# LORE: Continual Logit Rewriting Fosters Faithful Generation

**Charles Yu**[♡] **Qingyun Wang**[♡] **Yuting Hu**[♢] **Jinjun Xiong**[♢] **Heng Ji**[♡]
[♡]University of Illinois Urbana-Champaign [♢]University at Buffalo
{ctyu2,hengji}@illinois.edu

## Abstract

As autonomous agents and assistants, large language models (LLMs) often struggle with "hallucinations." Fundamentally, the problem is one of prioritization and balance: the LLM needs to understand or infer when it needs to be creative and balance that with its need to be accurate. Most efforts focus on either updating intrinsic knowledge via targeted post-training or by adding external knowledge sources which the LLM can reference neurosymbolically (e.g., via retrieval-augmented generation). However, these all eventually rely on the LLM's implicit reasoning ability during generation, still allowing for these random hallucinations despite high-quality training examples and references. Using aspect-oriented summarization as a case study, we propose **LOgit REwriting** (**LORE**), a new controlled generation paradigm which can simultaneously be faithful to external knowledge and to the LLM's intentions. LORE works by adding a rewriting module at left-to-right inference time, continuously reflecting on the newest prediction and trying to find a replacement that is more faithful to the source document. Then, it merges the logits of the replacement with those of the original prediction to generate the next token. We proposed a new long-context aspect-oriented summarization dataset, **SLPAspect**, and find that LORE generates 5.8% better summaries compared to the LLM without LORE-rewriting. [1]

## 1 Introduction

The most widely used LLMs are capable of generating extremely coherent and reasonable text, often surpassing the efficiency and eloquency of humans (Bojic et al., 2023; Mittelstädt et al., 2024; Luo et al., 2025; Samaan et al., 2024). However, for domains like education and medicine, where providing the right or wrong treatment or applying an effective or ineffective intervention can have long-lasting effects on the student or patient, hallucinations often undermine the efficacy of LLMs, eroding the trust of users (Asgari et al., 2025). Due to the infinite input space of prompts and their effects on the generated output, developers continuously update their prompts to reduce perceived hallucinations in document understanding or summarization. However, it is especially in these cases where the dangers of hallucinations prove to be most harmful: when the "95%" of a generation that a human can easily verify looks good, it's simple to take the remaining 5% at face value. A human may tune a prompt to seemingly eliminate hallucinations[2], but without a mechanism to *constantly avoid* them at inference time, this can lead to catastrophic failures when deployed.

These risks are widely pertinent in both the simplest case of prompting an LLM to operate on an explicit source document and in larger systems where retrieval-augmented generation (RAG) is used (Gao et al., 2024). To reduce hallucinations in text generated based on source documents, we propose Logit Rewriting (LORE), a framework for controlling the generation of an LLM and ensuring that the output text is faithful to both the source document and the LLM's creativity and language modeling ability. LORE does this by using a Natural Language Inference (NLI) model to continuously ground the generated output to the source document. Then, in cases of hallucination, LORE changes the offending output into text that can be grounded back to the source. Unlike many other frameworks, LORE allows for rewriting across any token in the vocabulary of the NLI model, as opposed to only reranking options retained in the current generation beam. To evaluate the efficacy of LORE, we also introduce a new dataset for

[1]Code for LORE and collecting SLPAspect are available at https://github.com/CharlesYu2000/LORE-LogitRewriting.

[2]In the context of this paper, we care about hallucinations from the perspective of an LLM introducing information not present in some reference piece of text, i.e., faithfulness.



*Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21314–21328

aspect-oriented summarization called **SLPAspect** focused on the Speech-Language Pathology domain[3]. The Speech-Language Pathology domain is a quintessential usecase where LLMs need to be highly accurate and hallucination-free with respect to their source to reduce risk of harm (we will discuss this further in §2).

To summarize, our primary contributions are:

- We propose a novel paradigm called **LORE** which adds a rewriting mechanism to any open-source LLM to ensure that the generation is faithful to a source document.
- We introduce a new dataset called **SLPAspect** for aspect-oriented summarization. This dataset contains over 4000 scholarly documents with nearly 15000 ground-truth aspect-oriented summaries.
- We apply LORE to SLPAspect and two other aspect-oriented summarization benchmarks, showcasing the efficacy of our rewriting mechanism across different settings.

## 2 Task & Data

Aspect-oriented summarization (Tan et al., 2020; Ahuja et al., 2022) is a task where faithfulness is particularly important, as each document may have multiple different aspects for which different parts of the source document are relevant. Given an aspect (a small subtopic of interest) and a source document, we aim to generate a summary of the source document's information relating to only that aspect, which we will call an "aspect summary." Existing aspect-oriented summarization benchmarks, such as AspectNews (Ahuja et al., 2022), MASSW (Zhang et al., 2025a), and ACLSum (Takeshita et al., 2024) focus on domains with relatively low risk (news and NLP scholarly documents respectively) compared to higher risk domains such as the medical and education domains.

As with other medical practitioners, Speech-Language Pathologists (SLPs) are constantly trying new ways for evaluating and treating patients, often school students with speech and language disorders. However, the patient load for SLPs is extraordinarily high among medical professions, with the average SLP reporting a caseload 25% larger than manageable and 76% of SLPs believing there to be barriers to maintaining manageable caseloads; of the 76%, a third identified the primary barrier being a shortage of SLPs (Association, 2024). Thus, most SLPs rarely have time to read the latest literature to find evidence-based diagnosis methods and interventions and have turned toward LLMs to reduce this burden by extracting information to summarize or synthesize scholarly articles (Hu et al., 2024). In these cases, the risk of harm is high, as LLMs are highly capable of producing reasonable-sounding aspect summaries, but if these contain hallucinations, the already time-constrained SLPs may not realize this and end up applying interventions which are not supported by the literature.

To help research in the intersection of NLP and Speech-Language Pathology, as well as to highlight the difficulty of faithfulness in a domain where most texts seem logical but are only high quality when evidence-based, we introduce **SLPAspect** as a new long context aspect-oriented summarization dataset. In this higher risk domain, hallucinations can have more consequences while simultaneously being more difficult to recognize. SLPAspect is constructed by collecting articles from three journals: *International Journal of Speech-Language Pathology* (**IJSLP**), *Language, Speech, and Hearing Services in Schools* (**LSHSS**), and *American Journal of Speech-Language Pathology* (**AJSLP**). These are three of the top journals published by the American Speech-Language-Hearing Association (ASHA), the leading organization on SLP research and practices. In a subset of these articles (as this is not required by the journals), the original authors wrote "headline"-type abstracts summarizing key aspects of their articles (e.g., purpose, method, results, conclusion, etc.). So, to construct SLPAspect, we collected all volumes of the associated journals and parsed their text to use these headline abstracts as ground-truth aspect summaries.

A summary of dataset statistics is presented in Tables 1 and 2, with example ground-truth aspect summaries shown in Appendix A.1.

## 3 Methodology

Architecturally, LORE uses the concept of "rewriting" as a way to control the generation. By continuously grounding the output back to the source document as we perform left-to-right generation, we can have a fine-grained control over faithfulness to both the LLM doing the generation as well as to the source document itself with built-in source

[3]Speech-Language Pathology is a medical discipline focused on evaluating, diagnosing, and treating communication and language disorders.

| Journal | Start Year | # of Issues | # of Articles | # of Articles with Ground-Truth Aspects |
|---|---|---|---|---|
| JSHLR | 1990 | 179 | 3652 | 1831 |
| LSHSS | 1990 | 133 | 1316 | 789 |
| AJSLP | 1991 | 143 | 1895 | 1476 |

Table 1: Statistics of the 3 ASHA journals included in SLPAspect.

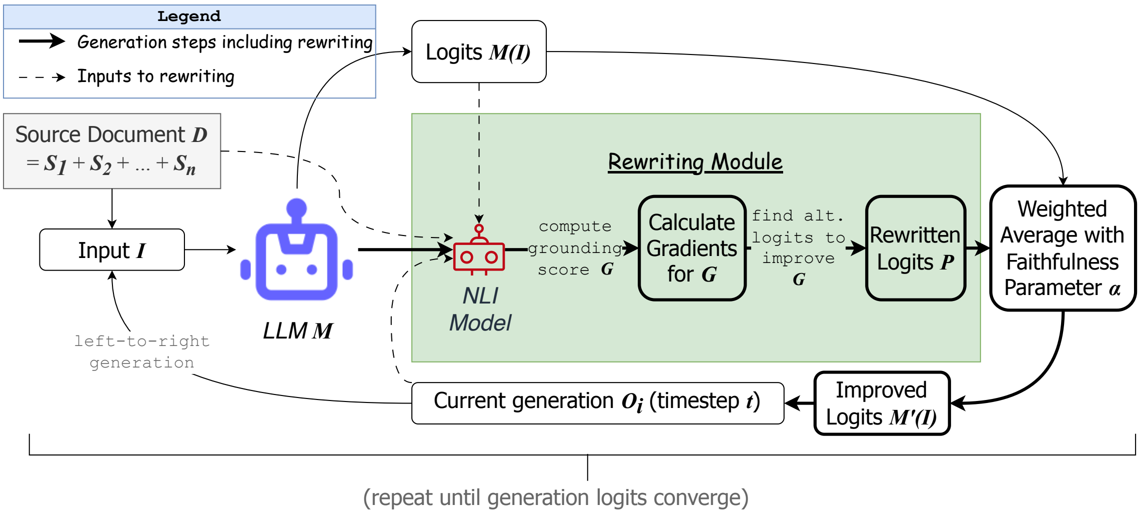


Figure 1: This diagram shows the high-level flow of LORE. The source document is split into segments which are then used by the LLM to generate the aspect summaries. While generating each aspect summary, we continuously rewrite the output by passing it through the NLI model, backpropagating to find improved embeddings, and combining with the original logits to generate a new set of tokens to be sampled. A detailed diagram of the rewriting steps can be found at Figure 2.

| Aspect | Count |
|---|---|
| Purpose | 4037 |
| Method | 3974 |
| Conclusion | 3891 |
| Result | 3727 |
| (other aspects) | 288 |

Table 2: Aspects and number of examples included in SLPAspect.

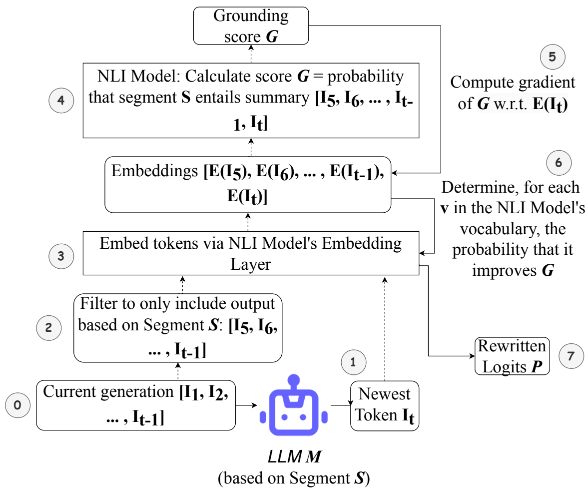


Figure 2: The steps to determining logits to rewrite the generation output. This fits into the Rewriting Module portion of Figure 1.

attribution. As it grounds the output to the source document, LORE finds cases where the generation is not faithful and, by approximating the effects of the embedding space on that faithfulness, looks across the entire vocabulary of the language model to find replacement tokens which best revise the generation toward a more faithful generation.

Although context length is often a factor when evaluating faithfulness, LORE does not rely on long context capabilities to function effectively. Instead, the rewriting of LORE is based on shorter targeted segments which are partitioned during pre-processing. Thus, LORE can be used plug-and-play with other methods across domains with different faithfulness measures for short or long contexts.

### 3.1 Rewriting based on Natural Language Inference at a High Level

The core methodology behind LORE is rewriting the generation based on continuous grounding. To ground the generation back to the source document, we employ the `roberta-large-mnli` (Liu et al., 2019) NLI model to detect unfaithful generations. After our LLM has produced a set of logits and chosen the best next token for the generation, we pass the source document and current generation output through our NLI model to produce a grounding score indicating whether or not the source document entails the current generation. In successful cases, the current generation continues to be entailed, but in cases where hallucinations are being created, the current generation will not be entailed. So, for this latter case, we want to rewrite the token such that the generation continues to be entailed.

To rewrite the token, we introduce a procedure which constructs a first-order approximation of the improvement that any other token would have on this generation and its probability of most improving the grounding score. This procedure will be detailed in the following section. Then, with these new probabilities, we mix them back in with the original logits produced by the generation LLM using an average weighted by a faithfulness hyperparameter $\alpha$. This produces a new set of probabilities which we can sample tokens from to replace the originally generated next token, and we repeat the rewriting step until there is no further improvement to the token. This full procedure is illustrated in Figure 1.

Note that $\alpha$ can be interpreted as how much we value faithfulness to the original source document vs faithfulness to the desires of the generation LLM. For higher creativity, we weight toward the LLM (decrease $\alpha$) and for better source attribution, we weight toward the source document (increase $\alpha$).

### 3.2 Producing Rewritten Logits from NLI

To rewrite this token, we employ an approach similar in conception to AutoPrompt (Shin et al., 2020). To start, we backpropagate the grounding score to produce a gradient with respect to the input embeddings by the NLI model. Specifically, when looking at the gradient with respect to the embedding of the current token, we know the direction in which those embedding values should move to locally increase the grounding score. Since these embeddings in modern Transformer-based models are based on lookups, we cannot directly translate this gradient to a better token, but by utilizing the HotFlip technique (Ebrahimi et al., 2018), we can find which tokens are in the direction hinted at by the gradient and find a best match.

Formally, given a text sentence $s = [s_1 s_2 \dots s_t]$, where $s_i$ is a single token for each $i$, let $t$ be the index of the token being generated at the current time-step. We pass the prompt $s$ (composed of the premise and hypothesis) through the NLI model $M$ and compute a logistic loss $\mathcal{L}(s)$ where the target label is 1 for being grounded/entailed:

$$\mathcal{L}(s) = -\log M(s)$$

The NLI model produces logits with respect to each $1 \times D$ embedding in the embedding layer $E$ (where $D$ is the embedding size). So, we capture the gradient of $\mathcal{L}(s)$ with respect to $s_t$'s embedding (only the token embedding, no positional embeddings) and call this gradient $\nabla_t$ (in PyTorch, we can get $\nabla_t$ by attaching a backward hook to the embedding layer and setting it to track the gradient).

$\nabla_t$ tells us the direction in which the $t^{th}$ token's embedding should move for the maximum change in the NLI model's loss, but there may not be any actual token with an embedding that matches. For token $s_t$, we can find the token's embedding $E(s_t)$ and find the vector difference between it and every other embedding in the model's vocabulary $(E - E(s_t))$. Then, by taking the cosine similarity between $\nabla_t$ and each of the vector differences, we can determine which tokens have embeddings closest in the direction of the maximal change in loss.

$$\cos(\nabla_t, E(v) - E(s_t)),$$

where $v$ is any token in the vocabulary of the NLI model. This allows us to truly “rewrite" this token as opposed to other methods which only rerank existing candidates.

Note though that the cosine similarity is only looking at the directional change between the embeddings, but there of course may be multiple tokens whose embeddings are very close to the correct angle from the current embedding. Despite being close in angle, the actual distance in the embedding space can be far, which would not satisfy the assumptions of the first-order approximation. So, we penalize those embeddings that are farther from the current embedding by additionally scaling according to the inverse of their distance. So, for each token $v$ in the vocabulary, we calculate:

$$p_v := \frac{\cos(\nabla_t, E(v) - E(s_t))}{\max\{||E(v) - E(s_t)||, 1\}} \quad (1)$$

to produce a new set of logits. We then normalize these to produce a set of probabilities $\mathcal{P}'_t$ over the model's vocabulary:

$$P'_t[v] := \frac{p_v}{\sum_v p_v} \quad (2)$$

We combine this with the original probabilities $\mathcal{P}_t$ produced during the generation according to the faithfulness parameter $\alpha$ according to

$$\mathcal{P}_t := (1 - \alpha)\mathcal{P}_t + \alpha\mathcal{P}'_t. \quad (3)$$

This gives a linear interpolation between the original generation and the grounded rewritten generation, allowing for the operator to decide the weighting between faithfulness to the original source document and faithfulness to the LLM's parametric and natural inference.

### 3.3 Improving NLI by attending to source document segments

When a document is long, we cannot assume that the NLI model will classify entailment accurately. Inherently, there's a disconnect, as the generation LLM is the one generating the hypothesis based on some portion of the source document, while a separate language model (the NLI model) is determining which part of the source document to ground back to.

To solve this conundrum, we first partition the source document into an arbitrary number of segments by prompting GPT-4. Then, during left-to-right generation, we use the attention matrices to determine which segments of the source document are being focused on by the LLM to generate the current token. Specifically, we use the first layer of the attention matrices, aggregating across all attention heads and over the tokens of each segment, to compute a probability for each segment of the base text. We pick the top segment only and use this as the premise for the NLI model. The intuition is that the LLM is generating the current token based on that segment, and thus that segment is what the token and generation must be primarily faithful to.

With an arbitrary source document, we cannot make any assumptions about the structure of the document nor the contents. For aspect-oriented summarization, there might not be any inherent structure or relation between parts of the text that are relevant to that aspect. Thus, in many cases, a piece of the generated summary may already include information from one segment but is now generating based on information from another segment. For example, if we were summarizing a document with only had two segments, suppose the first half of our summary is entailed only by the first segment and the second half by the second segment. Upon generating the second half of the summary, the NLI model accurately determines that the second segment does not entail the first half of our summary.

To prevent this, we keep a mapping from each token in the generation to the attributed segment. Then, when employing the NLI model, the current segment is used as the premise, while only the subset of the generation related to that same segment is used as the hypothesis (in the prior example, we would only ever look at the first and second halfs of the summary at a time). This is shown in Figure 2.

## 4 Results

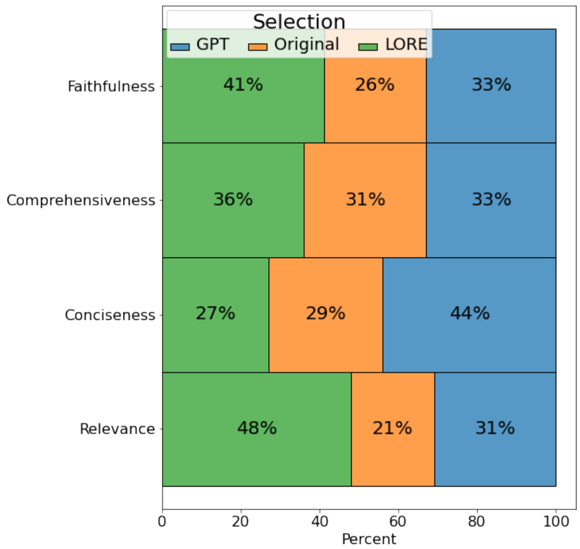


Figure 3: When compared head-to-head with GPT-4 rewrites, LORE rewrites are often higher quality.

### 4.1 Baselines

We compare LORE, using both a `DeepSeek-R1-Distill-Qwen-1.5B` and a `Qwen3-1.7B` backbone (thinking off), with 5 baselines across different levels of resources. We compare with 1) BertSum (Liu and Lapata, 2019), a BERT-based model fine-tuned on the

| Model | SLPAspect | | | | AspectNews | | | | ACLSum | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R-1 | R-2 | R-L | GPT | R-1 | R-2 | R-L | GPT | R-1 | R-2 | R-L | GPT |
| BertSum | 22.1 | 11.1 | 12.7 | 52.0 | 61.2 | 44.3 | 48.0 | 81.5 | **33.7** | 26.0 | 27.4 | 41.0 |
| CTRLSum | 24.3 | 12.2 | 13.7 | 67.2 | 38.1 | 26.9 | 27.4 | 52.8 | 27.9 | 20.5 | 21.0 | 40.7 |
| AOSumm | 23.6 | 14.0 | 14.2 | 65.8 | 64.3 | 45.6 | 49.2 | 82.7 | 31.3 | 24.3 | 25.7 | 39.2 |
| DeepSeek-R1 | 22.7 | 13.5 | 13.2 | 67.1 | 64.5 | 46.1 | 48.4 | 85.2 | 31.1 | 25.4 | 26.7 | 46.3 |
| GPT-4o-mini | 26.8 | 13.8 | 15.1 | 68.0 | 62.9 | 45.9 | 49.3 | 87.7 | 31.7 | 25.9 | 26.5 | 45.8 |
| LORE (Qwen3) | 30.8 | 15.8 | 18.0 | 69.3 | **67.0** | 49.6 | 50.3 | 86.2 | 31.8 | 26.7 | 28.0 | 47.8 |
| LORE (R1) | **32.5** | **16.2** | **18.8** | **71.0** | 66.7 | **50.1** | **52.1** | **90.3** | 32.1 | **27.7** | **28.9** | **50.3** |

Table 3: Performance of LORE using `DeepSeek-R1-Distill-Qwen-1.5B` and `Qwen3-1.7B` for generation compared to baseline models across three aspect-oriented summarization datasets including SLPAspect. $R-1$, $R-2$, and $R-L$ are the respective ROUGE scores compared to ground-truth while $GPT$ is the percentage of examples where the GPTScore evaluated from OpenAI o3-mini is higher than that of the ground-truth.

CNN/DailyMail dataset, 2) CtrlSum (He et al., 2022), a pretrained summarization model which generates summaries conditioned on additional prompts, and 3) AOSumm (Ahuja et al., 2022), a model based on BertSum which additionally finetunes to match extracted keywords with summaries, specifically for the aspect-oriented summarization task.

Finally, we compare with representative open- and closed-source LLMS, `DeepSeek-R1-Distill-Qwen-1.5B` (without LORE rewrites) and `GPT-4o mini`.

### 4.2 Evaluating Aspect Summaries

To evaluate generated aspect summaries on SLPAspect, we employ the ROUGE metrics to compare with the ground truth from the dataset. Furthermore, for each method, we compute a GPTScore (Fu et al., 2024) using probabilities from OpenAI o3-mini (reporting these as the percentage of outputs evaluated as better than the ground-truth).

We also evaluate on AspectNews (Ahuja et al., 2022) and ACLSum (Takeshita et al., 2024), two aspect-oriented summarization datasets spanning news and NLP scholarly documents respectively.

These results are reported in Table 3. Overall, we see that LORE achieves the best results in nearly all instances, showcasing the benefit of rewriting when generating necessarily extractive information.

### 4.3 Rewriting Ablations

The `DeepSeek-R1` and `LORE (R1)` rows of Table 3 showcase that rewriting improves the generation performance considerably. Furthermore, we can see that, compared to ground truth, our method outperforms GPT-4.

Rewriting the entire generation output with a LLM is trivial and we might expect to see similar improvements. Using the same evaluation criteria, we see that prompting GPT-4 to rewrite its own generation output yields improvements, and even generates a higher proportion of better rewrites than LORE does. These numbers are reported in Table 5.

The true benefit of LORE shines when evaluating the *quality* of the rewritten summaries. Inspired by Fu et al. (2024) and using GPTScore for evaluation, we score the quality of each rewrite along four dimensions: **Faithfulness**, **Comprehensiveness**, **Conciseness**, and **Relevance**.

To validate the quality of LORE rewrites compared to LLM rewrites, we performed a three-way comparison between the non-rewritten generation, LORE-rewritten, and GPT4-rewritten summaries along these four dimensions. These results are reported in Figure 3.

From these results, LORE is clearly better than GPT-4 rewriting along the Faithfulness and Relevance dimensions. For Comprehensiveness, LORE also pulls slightly ahead. However, LORE falls short with respect to Conciseness. The shortcomings for conciseness seem to primarily stem from redundant (yet faithful) information, rather than the summaries being too long. This is likely caused by the lack of a deletion mechanism in LORE, which can only rewrite a single token at a time, left-to-right. Thus, in some cases where there might be redundant information, LORE prioritizes faithfulness over brevity. In contrast, LLMs rewrite the entire output in one fell swoop, allowing for arbitrary removals.

An example of the above behavior is shown in Table 4. In this example, we see that LORE can

| Source | Summary |
|---|---|
| Ground Truth | The participants were 2 Canadian English-speaking adolescents who had not yet acquired /r/. The study included an initial period without ultrasound and 13 treatment sessions, each 1 hr long, using ultrasound. Speech samples were recorded at screening and immediately before and after treatment. Samples were analyzed acoustically and with *listener judgments*. *Ultrasound images* were obtained before, during, and after the treatment period. |
| DeepSeek-R1 | Two Canadian teens were assessed before and after treatment using **speech tasks**, audio recordings, and ultrasound. The assessments included an *oral exam, /r/ sound discrimination, and word lists in set phrases*. Ultrasound showed *tongue movement* during speech, using the same setup before and after treatment. One teen had extra single-word recordings after treatment. |
| DeepSeek-R1 with LORE | Two Canadian teens were assessed before and after treatment using **speech samples**. Audio recordings were assessed with an *oral mechanism exam, /r/ sound discrimination, and word lists*. Ultrasound recordings showed *midsagittal tongue images and coronal images of the tongue*. Both an *acoustic formant analysis* and a listener judgment task were conducted. |

Table 4: Generated summaries for the "Method" aspect of Adler-Bock et al. (2007). The LORE rewrite diverges at the word "tasks," determining that the most faithful generation is actually "speech samples" rather than "speech tasks." Despite this divergence, it eventually generates what the original model would generate, and intersperses more details as needed.

cause similar information to be generated to avoid hallucinations. Even after this first rewriting, the generated summary still roughly follows the originally intended generation, although it is not identical due to the change in context from the rewriting.

### 4.4 Adjusting the Faithfulness Parameter

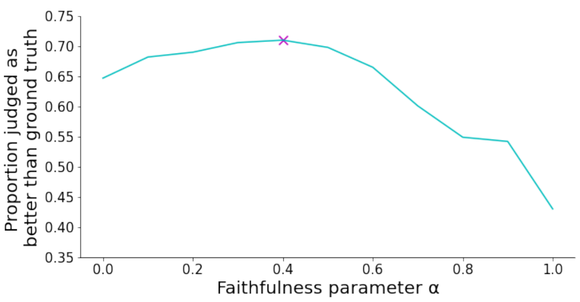


Figure 4: Proportion of rewritten summaries judged as better than original when varying the faithfulness parameter value $\alpha$ of LORE. Of increments of 0.1, the best performance on SLPAspect came from $\alpha = 0.4$ in our setup.

During the course of rewriting, we can vary the faithfulness parameter $\alpha$ from 0 to 1, with 0 being fully faithful to the language model's original generation and 1 being fully faithful to the rewriting module and the source document. Note that an $\alpha$ value of 1 does not mean that the generation model is completely ignored; in fact, we still fully utilize the models aggregation abilities and still only rewrite according to the model's intention (i.e., "generate this next token based on this segment").

When varying $\alpha$, we can see in Figure 4 that the overall performance of the model (aggregated over the four rewriting criteria) changes. For the best comprehensive results, we find that more equal weightings of faithfulness toward the model and rewriting modules yields the best results. Notably, performance skews toward trusting the language model more, which is not surprising given the relative sizes of the models and the fact that the rewriting module is still largely dependent on the original generation model's aggregation ability.

### 4.5 Computational Cost

The computational overhead of adding LORE largely depends on the size of the models, so this is ultimately a tradeoff that users should consider for their domains/tasks, much like the considerations they should be making when choosing any LLM of differing sizes.

Using the general heuristic where the backward pass is approximately twice the computational cost

| Generation Model | Rewriting Method | R-1 | R-L | GPT |
|---|---|---|---|---|
| DeepSeek-R1 | Prompting GPT-4 | 26.7 | 14.2 | 71.6 |
| DeepSeek-R1 | LORE | 32.5 | 18.8 | 71.0 |

Table 5: Scores over SLPAspect when rewriting using prompting compared to LORE. Prompt instructions can be found in Appendix A.5. Here, across all evaluated samples, GPT-4 rewrites result in more of those samples being evaluated as better than the ground-truth.

as the forward pass, with $\mathcal{O}(M)$ being the inference cost of the LLM and $\mathcal{O}(N)$ being the inference cost of the NLI model, adding LORE to the LLM changes the generation cost from $\mathcal{O}(M)$ to $\mathcal{O}(M + 3cN)$, where $c$ is the number of iterations of rewriting. In practice, $c$ depends on $\alpha$ (the faithfulness parameter) and the amount of mistakes of the base LLM. In our experiments, $c$ was just over 1 on average (as only a small portion of tokens need rewriting, after which the LLM will start self correcting). Omitting the input size factor from the calculation (the NLI model input is only a single segment vs the entire source), with a 1.5B parameter LLM and 355M parameter NLI model, the overhead is about 75%, whereas a 7B parameter LLM with the same 355M parameter NLI model would only incur an overhead of about 14%. Even larger models would see a proportionally smaller overhead (and, presumably, would make fewer mistakes that need correcting, reducing $c$ further).

## 5 Related Work

**Aspect-oriented Summarization.** Most aspect-oriented summarization methods focus on shorter documents in specific formats, such as question-answer forums (Chaturvedi et al., 2024) or reviews (Angelidis and Lapata, 2018; Bražinskas et al., 2020; Bhaskar et al., 2023; Tang et al., 2024). Fewer methods focus on general long documents summarized for generic aspects. Other approaches to aspect-oriented summarization include weak supervision (Tan et al., 2020), self-supervised pre-training (Soleimani et al., 2022), keyword-guided summarization (He et al., 2022), prompt engineering (Bhaskar et al., 2023), and extractive summarization (Qi et al., 2022; Tang et al., 2024; Hu et al., 2025). In contrast to these methods, we utilize a rewriting mechanism to improve the natural generation capabilities of a pretrained large language model.

**Faithfulness Rewriting.** Faithfulness, which represents the factual consistency with the input text, plays a vital role in summarization. Previous work has explored the role of post-editing (Cao et al., 2020), namely rewriting, to ensure such faithfulness. Despite this, those post-editing techniques usually require training revision models (Dong et al., 2020; Adams et al., 2022; Fabbri et al., 2022; Gao et al., 2023). Inspired by the success of logit editing in decoding phrases (Aralikatte et al., 2021; Wan et al., 2023), we use the NLI model as a signal to improve the consistency between source document segments and generation results. One direction close to our work is embedding steering (Jahanian* et al., 2020; Li and Liang, 2021; Subramani et al., 2022; Han et al., 2024). However, these efforts rely on external signals such as sentiment and toxicity to control the model's generation, while LORE relies on the knowledge consistency between generation results and source document. Separately, Qiu et al. (2024) utilized general hypothesis verification models (of which NLI is a subset) to rerank generations based on verification but can only operate over candidate sequences, while Yu et al. (2023) used a similar classification model-based signal but for post-training the pretrained language model itself.

**Hallucination Prevention, Detection and Mitigation.** Factuality hallucination detection in LMs typically involves external fact-checking methods, such as FACTSCORE (Min et al., 2023) and FacTool (Chern et al., 2023), or internal uncertainty analysis. The latter includes Chain-of-Verification (Dhuliawala et al., 2024), logit-based assessments (Kadavath et al., 2022; Zhang et al., 2024c), and leveraging LM internal states (Varshney et al., 2023; Luo et al., 2024). When internal states are unavailable, self-consistency probing (Manakul et al., 2023; Agrawal et al., 2024) or multi-LM corroboration (Cohen et al., 2023) can provide alternative signals. In a related direction, Zhang et al. (2024d) introduce a RESET token during training which allows LLMs to completely take back its generation if it discovers issues. Our work is also related to prior studies on mitigating hallucinations. Shen et al. (2021) addresses the issue by filtering out low-quality training data. Several approaches enhance model factuality through external knowledge (Niu et al., 2024; Xie et al., 2024; Lyu et al., 2023; Asai et al., 2024), and knowledge-aware tuning (Li et al., 2023). Some studies tackle hallucination by enforcing LLMs to adhere to input (Tian et al., 2019; Aralikatte et al., 2021), modifying internal states (Chen et al., 2023; Azaria and Mitchell, 2023; Gottesman and Geva, 2024), and adopting refusal-awareness (Zhang et al., 2024a). Recent work (Zhang et al., 2024b, 2025b) prevents hallucination by modeling it quantitatively, incorporating fine-grained factors like knowledge popularity, length, and model size. Compared to them, our work aligns with advanced decoding strategies (Wan et al., 2023; Shi et al., 2024) to enhance factuality. Most similar to our work is that by Aichberger et al. (2024) who use a similar mechanism of finding token replacements, but rather than merge with the original predictions, do a direct replacement to generate a semantically different output.

## 6 Conclusions and Future Work

In this paper, we introduced a paradigm called Logit Rewriting (LORE) for generating faithful summarizations based on a canonical source document. We showed that LORE's performance exceeds that of specially trained models and the latest closed- and open-source LLMs by utilizing "Rewriting" during inference time to control the generation. We also introduced a new dataset called SLPAspect for the aspect-oriented summarization task, which contains long-context data in the medical/education domains along with expert ground-truth aspect summaries.

Future directions for this line of work include applying LORE to general-domain RAG applications. Furthermore, increasing the rewriting context (instead of only the current token) of LORE and adding a deletion mechanism would help bridge the gap toward the seq2seq behavior of LLM-based rewriting. Also interesting would be rewriting using domain-specific classification models rather than only NLI.

Despite our method's attempts to curtail hallucinations, application of LORE is not enough to guarantee that generations will be accurate, safe, and hallucination free. We urge all readers to remain vigilant.


## Acknowledgments

This research is based upon work supported by the AI Research Institutes program by National Science Foundation and the Institute of Education Sciences, U.S. Department of Education through Award No. 2229873 - AI Institute for Transforming Education for Children with Speech and Language Processing Challenges, DARPA SemaFor Program No. HR001120C0123, and DARPA Seedling BRIES No. HR0011-24-3-0325. The opinions, views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the National Science Foundation, the Institute of Education Sciences, the U.S. Department of Education, DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.


## Limitations

There are a few core limitations of the methodology in this paper. The first limitation is in LORE's application to open-source large language models. Although open-source LLMs are widely used, many researchers, engineers, and users alike might not have the resources or expertise to deploy and use these LLMs. Thus, it would be ideal if LORE could be extended to apply to closed-source large language models. The second core limitation is in LORE's inability to delete or rewrite longer lengths of the generation. This leads to hyperlocal edits which, although effective, mean that some hallucinations cannot possibly be treated (e.g., there is no backtracking). Another limitation of our methodology is that rewriting involves continual iteration and essentially creates an optimization problem at each timestep of generation: optimizing each token for more timesteps might be more effective but more expensive. Furthermore, the optimization problem is not convex and therefore inherently difficult to solve well. One solution might be to maintain rewriting beams on top of generation beams vs a greedy decoding. Finally, one limitation of our experiments is that evaluation by language models is inherently biased and may itself be prone to hallucination. Although we have done our best to mitigate these effects, there is no way to prevent them completely.

## Ethical Considerations

There are no ethical considerations introduced by the methodology in this paper apart from those already existing for Large Language Models.

## References


Griffin Adams, Han-Chin Shing, Qing Sun, Christopher Winestock, Kathleen McKeown, and Noémie Elhadad. 2022. Learning to revise references for faithful summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4009–4027, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marcy Adler-Bock, Barbara May Bernhardt, Bryan Gick, and Penelope Bacsfalvi. 2007. The use of ultrasound in remediation of north american english /r/ in 2 adolescents. *American Journal of Speech-Language Pathology*, 16(2):128–139.

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Kalai. 2024. Do language models know when they're hallucinating references? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 912–928, St. Julian's, Malta. Association for Computational Linguistics.

Ojas Ahuja, Jiacheng Xu, Akshay Gupta, Kevin Horecka, and Greg Durrett. 2022. ASPECTNEWS: Aspect-oriented summarization of news documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6494–6506, Dublin, Ireland. Association for Computational Linguistics.

Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. 2024. Semantically diverse language generation for uncertainty estimation in language models. *Preprint*, arXiv:2406.04306.

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Rahul Aralikatte, Shashi Narayan, Joshua Maynez, Sascha Rothe, and Ryan McDonald. 2021. Focus attention: Promoting faithfulness and diversity in summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6078–6095, Online. Association for Computational Linguistics.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. 2025. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digital Medicine*, 8(1):1–15. Publisher: Nature Publishing Group.

American Speech-Language-Hearing Association. 2024. 2024 schools survey report: Slp caseload and workload characteristics. *American Speech-Language-Hearing Association*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.

Adithya Bhaskar, Alex Fabbri, and Greg Durrett. 2023. Prompted opinion summarization with GPT-3.5. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9282–9300, Toronto, Canada. Association for Computational Linguistics.

Ljubisa Bojic, Predrag Kovacevic, and Milan Cabarkapa. 2023. Gpt-4 surpassing human performance in linguistic pragmatics. In *arxiv*.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Unsupervised opinion summarization as copycat-review generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5151–5169, Online. Association for Computational Linguistics.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6251–6258, Online. Association for Computational Linguistics.

Rochana Chaturvedi, Abari Bhattacharya, and Shweta Yadav. 2024. Aspect-oriented consumer health answer summarization. *Preprint*, arXiv:2405.06295.

Anthony Chen, Panupong Pasupat, Sameer Singh, Hongrae Lee, and Kelvin Guu. 2023. Purr: Efficiently editing language model hallucinations by denoising language model corruptions. *Preprint*, arXiv:2305.14908.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative ai – a tool augmented framework for multi-task and multi-domain scenarios. *Preprint*, arXiv:2307.13528.

Roi Cohen, May Hamri, Mor Geva, and Amir Globerson. 2023. LM vs LM: Detecting factual errors via cross examination. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12621–12640, Singapore. Association for Computational Linguistics.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

Yue Dong, Shuohang Wang, Zhe Gan, Yu Cheng, Jackie Chi Kit Cheung, and Jingjing Liu. 2020. Multi-fact correction in abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9320–9331, Online. Association for Computational Linguistics.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 31–36, Melbourne, Australia. Association for Computational Linguistics.

Emma Everaert, Iris Selten, Tessel Boerma, Michiel Houben, Jacob Vorstman, Hester de Wilde, Desiree Derksen, Sarah Haverkamp, Frank Wijnen, and Ellen Gerrits. 2023. The language profile of preschool children with 22q11.2 deletion syndrome and the relationship with speech intelligibility. *American Journal of Speech-Language Pathology*, 32(1):128–144.

Alex Fabbri, Prafulla Kumar Choubey, Jesse Vig, Chien-Sheng Wu, and Caiming Xiong. 2022. Improving factual consistency in summarization with compression-based post-editing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9149–9156, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Daniela Gottesman and Mor Geva. 2024. Estimating knowledge in large language models without generating a single token. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3994–4019, Miami, Florida, USA. Association for Computational Linguistics.

Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher, and Heng Ji. 2024. Word embeddings are steers for language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16410–16430, Bangkok, Thailand. Association for Computational Linguistics.

Junxian He, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2022. CTRLsum: Towards generic controllable text summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5879–5915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yuntong Hu, Zhuofeng Li, Zheng Zhang, Chen Ling, Raasikh Kanjiani, Boxin Zhao, and Liang Zhao. 2025. Hireview: Hierarchical taxonomy-driven automatic literature review generation.

Yuting Hu, Dancheng Liu, Qingyun Wang, Charles Yu, Heng Ji, and Jinjun Xiong. 2024. Automating knowledge discovery from scientific literature via llms: A dual-agent approach with progressive ontology prompting. *Preprint*, arXiv:2409.00054.

Ali Jahanian*, Lucy Chai*, and Phillip Isola. 2020. On the "steerability" of generative adversarial networks. In *International Conference on Learning Representations*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. 2023. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1774–1793, Toronto, Canada. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of* *the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Junyu Luo, Cao Xiao, and Fenglong Ma. 2024. Zero-resource hallucination prevention for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3586–3602, Miami, Florida, USA. Association for Computational Linguistics.

Xiaoliang Luo, Akilles Rechardt, Guangzhi Sun, Kevin K Nejad, Felipe Yáñez, Bati Yilmaz, Kangjoo Lee, Alexandra O Cohen, Valentina Borghesani, Anton Pashkov, and 1 others. 2025. Large language models surpass human experts in predicting neuroscience results. volume 9, pages 305–315. Nature Publishing Group UK London.

Xiaozhong Lyu, Stefan Grafberger, Samantha Biegel, Shaopeng Wei, Meng Cao, Sebastian Schelter, and Ce Zhang. 2023. Improving retrieval-augmented large language models via data importance learning. *Preprint*, arXiv:2307.03027.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *Preprint*, arXiv:2305.14251.

Justin M Mittelstädt, Julia Maier, Panja Goerke, Frank Zinn, and Michael Hermes. 2024. Large language models can outperform humans in social situational judgments. volume 14, page 27449. Nature Publishing Group UK London.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

Siya Qi, Lei Li, Yiyang Li, Jin Jiang, Dingxin Hu, Yuze Li, Yingqi Zhu, Yanquan Zhou, Marina Litvak, and Natalia Vanetik. 2022. SAPGraph: Structure-aware extractive summarization for scientific papers with heterogeneous graph. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 575–586, Online only. Association for Computational Linguistics.

Yifu Qiu, Varun Embar, Shay Cohen, and Benjamin Han. 2024. Think while you write: Hypothesis verification promotes faithful knowledge-to-text generation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1628–1644, Mexico City, Mexico. Association for Computational Linguistics.

Jamil S Samaan, Samuel Margolis, Nitin Srinivasan, Apoorva Srinivasan, Yee Hui Yeo, Rajsavi Anand, Fadi S Samaan, James Mirocha, Seyed Amir Ahmad Safavi-Naini, Bara El Kurdi, and 1 others. 2024. Multimodal large language model passes specialty board examination and surpasses human test-taker scores: A comparative analysis examining the stepwise impact of model prompting strategies on performance. In *medRxiv*.

Lei Shen, Haolan Zhan, Xin Shen, Hongshen Chen, Xiaofang Zhao, and Xiaodan Zhu. 2021. Identifying untrustworthy samples: Data filtering for open-domain dialogues with bayesian optimization. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1598–1608.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Amir Soleimani, Vassilina Nikoulina, Benoit Favre, and Salah Ait Mokhtar. 2022. Zero-shot aspect-based scientific document summarization using self-supervised pre-training. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 49–62, Dublin, Ireland. Association for Computational Linguistics.

Nishant Subramani, Nivedita Suresh, and Matthew Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 566–581, Dublin, Ireland. Association for Computational Linguistics.

Sotaro Takeshita, Tommaso Green, Ines Reinig, Kai Eckert, and Simone Ponzetto. 2024. ACLSum: A new dataset for aspect-based summarization of scientific publications. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6660–6675, Mexico City, Mexico. Association for Computational Linguistics.

Bowen Tan, Lianhui Qin, Eric Xing, and Zhiting Hu. 2020. Summarizing text on any aspects: A knowledge-informed weakly-supervised approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6301–6309, Online. Association for Computational Linguistics.

An Tang, Xiuzhen Zhang, and Minh Dinh. 2024. Aspect-based key point analysis for quantitative summarization of reviews. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1419–1433, St. Julian's, Malta. Association for Computational Linguistics.

Ran Tian, Shashi Narayan, Thibault Sellam, and Ankur P Parikh. 2019. Sticking to the facts: Confident decoding for faithful data-to-text generation. *arXiv preprint arXiv:1910.08684*.

Neeraj Varshney, Wenlin Yao, Hongming Zhang, Jianshu Chen, and Dong Yu. 2023. A stitch in time saves nine: Detecting and mitigating hallucinations of llms by validating low-confidence generation. *arXiv preprint arXiv:2307.03987*.

David Wan, Mengwen Liu, Kathleen McKeown, Markus Dreyer, and Mohit Bansal. 2023. Faithfulness-aware decoding strategies for abstractive summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2864–2880, Dubrovnik, Croatia. Association for Computational Linguistics.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *The Twelfth International Conference on Learning Representations*.

Charles Yu, Sullam Jeoung, Anish Kasi, Pengfei Yu, and Heng Ji. 2023. Unlearning bias in language models by partitioning gradients. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6032–6048, Toronto, Canada. Association for Computational Linguistics.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024a. R-tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024b. How language model hallucinations can snowball. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 59670–59684. PMLR.

Xiaoying Zhang, Baolin Peng, Ye Tian, Jingyan Zhou, Lifeng Jin, Linfeng Song, Haitao Mi, and Helen Meng. 2024c. Self-alignment for factuality: Mitigating hallucinations in LLMs via self-evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1965, Bangkok, Thailand. Association for Computational Linguistics.

Xingjian Zhang, Yutong Xie, Jin Huang, Jinge Ma, Zhaoying Pan, Qijia Liu, Ziyang Xiong, Tolga Ergen, Dongsub Shim, Honglak Lee, and Qiaozhu Mei. 2025a. MASSW: A new dataset and benchmark tasks for AI-assisted scientific workflows. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2373–2394, Albuquerque, New Mexico. Association for Computational Linguistics.

Yiming Zhang, Jianfeng Chi, Hailey Nguyen, Kartikeya Upasani, Daniel M. Bikel, Jason Weston, and Eric Michael Smith. 2024d. Backtracking improves generation safety. *Preprint*, arXiv:2409.14586.

Yuji Zhang, Sha Li, Cheng Qian, Jiateng Liu, Pengfei Yu, Chi Han, Yi R. Fung, Kathleen McKeown, Chengxiang Zhai, Manling Li, and Heng Ji. 2025b. The law of knowledge overshadowing: Towards understanding, predicting, and preventing llm hallucination. In *Proc. The 63rd Annual Meeting of the Association for Computational Linguistics (ACL2025) Findings*.

# A Appendix

## A.1 Example Aspects and Associated Summaries

As an example, the ground-truth aspects and summaries for Everaert et al. (2023) (open access online in the ASHA portal) are shown in Table 6.

## A.2 Example Segmentation

An example of segmenting a portion of Adler-Bock et al. (2007) is shown in Figure 5.

## A.3 Interoperability

The core of LORE is adding a rewriting module into the inference stage of a language model. Notably, this means that the actual generation model,

| Aspect | Summary |
|---|---|
| Purpose | This study evaluated whether developmental reading disability could be predicted in children at the age of 30 months, according to 3 measures of speech production: speaking rate, articulation rate, and the proportion of speaking time allocated to pausing. |
| Method | Speech samples of 18 children at high risk and 10 children at low risk for reading disability were recorded at 30 months of age. High risk was determined by history of reading disability in at least 1 of the child's parents. In grade school, a reading evaluation identified 9 children within the high-risk group as having reading disability and 9 children as not having reading disability. The 10 children at low risk for reading disability tested negative for reading disability. |
| Result | Children with reading disability showed a significantly slower speaking rate than children at high risk without reading disability. Children with reading disability allocated significantly more time to pausing, as compared with the other groups. Articulation rate did not differ significantly across groups. |
| Conclusion | Speaking rate and the proportion of pausing time to speaking time may provide an early indication of reading outcome in children at high risk for reading disability. |

Table 6: These aspects are extracted from "headline"-style abstracts for a subset of ASHA journal articles.

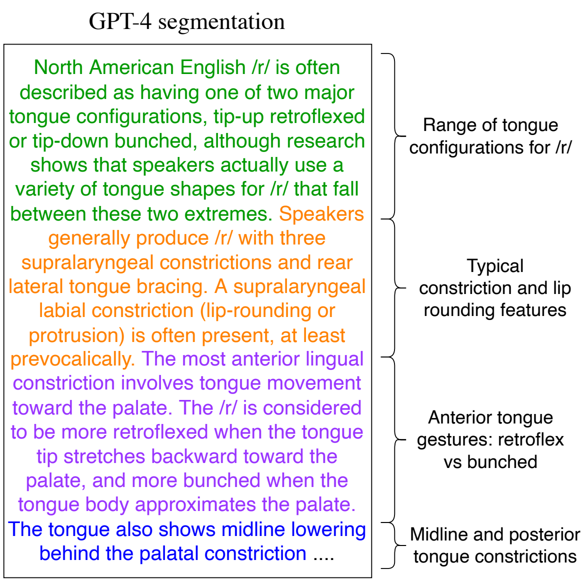


Figure 5: GPT-4, when prompted to segment the document based on broad topics, generates slightly coarser-grained segments than sentence-level segmentation.

assuming it's a Transformer, can be switched to whatever model works best for the task at hand. Furthermore, any prompting procedure can and should be modified accordingly, and similarly, for certain tasks, any other hypothesis verification model could potentially be used in place of the NLI model.

In our study, we use a pretrained and open-source NLI model, `roberta-large-mnli` (Liu et al., 2019), to determine if the relevant segments of the source document entail the generation. For generation, we use the open-source LLM `DeepSeek-R1-Distill-Qwen-1.5B` and generated aspect summaries by concatenating the source document and a prompt containing the aspect: "Given the following source document, generate a summary of the document's Purpose."

### A.4 Rewriting Metrics

1. **Faithfulness.** Is the summary accurate in relation to the source document? Are there any hallucinations, contradictions, or misinterpretations?

2. **Comprehensiveness.** Does the summary cover all the important points from the source document that are relevant to the aspect?

3. **Conciseness.** Is the summary free of redundant information or unnecessary elaboration?

4. **Relevance.** Does the summary include only information pertinent to the original document and task? Is irrelevant or tangential information excluded?

### A.5 Evaluation Prompts

The prompts used for calculating GPTScore are based on the original prompts in Fu et al. (2024). These are detailed in Table 7. For three way comparisons, the ground-truth is scored using the $src \rightarrow summ$ template and rewrites with the $ref \rightarrow summ$ template.

| Dimension | Function | Instruction |
|---|---|---|
| Overall | $src \rightarrow summ$ | Generate a summary consistent in relation to the following text's $\{aspect\}$, ensuring it is concise, comprehensive, relevant, consistent, and has faithful details: $\{src\}$\n\nTl;dr $\{summ\}$ |
| | $ref \rightarrow summ$ | Rewrite the following text to have concise, comprehensive, relevant, consistent, and faithful details. $\{ref\}$ In other words, $\{summ\}$ |
| Faithfulness | $src \rightarrow summ$ | Generate a summary consistent in relation to the following text's $\{aspect\}$, avoid making details up, contradictions, and misinterpretations: $\{src\}$\n\nTl;dr $\{summ\}$ |
| | $ref \rightarrow summ$ | Rewrite the following text with consistent and faithful details. $\{ref\}$ In other words, $\{summ\}$ |
| Comprehensiveness | $src \rightarrow summ$ | Generate a comprehensive summary in relation to the following text's $\{aspect\}$: $\{src\}$\n\nTl;dr $\{summ\}$ |
| | $ref \rightarrow summ$ | Rewrite the following text with comprehensive details. $\{ref\}$ In other words, $\{summ\}$ |
| Conciseness | $src \rightarrow summ$ | Generate a concise summary in relation to the following text's $\{aspect\}$: $\{src\}$\n\nTl;dr $\{summ\}$ |
| | $ref \rightarrow summ$ | Rewrite the following text with comprehensive details. $\{ref\}$ In other words, $\{summ\}$ |
| Relevance | $src \rightarrow summ$ | Generate a relevant summary with consistent details in relation to the following text's $\{aspect\}$: $\{src\}$\n\nTl;dr $\{summ\}$ |
| | $ref \rightarrow summ$ | Rewrite the following text with relevant details. $\{ref\}$ In other words, $\{summ\}$ |

Table 7: Instruction design on different dimensions. $aspect$ represents the name of the aspect, $src$ is the source document, $ref$ is the ground-truth summary, and $summ$ is the generated summary. These dimensions are further detailed in Appendix A.4.