#### 3.0.3 Finetuning the model on more text summarization data

Tables 1 and 4 display results for the t5-large model fine-tuned on XSum and CNN summaries. Notably, the t5-large outperforms both XSum and XSum-CNN in hallucination scores. However, Table 4 reveals that t5-large-xsum-cnn generates superior quality summaries when evaluated on the gold XSum summaries. These preliminary findings suggest that fine-tuning the model on a broader range of summarization data could reduce factual inaccuracies and enhance summarization quality.

## 4 Conclusion

This paper introduced an ensemble of metrics to measure factual hallucinations in text summaries and we used it to assess hallucination mitigation approaches. We provide evidence that an ensemble of metrics is more suitable than any single metric. Our ensemble favor summaries matching source text verbatim, suggesting it may perform better with extractive summarization. We observed that smaller models do not hallucinate more than larger models. Fine-tuning on diverse datasets can partly mitigate hallucinations whereas prompt engineering approaches didn't prove effective.

Future research may involve replicating our experiments on diverse summarization datasets with less compressed summaries than XSum. Interventions in the LLM's decoding process and forcing models to generate more detailed summaries could also improve factuality scores. We are also curious whether fine-tuning LLMs on QA tasks is enough to benefit from chain-of-verification approach to mitigate hallucinations.

## References


Isabelle Augenstein, Timothy Baldwin, Meeyoung Cha, Tanmoy Chakraborty, Giovanni Luca Ciampaglia, David Corney, Renee DiResta, Emilio Ferrara, Scott Hale, Alon Halevy, et al. 2023. Factuality challenges in the era of large language models. *arXiv preprint arXiv:2310.05189*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models.

Esin Durmus, He He, and Mona T. Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *CoRR*, abs/2005.03754.

Hoon Heo. 2021. Factsumm: Factual consistency scorer for abstractive summarization.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. *arXiv preprint arXiv:1706.03872*.

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.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

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.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation.


# Appendix

## A Summarization quality results

| Models | Rouge-1 | Rouge-2 | Rouge-L |
|---|---|---|---|
| t5-small | 0.1248 | 0.0161 | 0.0998 |
| t5-small-xsum | 0.2859 | 0.0948 | 0.2456 |
| t5-large | 0.1434 | 0.0227 | 0.1130 |
| t5-large-xsum | 0.1966 | 0.0457 | 0.1666 |
| t5-large-xsum-cnn | **0.3020** | **0.1023** | **0.2578** |

Table 4: ROUGE scores of each of the models we considered evaluated on XSum gold summaries for the same corresponding articles as considered in Table 1.

## B XSum article and summarization examples

| **Article** |
|---|
| The injured pedestrian - a young man - is thought to have been walking with a group of people from a graduation ceremony at the Caird Hall. The incident took place on High Street at about 18:00. The man's injuries are believed not to be life-threatening. The driver of the taxi is thought to be uninjured. |
| **Gold summary** |
| A pedestrian has been struck by a taxi in Dundee after it mounted the pavement. |
| **Non-Hallucinated Summary** |
| A pedestrian has been struck by a taxi in High Street after it mounted the pavement. |
| **Non-Hallucinated Summary - long** |
| A young man, part of a group returning from a Caird Hall graduation ceremony, was injured on High Street around 18:00. His injuries are not life-threatening, and the taxi driver is uninjured. |
| **Hallucinated Summary** |
| During a Caird Hall graduation ceremony, a pedestrian accident on High Street at 18:00 left a young man and a taxi driver in critical condition. |

Table 5: Four examples of summaries based on an article. The gold summary is the article's corresponding human-written summary, the rest are generated. The parts in red are hallucinations. Note that we consider that the gold summary includes a hallucination as it was never mentioned in the article that Caird Hall is in Dundee.

| Metrics | Gold | NH | NH long | H |
|---|---|---|---|---|
| QAGS | 0.000 | 1.000 | 1.000 | 1.000 |
| Rouge-L | 0.078 | 0.145 | 0.280 | 0.189 |
| Fact | 0.000 | 0.000 | 0.000 | 0.000 |
| BLEURT | 0.279 | 0.331 | 0.666 | 0.489 |
| SummaC | 0.204 | 0.203 | 0.206 | 0.204 |
| Ensemble | 0.112 | 0.336 | 0.430 | 0.376 |

Table 6: Different evaluation metrics for the four different generated summaries: hallucinated (H), non-hallucinated (NH) shorter and longer as in Table 5.

| SUM | **Question** | **Source Answer** | **Summary Answer** |
|---|---|---|---|
| H | Where was the graduation ceremony held? | Caird Hall | Caird Hall |
| | On what street was a pedestrian accident at Caird Hall? | High Street | High Street |
| | At what time was the pedestrian accident on High Street? | 18.00 | 18.00 |
| NH s | In what city was a pedestrian struck by a taxi? | Dundee | <*unanswerable*> |
| NH l | Where was the graduation ceremony held? | Caird Hall | Caird Hall |
| | On what street was a man injured? | High Street | High Street |
| | When was the injured man on High Street? | 18.00 | 18.00 |

Table 7: QA generated questions and answers for the example in Table 5

.

| **Prompt** |
| --- |
| *What county in Northern Ireland has been affected by a power supply fault?*<br>Northern Ireland has been affected by a power supply fault in the power supply of a major power supply in County Antrim, Northern Ireland.<br>*What county in Northern Ireland has been affected by a power supply fault?*<br>Northern Ireland has been affected by a power supply fault in the power supply of a major power supply in County Antrim, Northern Ireland.<br>*Where is the power plant located?*<br>A power plant in the north east of England has been shut down after a major power plant was damaged by a nuclear power plant, a company has said.<br><br>Summarize the following text: Areas in Counties Londonderry, Antrim and Down were affected. A spokesperson for Northern Ireland Electricity said was an equipment fault was detected at 21.40 BST. All properties have had power restored had their power restored by 22.14 BST. |
| **Gold summary** |
| Several thousand customers were left without electricity for a time on Wednesday night.<br>**Ensemble score: 0.1187** |
| **Summary with CoVe** |
| Northern Ireland has been hit by a power supply fault in a major power plant in County Antrim, causing power to be shut down in Northern Ireland.<br>**Ensemble score: 0.2129, Rouge-1: 0.0** |
| **Summary without CoVe** |
| Counties Antrim and County Antrim have been affected by a fault in the power supply of a major power plant in Northern Ireland, a spokesperson has said.<br>**Ensemble score: 0.2374, Rouge-1: 0.0** |

Table 8: An example article and summaries, generated by t5-large-xsum, demonstrating the improvement achieved with CoVe compared to the no-CoVe summary approach. The italicized text represents the generated questions. We present ensemble scores for each summary. This XSum instance highlights that the gold summary lacks significant information from the source article. It is unsurprising, then, that the rouge-1 scores of the predicted (and seemingly more factual) summaries calculated w.r.t. to the gold summary are 0.

| **Prompt** |
| --- |
| *Who beat the National League club to reach the last eight of the European Champions Cup?*<br>National League club Leyton Orient secured a place in the last eight of the European Champions Cup with a 2-1 win over Manchester United.<br>*How many places did Viljoen hope to reach in the European Champions Cup?*<br>League One side Viljoen have been drawn in the European Champions Cup with a draw in the first round of the Europa Champions Cup in Londonderry on Saturday.<br>*What competition did Viljoen win?*<br>Viljoen has been crowned as the first winner of the competition to win the coveted trophy in the competition of the year in the final round of the European Championships.<br>*Where did Viljoen beat the National League club?*<br>League One side Viljoen thrashed the National League club in a dramatic win over over the weekend's relegation game.<br>*What league did Viljoen beat to advance to the last eight of the European Champions Cup?*<br>League One side Viljoen have been promoted to the last eight of the European Champions Cup with a 2-1 win over relegation-threatened Manchester United.<br><br>Summarize the following text: The hosts lost opener Chris Dent for a duck in the fourth over and Gareth Roderick (23) shortly after. Will Tavare (20) and Michael Klinger (10) continued a steady fall of wickets until Hamish Marshall (58) and Phil Mustard (38) led the hosts' recovery. However, Viljoen swept through the tail to end with impressive figures of 5-55. |
| **Gold summary** |
| South Africa fast bowler Hardus Viljoen took five wickets on his Kent debut as they bowled Gloucestershire out for 221 on day one in Bristol.<br>**Ensemble score: 0.1206** |
| **Summary with CoVe** |
| League One side Viljoen thrashed relegation-threatened Manchester United to reach the last eight of the European Champions Cup.<br>**Ensemble score: 0.0909, Rouge-1: 0.0741** |
| **Summary without CoVe** |
| Viljoen boosted their hopes of reaching the last eight of the European Champions Cup with a win over at the Clubhouse of the National League club.<br>**Ensemble score: 0.1013, Rouge-1: 0.0741** |

Table 9: In this example, the t5-large-xsum model initially provides incorrect answers to the "verification" questions and subsequently incorporates these inaccuracies into the CoVe summary. The italicized text represents the generated questions. We present our hallucination ensemble scores for each summary.

## C Models

- T5-small: https://huggingface.co/t5-small
- T5-small-XSUM: https://huggingface.co/pki/t5-small-fine-tuned_xsum
- T5-large: https://huggingface.co/t5-large
- T5-large-XSUM https://huggingface.co/sysresearch101/t5-large-fine-tuned-xsum
- T5-large-XSUM-CNN
  https://huggingface.co/sysresearch101/t5-large-fine-tuned-xsum-cnn

## D Hallucination metric pipelines

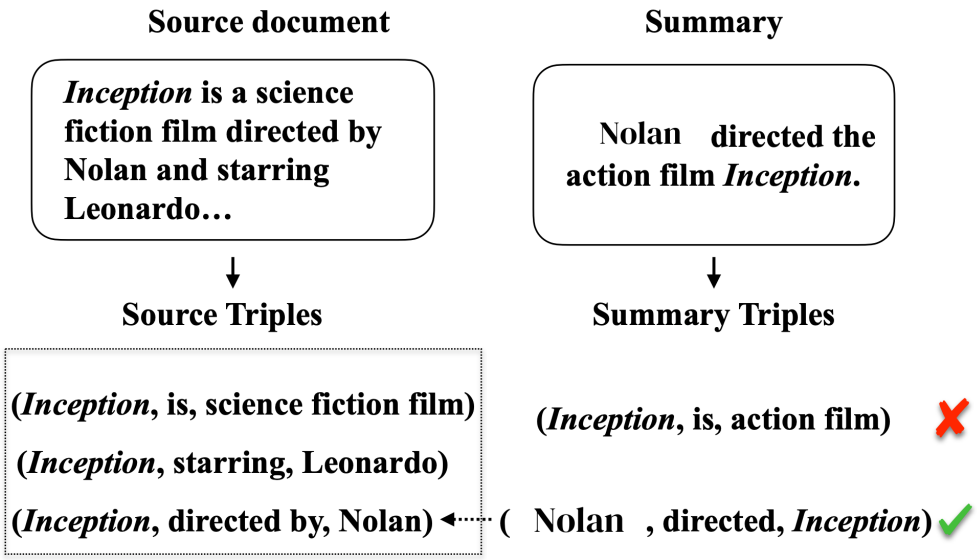


Figure 1: Pipeline for computing Factscore. Figure taken from (Heo, 2021)

.

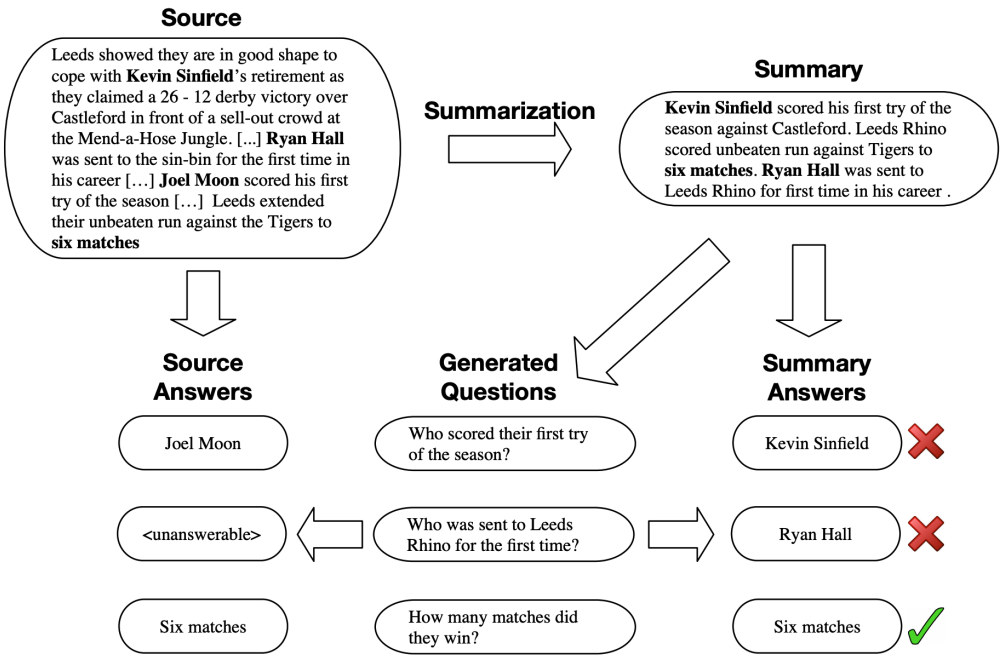


Figure 2: Pipeline for computing QAGS score. Figure taken from (Heo, 2021)

.

## E Chain-of-verification pipeline

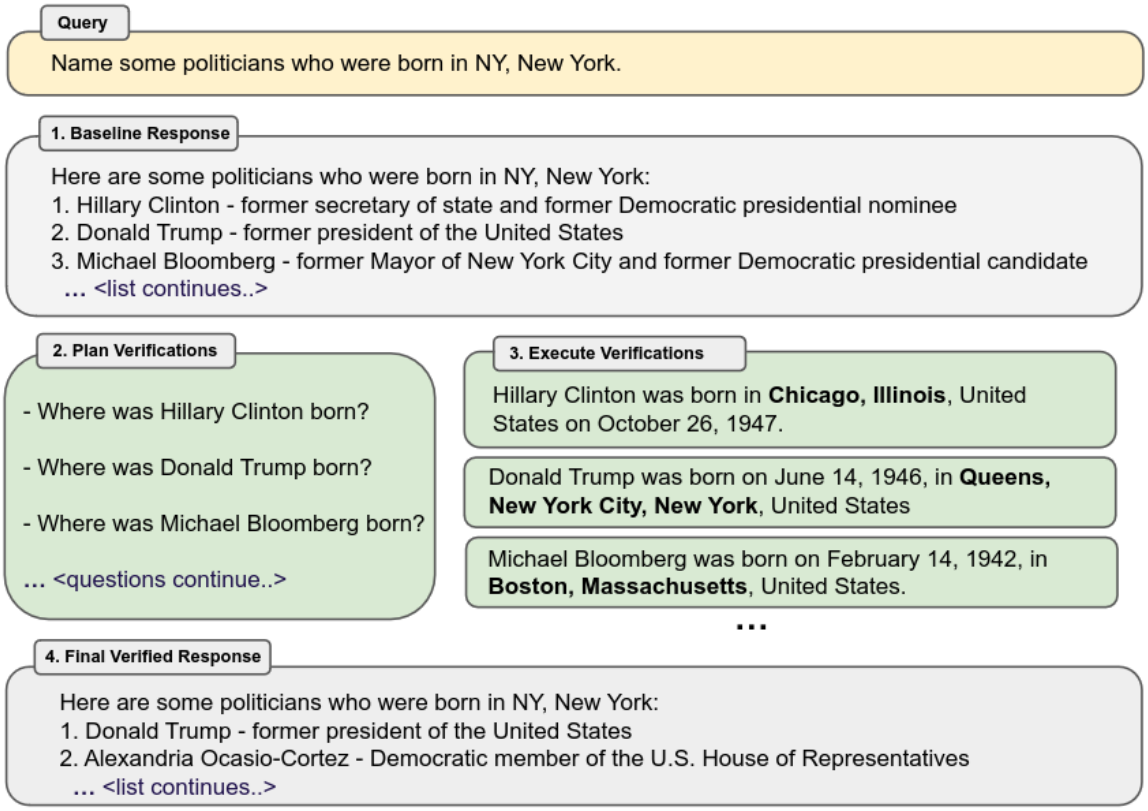


Figure 3: Figure taken from (Dhuliawala et al., 2023). The CoVe method enhances language model responses by using verification questions. These questions are generated to check and improve factual accuracy, resulting in more accurate responses compared to the initial generation, potentially without repeating information from the original response.