Mur, M., Bandettini, P. A., and Kriegeskorte, N. Revealing representational content with pattern-information fmri—an introductory guide. *Social cognitive and affective neuroscience*, 4(1):101–109, 2009.

Nan, F., Nallapati, R., Wang, Z., Santos, C. N. d., Zhu, H., Zhang, D., McKeown, K., and Xiang, B. Entity-level factual consistency of abstractive text summarization. *arXiv preprint arXiv:2102.09130*, 2021.

Nanda, N., Lee, A., and Wattenberg, M. Emergent linear representations in world models of self-supervised sequence models. *arXiv preprint arXiv:2309.00941*, 2023.

Narayan, S., Cohen, S. B., and Lapata, M. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*, 2018.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Rai, D., Zhou, Y., Feng, S., Saparov, A., and Yao, Z. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024. URL https://arxiv.org/abs/2407.02646.

Research, A. Hallushield: A mechanistic approach to hallucination–resistant models. https://apartresearch.com/project/hallushield-a-mechanistic-approach-to-hallucination-resistant-models, 2025. White paper.

See, A., Liu, P. J., and Manning, C. D. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*, 2017.

Simhi, A., Herzig, J., Szpektor, I., and Belinkov, Y. Constructing benchmarks and interventions for combating hallucinations in llms. *arXiv preprint arXiv:2404.09971*, 2024.

Simhi, A., Itzhak, I., Barez, F., Stanovsky, G., and Belinkov, Y. Trust Me, I'm Wrong: High-Certainty Hallucinations in LLMs. *arXiv preprint arXiv:2502.12964*, 2025. URL https://arxiv.org/abs/2502.12964.

Slobodkin, A., Goldman, O., Caciularu, A., Dagan, I., and Ravfogel, S. The curious case of hallucinatory (un) answerability: Finding truths in the hidden states of over-confident large language models. *arXiv preprint arXiv:2310.11877*, 2023.

Sun, Z., Zang, X., Zheng, K., Song, Y., Xu, J., Zhang, X., Yu, W., and Li, H. Redeep: Detecting hallucination in retrieval–augmented generation via mechanistic interpretability. *arXiv preprint arXiv:2410.11414*, 2024. URL https://arxiv.org/abs/2410.11414.

Valentin, S., Fu, J., Detommaso, G., Xu, S., Zappella, G., and Wang, B. Cost-Effective Hallucination Detection for LLMs, 2024. URL https://arxiv.org/abs/2407.21424.

Yu, L., Cao, M., Cheung, J. C. K., and Dong, Y. Mechanistic understanding and mitigation of language model non–factual hallucinations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7943–7956, Miami, USA, 2024. doi: 10.18653/v1/2024.findings-emnlp.466. URL https://aclanthology.org/2024.findings-emnlp.466.

Yuksekgonul, M., Chandrasekaran, V., Jones, E., Gunasekar, S., Naik, R., Palangi, H., Kamar, E., and Nushi, B. Attention satisfies: A constraint-satisfaction lens on factual errors of language models, 2024. URL https://arxiv.org/abs/2309.15098.

A Generation Prompts

This section details the prompts used to generate hallucinated and factual continuations for our datasets. We used the OpenAI API with GPT-4.1-mini to create both hallucinated and factual continuations by providing specific instructions through system and user messages.

A.1 Hallucination Generation Prompts

For generating hallucinated continuations (introducing unsupported or contradictory information), we used the following system and user prompts:

```
# System Message
You are an expert summarization assistant helping to create a hallucination dataset.
Your task is to produce exactly ONE sentence that *appears* relevant to the given article
    but introduces at least one factual detail that cannot be inferred from the article.
The sentence must be grammatically correct, and <= 40 words.

# User Message
Original article:
<article>
{document}
</article>

Original correct summary:
<summary>
{original_summary}
</summary>

We are going to replace this sentence: <sentence_to_replace>{replaced_sentence}</
    sentence_to_replace> with a hallucinated sentence that you generate.

The new sentence must contain a made up factual detail that is not present in the original
     article.

Return JUST the new sentence, without quotation marks, xml tags, or any additional
    commentary.
```

A.2 Factual Generation Prompts

For generating factual continuations (containing only information supported by the source), we used the following system and user prompts:

```
# System Message
You are an expert summarization assistant helping to create a dataset.
You will be given an existing summary and asked to replace a single sentence with one
    written by you.
The sentence you write must be factually accurate, grammatically correct, and only contain
     information directly from the original article.
Your sentence should be concise and <= 40 words.

# User Message
Original article:
<article>
{document}
</article>

Original correct summary:
<summary>
{original_summary}
</summary>

We are going to replace this sentence: <sentence_to_replace>{replaced_sentence}</
    sentence_to_replace> with a new sentence that you generate.
```

```
The new sentence must be completely factual, only containing information directly from the
     original article.

Return JUST the new sentence, without quotation marks, xml tags, or any additional
    commentary.
```

A.3 Hallucination Evaluation Prompt

For evaluating hallucination rates during the steering experiments in Section 4.4, we used GPT-4.1 with the following prompt:

```
You are an expert at detecting hallucinations in summaries.

I will provide you with an article and a summary. Your task is to determine if the summary
     contains any hallucinations - information that is not supported by or contradicts the
     article.

A hallucination in a summary is when:
1. It contains information NOT present in the article
2. It contradicts information in the article
3. It makes claims that require outside knowledge not in the article

Answer with ONLY "hallucination" or "no hallucination".

Article:
{article}

Summary:
{summary}
```

These prompts were implemented in our data generation pipeline to create paired examples of source contexts and either factual or hallucinated continuations across all datasets described in Section 3.1.

B ContraTales Dataset Generation

The ContraTales dataset, comprising 2000 examples of stories with purely logical contradictions, was generated to test the contextual understanding of hallucination detectors. The generation process involved using a large language model, specifically `o4-mini`, guided by a detailed prompt that included instructional guidelines and few-shot examples.

B.1 Generation Process

Each example in ContraTales consists of three main parts:

1. **Story Prefix**: An initial narrative segment (typically 7-10 sentences). The first sentence of this prefix establishes an unambiguous constraint or fact about a character or situation (e.g., “Jack had been bald for 10 years,” “Sarah was allergic to peanuts”). The subsequent sentences in the prefix develop a neutral, everyday scene, intentionally avoiding any direct reference to, or negation of, the established constraint.
2. **Correct Concluding Sentence**: A single sentence that provides a logical and coherent continuation of the story prefix, respecting the initial constraint.
3. **Contradictory Concluding Sentence**: A single sentence that is structurally and tonally similar to the correct concluding sentence but introduces a detail that subtly and logically contradicts the constraint established in the first sentence of the story prefix.

The generation was performed using the `o4-mini` model. The model was prompted with a set of instructions that specified the desired structure and characteristics of the ContraTales examples. To guide the model effectively, the prompt also included a random selection of 5 few-shot examples from a seed set of 24 pre-existing ContraTales. These examples