Article:

Chrysostomou, G., Zhao, Z., Williams, M. et al. (1 more author) (2024) Investigating hallucinations in pruned large language models for abstractive summarization. Transactions of the Association for Computational Linguistics, 12. pp. 1163-1181. ISSN 2307-387X

https://doi.org/10.1162/tacl_a_00695

Investigating Hallucinations in Pruned Large Language Models for Abstractive Summarization

George Chrysostomou[*♣†] **Zhixue Zhao**[*♢] **Miles Williams**[*♢] **Nikolaos Aletras**[♢]
[♢]University of Sheffield, UK [♣]AstraZeneca, UK
{zhixue.zhao, mwilliams15, n.aletras}@sheffield.ac.uk

Abstract

Despite the remarkable performance of generative large language models (LLMs) on abstractive summarization, they face two significant challenges: their considerable size and tendency to hallucinate. Hallucinations are concerning because they erode reliability and raise safety issues. Pruning is a technique that reduces model size by removing redundant weights, enabling more efficient sparse inference. Pruned models yield downstream task performance comparable to the original, making them ideal alternatives when operating on a limited budget. However, the effect that pruning has upon hallucinations in abstractive summarization with LLMs has yet to be explored. In this paper, we provide an extensive empirical study across five summarization datasets, two state-of-the-art pruning methods, and five instruction-tuned LLMs. Surprisingly, we find that hallucinations are less prevalent from pruned LLMs than the original models. Our analysis suggests that pruned models tend to depend more on the source document for summary generation. This leads to a higher lexical overlap between the generated summary and the source document, which could be a reason for the reduction in hallucination risk.[1]

1 Introduction

Abstractive summarization is the task of distilling the key information from a document into a summary that may contain novel text not present in the original document (Cohn and Lapata, 2008; Saggion and Poibeau, 2013; Lin and Ng, 2019). Generative large language models (LLMs) have demonstrated strong performance on abstractive summarization (Ouyang et al., 2022; Touvron etal., 2023; Almazrouei et al., 2023; OpenAI et al., 2024; Zhang et al., 2024). However, they face two significant challenges: Their substantial size requires extensive computational resources for training and inference; and they tend to hallucinate, i.e., generate nonfactual contents not supported by the source document (Zhao et al., 2020; Xu et al., 2023). Figure 1 shows an illustrative example of hallucinated content in a generated summary.

On the one hand, hallucinations not only undermine the performance of models but also introduce critical safety risks, ultimately eroding the trust of end users (Milintsevich and Agarwal, 2023; Tang et al., 2023a; Narayan et al., 2023; Zhao and Shan, 2024). For example, LLM-generated summaries in the legal or health domain can contain inaccurate information that poses real-life harms (Zhao et al., 2022a; Weidinger et al., 2022).

On the other hand, LLMs such as GPT-3.5 (Ouyang et al., 2022), GPT-4 (OpenAI et al., 2024), and Llama-2 (Touvron et al., 2023) demand substantial hardware resources. As an indication, GPT-3 (175B) requires at least five NVIDIA A100 GPUs with 80GB of memory each for half-precision inference (Frantar and Alistarh, 2023). This creates barriers for those without access to costly computational resources, ultimately hindering inclusivity in NLP (Schwartz et al., 2020; Weidinger et al., 2022). To tackle this issue, pruning techniques enable efficient sparse inference by removing redundant weights, while maintaining comparable performance (Sun et al., 2024). Pruned models therefore appear as attractive alternatives for abstractive summarization when computational resources are constrained.

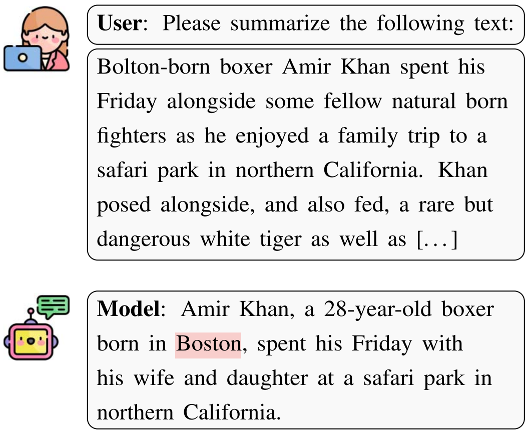
In abstractive summarization, model hallucinations are a thoroughly studied subject (Cao et al., 2020; Durmus et al., 2020; Raunak et al., 2021; Narayan et al., 2023; Laban et al., 2023). Similarly, the effect of pruning on model performance in abstractive summarization benchmarks was also explored more recently (Dun et al., 2023; Jaiswal et al., 2024). However, the relationship between pruning and hallucination

[*] Equal contribution.

[†] Work done independently of AstraZeneca.

[1] https://github.com/casszhao/PruneHall.

Transactions of the Association for Computational Linguistics, vol. 12, pp. 1163–1181, 2024. https://doi.org/10.1162/tacl_a_00695
Action Editor: Wenjie (Maggie) Li. Submission batch: 3/2024; Revision batch: 4/2024; Published 9/2024.


User: Please summarize the following text:

Bolton-born boxer Amir Khan spent his Friday alongside some fellow natural born fighters as he enjoyed a family trip to a safari park in northern California. Khan posed alongside, and also fed, a rare but dangerous white tiger as well as [...]

Model: Amir Khan, a 28-year-old boxer born in Boston, spent his Friday with his wife and daughter at a safari park in northern California.

Figure 1: An example of a hallucination (highlighted text) in abstractive summarization.

risk has yet to be explored. Given the appeal of greater efficiency with comparable downstream performance it is important to establish how trustworthy summaries generated from pruned models are. Therefore, we seek to answer the following question: *Are hallucinations more or less prevalent in LLMs after pruning?*

To this end, we empirically investigate the risk of generating hallucinated content in pruned models across five LLMs, two state-of-the-art pruning methods, and five summarization datasets. Surprisingly, our results show that pruned models are less prevalent in hallucinations compared to the original LLM. To understand this phenomenon, we further investigate the impact of different sparsity levels on hallucination patterns. Our analysis shows that hallucination risk decreases as sparsity increases, regardless of the pruning methods tested. Furthermore, our results suggest that pruning encourages the model to rely more on the source document during generation, resulting in summaries that are lexically more similar to the source document.

2 Related Work

2.1 Hallucinations in Summarization

In abstractive summarization, a model is expected to generate a concise summary of the source document. However, prior work observed that models tend to generate hallucinatory content that is not based on or cannot be entailed from the source document (Vinyals and Le, 2015; Rohrbach et al., 2018; Cao et al., 2018; Maynez et al., 2020; Raunak et al., 2021; Falke et al., 2019; Maynez et al., 2020; Zhao et al., 2022b; Chen et al., 2022). For example, Falke et al. (2019) found that 25% of the model generated summaries contain hallucinated content. On the other hand, automatic summary quality evaluation metrics such as ROUGE (Lin, 2004) and BERTScore (Zhang et al., 2020) do not correlate with the degree of hallucinations appearing in summaries (Zhou et al., 2021). For instance, Zhou et al. (2021) show that even if a summary contains a large amount of hallucinatory content, it can still achieve a high ROUGE score. This has opened up new research directions that develop approaches to detect and evaluate hallucinations (Zhou et al., 2021; Durmus et al., 2020; Guerreiro et al., 2023; Ji et al., 2023), as well as mitigate them (Xiao and Wang, 2021; Choubey et al., 2023; King et al., 2022).

2.2 Measuring Hallucination Risk

Evaluation metrics for measuring hallucination risk can be broadly categorized as: (a) entailment-based, (b) question-answering (QA), and (c) text-generation based. Entailment-based methods (Kryscinski et al., 2020; Laban et al., 2022) use pre-trained language models to compute the entailment score between the source and the generated summary. The higher the entailment score, the more consistent a summary is with respect to the source. QA methods decompose the task to a question answering problem (Wang et al., 2020; Deutsch et al., 2021; Durmus et al., 2020). Finally, text-generation based methods use off-the-shelf models to quantify the risk of hallucinations (Yuan et al., 2021; Son et al., 2022). A representative approach is the Hallucination Risk Measurement (HaRiM$^+$), which uses the log-likelihoods from a reference-free decoder model to estimate hallucination risk in a summary at the token level (Son et al., 2022). More recently, Laban et al. (2023) examined instruction-tuned LLMs as reasoners for factual assessments (i.e., assessors of hallucination prevalence) in abstractive text summarization. They demonstrated that many of these LLMs struggle to compete with previous entailment-based methods.

2.3 Pruning Large Language Models

Model compression is the task of reducing the memory footprint of a model (Ganesh et al., 2021). Pruning is a popular technique that removes redundant weights from the model (LeCun et al., 1989). Weights may be removed individually (unstructured pruning), according to defined blocks (semi-structured pruning), or in relation to model components (structured pruning) (Blalock et al., 2020; Mishra et al., 2021; Ma et al., 2023).

As the size of LLMs surpasses billions of parameters, pruning techniques that require re-training become impractical. Instead, post-training compression aims to reduce model size using only a small calibration dataset (Nagel et al., 2020; Williams and Aletras, 2023). In this setting, Frantar and Alistarh (2022) define the layer-wise compression problem, with the aim of creating a compressed version of a given layer that functions as closely as possible to the original. State-of-the-art post-training pruning techniques, such as SparseGPT (Frantar and Alistarh, 2023) and Wanda (Sun et al., 2024), build upon this, offering layer-wise solutions. SparseGPT introduces an efficient approximation that relies upon an iterative weight update process using Hessian inverses, inspired by Optimal Brain Surgeon (Hassibi et al., 1993). Wanda further improves upon efficiency by avoiding a weight update procedure, enabling pruning in a single forward pass.

In practice, the sparsity induced by pruning enables substantial improvements in inference performance across a variety of hardware. On a CPU, Frantar and Alistarh (2023) demonstrate a 1.82× speedup with 50% unstructured sparsity, using the DeepSparse engine (Neural Magic, 2021). Separately, they observe a 1.54-1.79× speedup for feed-forward layers on an NVIDIA Ampere GPU, using 2:4 semi-structured sparsity (Mishra et al., 2021).

Recent pruning approaches (such as SparseGPT and Wanda) can be applied to decoder-only LLMs with minimal impact upon common-sense reasoning (Sun et al., 2024) or summarization performance (Jaiswal et al., 2024). Interestingly, related studies suggest that pruning can reduce social bias and toxicity (Xu and Hu, 2022) and improve resilience to 'jailbreaking' attacks (Hasan et al., 2024). However, it remains unclear how pruning affects hallucination risk in LLMs.

3 Methodology

3.1 Models

We experiment with the following publicly available LLMs: (1) the **Llama-2** (Touvron et al., 2023) model family (7B, 13B, and 70B); (2) **Mistral** 7B (v0.1) (Jiang et al., 2023); (3) **Falcon** 7B (Almazrouei et al., 2023); and (4) the **OPT-IML** (Iyer et al., 2023) model family (1.3B and 30B).

We opt for decoder-only instruction-tuned models due to their efficacy in zero-shot abstractive summarization tasks (Tang et al., 2023b; Adams et al., 2023; Laskar et al., 2023).

3.2 Pruning Methods

We consider three different pruning methods: one standard baseline (layer-wise magnitude) and two state-of-the-art techniques (SparseGPT and Wanda). Formally, these pruning methods provide a saliency score $\mathbf{S}_{ij}$ for each element of the weight matrix $\mathbf{W}_{ij}$ in a given layer. The elements corresponding to the k smallest saliency scores are the target weights to be pruned, where k is determined by the sparsity ratio. The primary distinction between our selected pruning methods lies in their saliency score calculation metrics. In a post-training setting, pruning metrics can additionally incorporate layer activations, $\mathbf{X}$. The activations for each layer of the model are computed through performing a forward pass with the calibration data. We follow Sun et al. (2024) in using the same calibration data for each model, specifically 128 examples randomly sampled from C4 (Raffel et al., 2020).

Magnitude (Hagiwara, 1994; Han et al., 2015) To offer a lower bound for the performance of pruned models, we employ layer-wise weight magnitude pruning. Here, the saliency score is simply the magnitude of each weight:

$$\mathbf{S}_{ij} = |\mathbf{W}_{ij}|$$

SparseGPT (Frantar and Alistarh, 2023) The SparseGPT algorithm is an iterative procedure that offers an efficient approximation to the exact layer reconstruction. The effective saliency criterion is

$$\mathbf{S}_{ij} = \left[|\mathbf{W}|^2/\mathrm{diag}\left((\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I})^{-1}\right)\right]_{ij}$$

where λ is a dampening parameter to enable inversion of the Hessian, $\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}$.[2]

Wanda (Sun et al., 2024) In contrast, Wanda avoids a computationally expensive weight update procedure, instead relying upon only the weight magnitudes and norm of the input activations:

$$\mathbf{S}_{ij} = |\mathbf{W}_{ij}| \cdot ||\mathbf{X}||_2$$

This approximates SparseGPT when considering only diagonal elements of the Hessian for $\lambda = 0$.

[2] We follow Frantar and Alistarh (2023) in using $\lambda = 0.01$.

#	Prompt Template
A	*Summarize in a single short paragraph the context below:* `[document]` *The summary is:* `[summary]`
B	*Summarize in a couple of sentences the document below:* `[document]` *The summary is:* `[summary]`
C	*Give me a short summary of the below:* `[document]` *The summary is:* `[summary]`

Table 1: Each prompt template consists of the task instructions (*italic*) and the source `[document]`. The LLM then generates the `[summary]`.

Sparsity Level Following previous work (Fantar and Alistarh 2023; Sun et al., 2024), we evaluate our pruning methods across both semi-structured and unstructured settings:

- **2:4 semi-structured sparsity**: Two weights in every contiguous block of four must be zero, providing a total of 50% sparsity. This sparsity pattern is required to enable hardware acceleration on GPUs (Mishra et al., 2021).
- **50% unstructured sparsity**: To enable comparison, we use a sparsity level of 50% for unstructured pruning, unless otherwise stated.

We do not explore pruning above 50% sparsity as language modeling performance collapses shortly beyond this threshold (Frantar and Alistarh, 2023; Sun et al., 2024). Maintaining language modeling performance is essential for the generation of high-quality summaries, enabling comparison between the models and their pruned counterparts.

3.3 Prompting

LLMs are known to be sensitive to prompt design (Petroni et al., 2019; Elazar et al., 2021; Fierro and Søgaard, 2022). To mitigate the effect of prompt variability, we summarize each document using three distinct prompt templates (Table 1). Each template instructs the model to summarize a given document in a slightly different manner, offering three summaries for each document. We then evaluate all three summaries by averaging the scores.

For each model family, we follow the prompt formatting used in the original work. In the case of Llama-2 and Mistral, this includes the use of `[INST]` and `[/INST]` tokens to delimit user instructions. For the Falcon and OPT-IML model families, which were not trained with a specific prompt format, we use the prompts as is (Table 1).

		Source		Reference	
Dataset	#	Mean	Max	Mean	Max
FactCC	311	634.2	1838	17.4	63
Polytope	634	575.1	1781	64.6	128
SummEval	100	407.8	589	65.1	101
Legal Contracts	85	237.8	1106	21.6	61
RCT	53	307.5	447	68.7	174

Table 2: The number of source documents in each dataset (#), and the mean and maximum length (in words) for the documents and reference summaries.

3.4 Summarization Datasets

We include the following summarization datasets: (1) **FactCC** (Kryscinski et al., 2020); (2) **Polytope** (Huang et al., 2020); (3) **SummEval** (Fabbri et al., 2021); (4) **Legal Contracts** (Manor and Li, 2019); and (5) **RCT** summaries (Wallace et al., 2021). FactCC, Polytope, and SummEval are all different subsets of the CNN/DailyMail news article dataset (Nallapati et al., 2016), covering a variety of topics. Legal Contracts consists of legal text snippets from the terms of service for various products and services. Finally, RCT combines the abstracts from randomized control trials with their corresponding human-written conclusions from systematic reviews, i.e., the conclusions are used as the target summary. For simplicity, we select instances in RCT where there is a one-to-one mapping between abstract and target summary.

We use the test set from each dataset and remove any duplicates if any exist. Table 2 provides detailed dataset statistics.

3.5 Evaluation of Summarization Quality

We evaluate the quality of generated summaries against the corresponding reference summary, using a subset of the ROUGE family of metrics (Lin, 2004) and BERTScore (Zhang et al., 2020).[3] From ROUGE, we use two n-gram overlap metrics (ROUGE-1 and ROUGE-2) and the longest sequence overlap metric (ROUGE-L).

3.6 Hallucination Risk Metrics

To automatically evaluate the hallucination risk in the generated summaries, we use standard

[3] For FactCC, we use the extracted claim as the reference.

automatic metrics that compare directly the source document and the corresponding generated summary.

HaRiM$^+$ (Son et al., 2022) HaRiM is based on the idea that over-reliance on decoder context during generation leads to hallucinations. Given a summary and a reference document, HaRiM$^+$ first uses a pre-trained sequence-to-sequence model (S2S, an encoder-decoder model) to calculate the token probabilities in the summary given the reference document as input. A pre-trained decoder-only model is used as a secondary model (Aux) to compute summary token probabilities, i.e., no input document is provided to summarize. HaRiM$^+$ therefore uses Aux token probabilities to regularize S2S token probabilities and detect hallucinations by:

$$\text{HaRiM} = \frac{1}{L}\sum_{i=0}^{L}(1 - p_{\text{S2S}})(1 - (p_{\text{S2S}} - p_{\text{Aux}}))$$

where L is the sequence length, p_{S2S} the predicted probability of a token generated by the model given the source document, and p_{Aux} the probability of the same generated token from the auxiliary model.

HaRiM$^+$ extends HaRiM through adding the S2S log-likelihood of tokens, and applying a scaling hyperparameter λ_H:[4]

$$\text{HaRiM}^+ = \frac{1}{L}\sum_{i}^{L}\log(p(y_i \mid y_{<i}; X)) - \lambda_H \text{HaRiM}$$

Intuitively, a higher HaRiM$^+$ score indicates that the summary is more likely to be faithful to the source document, i.e., less likely to contain hallucinations. Son et al. (2022) also showed that the first sequence-to-sequence model can also act as a secondary model, with equivalent performance.

SummaC (Laban et al., 2022) This metric uses an off-the-shelf entailment model to assess the consistency between a source document and a generated summary. First, the document and summary are split into sentences, with the document sentences (N) being the hypothesis and the generated summary sentences (K) being the premise. The second step is to create an $K \times N$ matrix of entailment scores from the pre-trained model. A generated sentence with a low entailment score to any of the document sentences is a potential hallucination.

SummaC$_{\text{ZS}}$ obtains the row-wise maximum entailment score, which leads to a vector E of size K. **SummaC$_{\text{Conv}}$** obtains vector E by using a convolutional model over each row K, to obtain a single score. In both metrics, each element in E can be interpreted as the consistency score for each sentence in the summary. E is averaged to obtain a single summary consistency score.

Hallucination Risk Ratio (HRR) To compare the hallucination risk of pruned models relative to the original, we compute a ratio using any one of the hallucination risk metrics:

$$\text{HRR} = \frac{\text{Hallucination Risk}_{\text{Original}}}{\text{Hallucination Risk}_{\text{Pruned}}}$$

A lower HRR indicates that the pruned model has a lower hallucination risk than the original. This contrasts the hallucination risk metrics, where a higher score indicates a lower risk for a given model.

3.7 Human Evaluation

We also conduct a human evaluation task to compare the hallucination prevalence between the original and pruned models. For this purpose, we randomly sample 100 distinct source documents from FactCC, Polytope, and SummEval. We selected these datasets because they consist of news articles, making them suitable for human evaluation without requiring extensive domain expertise. We recruited three participants who are native speakers or proficiently fluent in English. Following Lango and Dusek (2023), we ask them to answer the following questions for comparing the summaries generated by the original and pruned models:

Q1. **Hallucinations**: Which summary contains more hallucinations (i.e., content that is not supported by the source document)?

Q2. **Omission**: Which summary is missing more crucial information from the document?

Q3. **Repetition**: Which summary contains more repetitive information?

Q4. **Alignment**: Which summary is more semantically aligned with the source document?

Identifying hallucinations in text is challenging and requires careful reading and attention to nuanced facts (Laban et al., 2023). Therefore, we

[4] We follow Son et al. (2022) in using $\lambda_H = 7$.

		Magnitude		SparseGPT		Wanda	
Model	–	2:4	50%	2:4	50%	2:4	50%
Falcon 7B	19.93	303.22	482.11	52.11	37.10	85.68	38.93
Llama-2 7B	6.49	78.29	19.07	10.79	7.94	12.46	7.93
Llama-2 13B	5.71	10.73	7.98	8.68	6.80	9.58	6.94
Llama-2 70B	4.30	6.89	5.61	6.51	5.18	6.45	5.23
Mistral 7B	6.32	9.55	7.96	9.21	7.18	9.85	7.26
OPT-IML 1.3B	14.68	166.09	1391.46	24.92	18.03	25.11	17.94
OPT-IML 30B	10.56	246.42	57.88	11.61	10.74	12.44	10.74

Table 3: Perplexity (↓) of original and pruned models on the held-out set of WikiText.

first perform a calibration run on a held-out set of ten documents and their generated summaries. Two of the participants are then presented with the set of 100 original documents, alongside two generated summaries: one from a pruned model and the other from the original model. The order of the documents is shuffled and information about which model generated the summary is not disclosed to the participants. Similar to Xu et al. (2023), we use the third participant as an adjudicator for disagreements. The inter-annotator agreement is computed using Cohen's kappa IAA (κ), as the average between the two participants and the adjudicator.

3.8 Implementation Details

We use the model implementation and weights available from Hugging Face (Wolf et al., 2020). We perform experiments using either one or two NVIDIA A100 (SXM 80GB) GPUs. For the pruning methods, we use the hyperparameters from Frantar and Alistarh (2023) and Sun et al. (2024).

For summary generation we use greedy decoding (i.e., sampling the token with the highest probability) for better reproducibility. We continue to sample tokens until we reach either (a) the end of sequence token, or (b) the maximum sequence length of the model.

4 Results

4.1 Language Modeling

We first compare language modeling performance between the original and pruned models. Following Frantar and Alistarh (2023) and Sun et al. (2024), we compute perplexity on the WikiText test set (Merity et al., 2017), shown in Table 3.

Overall, pruned models consistently generate text with higher perplexity than their original counterparts. Unsurprisingly, magnitude pruning routinely produces the highest perplexity. In many cases, the increase over the original model (denoted by '-') is substantial. For example, we observe more than a twentyfold increase for OPT-IML 30B, from 10.56 to 246.42. In contrast, SparseGPT and Wanda achieve perplexity close to the original for the majority of the models. Surprisingly, Falcon 7B records higher perplexity across all pruning methods, e.g., 85.68 when applying Wanda from 19.93 without pruning.

Due to the substantial degradation in language modeling performance, we omit magnitude pruning from further analysis. For the same reason, we also exclude the Falcon 7B and OPT-IML 1.3B models.

4.2 Summarization

Table 4 shows summarization performance (ROUGE-1/2/L & BERTScore) across all datasets.[5] We first observe that the original models perform comparably for BERTScore across most datasets. For example, in Legal Contracts, Llama-2 13B records a BERTScore of 84.75 compared to 84.90 from OPT-IML 30B. We only observe larger performance deviations in the case of RCT, with the original Mistral 7B obtaining the highest BERTScore (88.46) and OPT-IML 30B the lowest (83.12). This suggests that all LLMs generate summaries that are equally semantically similar to the reference summary. Compared to BERTScore, the scores of the original models in lexical overlap metrics (ROUGE-1/2/L) differ largely not only across models, but also across datasets. For example, Llama-2 7B achieves the second highest ROUGE-L score in RCT (33.50) and the lowest score in FactCC (11.51). Similarly, in RCT, Mistral 7B records an increase of 34.65 (46.16) for ROUGE-L, making it the best performing original model for this metric.

Comparing the performance between original and pruned models, we find that they perform comparably in the majority of cases. For SparseGPT, the summaries score significantly higher (across all metrics) than those from the original model in 19 out of 100 comparisons, while they score significantly lower in 11 out of 100 (**bold** scores; paired t-test; $p < 0.05$). The results are similar for Wanda, where pruned models perform significantly higher in 20 out of 100 comparisons and significantly lower (underlined scores) in 26

[5] We obtain comparable results using 50% unstructured sparsity, which are omitted for brevity.

Dataset	Method	Llama-2 7B		Llama-2 13B		Llama-2 70B		Mistral 7B		OPT-IML 30B	
		ROUGE-1/2/L	BS	ROUGE-1/2/L	BS	ROUGE-1/2/L	BS	ROUGE-1/2/L	BS	ROUGE-1/2/L	BS
FactCC	–	13.99 / 6.41 / 11.51	84.60	15.14 / 6.39 / 12.30	84.39	15.04 / 6.29 / 12.11	84.75	14.83 / 8.21 / 12.70	84.78	23.51 / 12.68 / 20.48	85.71
	SpGPT	12.46 / 6.07 / 10.55	84.15	15.34 / 6.62 / 12.75	**84.76**	14.78 / 6.80 / 12.29	84.68	14.43 / 8.52 / 12.62	84.45	18.52 / 12.05 / 16.89	85.04
	Wanda	11.04 / 5.94 / 9.53	80.57	15.64 / 7.32 / 13.09	**84.78**	15.09 / 6.88 / 12.47	84.72	13.67 / 8.30 / 12.02	84.34	17.91 / 11.68 / 16.38	83.94
Polytope	–	38.92 / 18.19 / 25.86	85.41	38.63 / 17.51 / 25.34	84.91	39.28 / 17.48 / 25.78	85.48	40.27 / 22.69 / 28.65	85.63	33.06 / 22.81 / 27.74	86.54
	SpGPT	33.98 / 18.14 / 24.45	84.88	35.99 / 16.74 / 25.01	85.01	38.16 / 18.51 / 25.89	85.31	39.07 / **24.21** / 29.54	85.58	33.39 / **26.32** / 29.02	**87.01**
	Wanda	30.88 / 15.39 / 21.77	83.09	37.33 / **19.29** / **26.68**	**85.23**	38.74 / **18.80** / 26.58	85.42	37.08 / 23.78 / 28.76	85.34	30.14 / 22.72 / 25.85	86.03
SummEval	–	40.39 / 18.73 / 26.61	85.42	40.36 / 18.00 / 25.88	84.78	41.52 / 18.78 / 26.82	85.58	43.94 / 26.34 / 32.04	86.05	51.93 / 36.55 / 41.38	86.94
	SpGPT	38.77 / **23.04** / 27.81	85.36	40.55 / 18.42 / 27.15	**85.33**	41.58 / 19.69 / 27.65	85.61	43.77 / 28.00 / 33.33	86.03	50.00 / 37.16 / 41.64	86.73
	Wanda	37.78 / **23.95** / 28.82	85.12	**44.31** / **23.51** / **31.58**	**86.03**	41.57 / 19.44 / 27.67	85.57	45.11 / 29.95 / 34.84	86.22	44.48 / 33.57 / 36.90	86.12
Legal Contracts	–	18.75 / 6.20 / 13.93	84.73	21.12 / 6.90 / 15.41	84.75	21.66 / 7.07 / 16.19	85.60	17.52 / 6.21 / 13.70	84.78	22.96 / 7.45 / 18.30	84.90
	SpGPT	16.84 / 5.98 / 12.80	84.17	18.99 / 6.11 / 14.41	84.90	21.74 / 7.42 / 16.73	85.33	18.56 / 6.90 / 14.51	84.76	21.18 / 7.22 / 17.15	84.49
	Wanda	14.22 / 4.94 / 11.14	81.52	18.80 / 6.37 / 14.53	84.41	22.13 / 7.51 / 16.72	85.55	18.14 / 6.37 / 13.83	84.79	19.10 / 6.79 / 15.36	81.86
RCT	–	45.29 / 26.89 / 33.50	86.97	39.87 / 22.01 / 28.56	86.43	37.79 / 20.98 / 28.05	86.25	53.66 / 40.66 / 46.16	88.46	24.62 / 18.20 / 21.33	83.12
	SpGPT	**50.57** / **37.40** / **43.12**	**87.89**	37.81 / 22.40 / 29.37	86.26	**40.19** / **25.35** / **31.97**	**86.57**	**56.93** / **47.79** / **52.45**	**89.17**	25.22 / **21.50** / 23.61	77.39
	Wanda	38.79 / 28.59 / 33.12	86.06	36.90 / 23.07 / 28.82	86.11	**39.61** / **24.79** / **31.60**	86.49	**59.29** / **50.02** / **54.83**	**89.40**	**31.59** / **28.84** / **30.49**	70.64

Table 4: ROUGE-1/2/L (↑) and BERTScore (BS; ↑) for the original models (–) and their pruned counterparts (SparseGPT and Wanda). Values in **bold** indicate that the pruned model scores significantly higher than the original while underlined values denote a significantly lower score (paired t-test; $p < 0.05$).

		Llama-2 7B				Llama-2 13B				Llama-2 70B				Mistral 7B				OPT-IML 30B			
		SparseGPT		Wanda		SparseGPT		Wanda		SparseGPT		Wanda		SparseGPT		Wanda		SparseGPT		Wanda	
Dataset	Metric	2:4	50%	2:4	50%	2:4	50%	2:4	50%	2:4	50%	2:4	50%	2:4	50%	2:4	50%	2:4	50%	2:4	50%
FactCC	HaRiM$^+$	**0.98**	**0.95**	**0.94**	**0.95**	**0.77**	**0.95**	**0.69**	**0.91**	**0.93**	**0.96**	**0.93**	**0.96**	**0.93**	**0.94**	**0.91**	**0.94**	**0.83**	**0.87**	**0.87**	**0.85**
	SummaC$_{\text{conv}}$	**0.64**	**0.82**	**0.56**	**0.81**	**0.76**	**0.83**	**0.64**	**0.84**	**0.76**	**0.92**	**0.77**	**0.90**	**0.79**	**0.88**	**0.74**	**0.86**	**0.80**	**0.86**	**0.84**	**0.83**
	SummaC$_{\text{zs}}$	**0.47**	**0.65**	**0.39**	**0.65**	**0.50**	**0.61**	**0.41**	**0.61**	**0.63**	**0.86**	**0.63**	**0.83**	**0.76**	**0.85**	**0.68**	**0.82**	**0.80**	**0.87**	**0.85**	**0.83**
Polytope	HaRiM$^+$	**0.97**	**0.97**	**0.97**	**0.97**	**0.78**	**0.93**	**0.71**	**0.85**	**0.94**	0.96	**0.95**	1.00	**0.95**	**0.95**	**0.94**	**0.96**	**0.87**	**0.93**	**0.92**	**0.88**
	SummaC$_{\text{conv}}$	**0.67**	**0.83**	**0.69**	**0.83**	**0.70**	**0.78**	**0.65**	**0.79**	**0.77**	**0.93**	**0.78**	**0.92**	**0.78**	**0.82**	**0.76**	**0.84**	**0.86**	**0.95**	**0.91**	**0.92**
	SummaC$_{\text{zs}}$	**0.64**	**0.85**	**0.64**	**0.75**	**0.58**	**0.69**	**0.56**	**0.69**	**0.75**	**0.88**	**0.74**	**0.83**	**0.76**	**0.81**	**0.75**	**0.84**	**0.88**	**0.95**	**0.92**	**0.93**
SummEval	HaRiM$^+$	**0.88**	**0.93**	**0.81**	**0.93**	**0.80**	**0.97**	**0.69**	**0.96**	**0.95**	0.98	**0.95**	0.98	**0.93**	**0.94**	**0.92**	**0.95**	**0.91**	**0.92**	**0.90**	**0.89**
	SummaC$_{\text{conv}}$	**0.55**	**0.81**	**0.46**	**0.76**	**0.67**	**0.81**	**0.59**	**0.81**	**0.78**	0.96	**0.79**	0.93	**0.79**	**0.85**	**0.77**	**0.87**	**0.86**	**0.88**	**0.83**	**0.85**
	SummaC$_{\text{zs}}$	**0.49**	**0.75**	**0.4**	**0.68**	**0.56**	**0.71**	**0.49**	**0.66**	**0.70**	**0.92**	**0.70**	**0.88**	**0.79**	**0.84**	**0.76**	**0.88**	**0.86**	**0.89**	**0.85**	**0.86**
Legal Contracts	HaRiM$^+$	**0.99**	**0.85**	**0.90**	**0.85**	**0.83**	**0.88**	**0.76**	**0.88**	**0.87**	**0.92**	**0.89**	**0.95**	**0.85**	**0.94**	**0.89**	**0.93**	**0.85**	**0.89**	**0.81**	**0.83**
	SummaC$_{\text{conv}}$	0.98	0.85	0.93	0.94	**0.82**	**0.81**	**0.76**	**0.81**	**0.79**	**0.88**	**0.83**	**0.91**	**0.83**	**0.92**	**0.92**	**0.89**	**0.85**	**0.88**	**0.81**	**0.86**
	SummaC$_{\text{zs}}$	1.01	**0.86**	0.96	**0.90**	**0.93**	**0.86**	**0.88**	**0.88**	**0.85**	0.93	**0.88**	0.95	**0.88**	**0.92**	**0.93**	**0.92**	**0.93**	0.96	**0.94**	1.00
RCT	HaRiM$^+$	**0.92**	**0.96**	**0.87**	**0.92**	**0.86**	**0.99**	**0.80**	**0.97**	**0.93**	**0.96**	**0.93**	**0.97**	**0.93**	**0.96**	**0.93**	**0.95**	**0.85**	**0.88**	**0.83**	**0.87**
	SummaC$_{\text{conv}}$	**0.69**	**0.86**	**0.70**	**0.88**	**0.78**	**0.89**	**0.79**	**0.88**	**0.82**	**0.92**	**0.82**	**0.93**	**0.82**	**0.88**	**0.81**	**0.87**	**0.83**	**0.88**	**0.79**	**0.88**
	SummaC$_{\text{zs}}$	**0.71**	**0.83**	**0.71**	**0.82**	**0.69**	**0.81**	**0.70**	**0.82**	**0.79**	**0.90**	**0.79**	**0.90**	**0.84**	**0.89**	**0.82**	**0.89**	**0.77**	**0.80**	**0.77**	**0.83**
Average	HaRiM$^+$	0.95	0.93	0.90	0.92	0.81	0.95	0.73	0.91	0.92	0.96	0.93	0.97	0.92	0.95	0.92	0.95	0.87	0.90	0.87	0.87
	SummaC$_{\text{conv}}$	0.70	0.83	0.67	0.85	0.74	0.82	0.68	0.83	0.78	0.92	0.80	0.92	0.80	0.87	0.80	0.87	0.84	0.89	0.84	0.87
	SummaC$_{\text{zs}}$	0.67	0.79	0.62	0.76	0.65	0.74	0.61	0.73	0.74	0.90	0.75	0.88	0.81	0.86	0.79	0.87	0.85	0.90	0.86	0.89

Table 5: Hallucination risk ratio (HRR) between the original and the pruned model (values less than one are highlighted, indicating that the pruned model has a lower hallucination risk than the original model), averaged across all data points over the three prompts for each dataset. **Bold** values denote significant differences between the pruned and the original model (paired t-test; $p < 0.05$).

out of 100. We also find that models pruned with SparseGPT perform more consistently compared to those pruned using Wanda. For example, Llama-2 7B pruned with SparseGPT records a BERTScore of 84.17 for Legal Contracts, compared to 81.52 with Wanda, and 84.73 from the original.

Comparing across model sizes for Llama-2, pruning seems to be less impactful as model size increases. For SparseGPT, we find that the pruned model is comparable (by any metric) in 15 out of 20 comparisons for Llama-2 7B, 18 out of 20 for Llama-2 13B, and in all 20 for Llama-2 70B.

These findings suggest that the summarization performance between pruned and original models is at least comparable.

4.3 Hallucination Risk

Table 5 shows the HRR (Section 3.6) for all models and datasets, using each hallucination risk metric.[6]

Pruning Reduces Hallucination Risk. In almost all cases, irrespective of the pruning method or sparsity pattern (i.e., 2:4 or 50%), the results show that pruned models have a lower hallucination risk (i.e., values lower than 1.0). We find only a single exception, Llama-2 7B pruned with SparseGPT (2:4) for Legal Contracts, with a SummaC$_{\text{ZS}}$ ratio of 1.01. More importantly,

[6] For reproducibility and transparency, we include the full results (i.e., absolute hallucination risk scores) `in this link` due to space constraints.

pruned models record significantly lower HRRs (paired t-test; $p < 0.05$). This applies to 284 out of 300 total comparisons across datasets, models, pruning methods, and sparsity patterns. For example, we observe significantly lower scores across all metrics for Llama-2 7B with SummEval. In particular, SummaC$_{\text{ZS}}$ scores more than halve for 2:4 semi-structured SparseGPT (0.55) and 2:4 semi-structured Wanda (0.49).

These findings seem counter intuitive, considering that pruned models typically perform comparably to original models in summarization (Table 4). As both language modeling and summarization performance remains comparable, we hypothesize that *the parametric knowledge removed by pruning (Namburi et al., 2023) ''forces'' the model to rely more on the source document during generation and in turn reducing hallucination risk.* We examine this further in Section 5.

Semi-structured Pruning Mitigates Hallucination Risk. We observe consistently lower HRRs when pruning with semi-structured sparsity (2:4 pattern), versus unstructured pruning at the same sparsity level (50%). Semi-structured pruning records a lower HRR across all three metrics in 59 out of 65 cases with SparseGPT, and in 55 out of 65 cases with Wanda. We note that semi-structured pruning sometimes produces a substantially lower HRR than unstructured pruning. For example, semi-structured pruning for Llama-2 13B with Wanda records an average SummaC$_{\text{ZS}}$ HRR of 0.61 versus 0.73 with unstructured pruning.

Unstructured pruning allows weights to be removed in any pattern, enabling pruning according to the optimal layer-wise solution. In contrast, semi-structured pruning constrains the solution space to only the subset that satisfies the desired sparsity pattern (e.g., 2:4, removing two weights in every contiguous block of four). Inevitably, even influential weights with relatively high layer-wise saliency scores may be removed. As semi-structured pruning deviates from the optimal layer-wise solution, a higher proportion of important weights are therefore removed. This likely includes relevant parametric knowledge (Namburi et al., 2023), potentially requiring such models to rely more on the source document for generation.

To investigate this, we compute lexical overlap (using ROUGE-1/2/L) between summaries and their source documents across all models, datasets and pruning methods. We find that summaries from models pruned with 2:4 sparsity result in higher lexical overlaps in 114 out of 150 comparisons (three ROUGE metrics, five datasets, five models, two pruning methods) compared to models with 50% unstructured pruning, supporting our hypothesis.

SummaC and HaRiM^{+} Moderately Agree. Considering the average results across datasets, we observe mixed signals from SummaC-based HRRs versus HaRiM^{+} HRRs. For example, SummaC$_{\text{Conv}}$ with SparseGPT (2:4) shows that on average, Llama-2 7B benefits most over the original (0.70), followed by Llama-2 13B (0.74). On the contrary, for HaRiM^{+}with 2:4 sparsity, summaries from Llama-2 13B appear to yield the largest reductions in hallucination risk on average (0.81 with SparseGPT and 0.73 with Wanda), followed by OPT-IML 30B (0.86 with both SparseGPT and Wanda). As the results between hallucination risk metrics differ, we want to shed light on how well they agree with each other. Therefore, we compute Pearson's correlation coefficient between all HRR metrics, across all datasets, models and pruning methods. Unsurprisingly, both SummaC-based metrics show a strong correlation between them (0.82 averaged across all datasets, models and pruning methods). We also find moderate correlations between HaRiM^{+} and SummaC metrics (0.45 between HaRiM^{+} and SummaC$_{\text{ZS}}$; 0.53 between HaRiM^{+}and SummaC$_{\text{Conv}}$).

This is expected, as each metric group computes hallucination risk with different motivations (SummaC-based metrics use entailment methods over the summary and document, while HaRiM^{+} uses token-level predictive likelihood). This explains partly the moderate correlation between them, also *highlighting that it can be beneficial to use HaRiM^{+} and SummaC in conjunction.*

4.4 Human Evaluation

Table 6 shows human evaluation results for the questions presented in Section 3. To offer a fair selection of models, we use summaries generated by the pair that benefited the most (Llama-2 7B) and the least (Mistral 7B) in terms of hallucination risk (i.e., the largest and smallest improvements in Table 5). We then select the corresponding summaries from the pruned counterpart, specifically SparseGPT (2:4)

Model	Halluc. Q1 (↓)	Omiss. Q2 (↓)	Repet. Q3 (↓)	Align. Q4 (↑)
Llama-2 7B	31	**5**	**0**	**28**
w/ SparseGPT	**14**	18	9	21
IAA (κ)	0.82	0.63	0.62	0.53
Mistral 7B	12	**9**	**0**	**31**
w/ SparseGPT	**10**	13	5	23
IAA (κ)	0.87	0.61	0.67	0.59

Table 6: Human evaluation results. Values denote the number (out of 100) of summary preferences by participants for the corresponding category. **Bold** denotes the best performing model per question.

which obtained the most consistent summarization performance (Section 4.2).

Original Models Hallucinate More. Summaries generated by the original Llama-2 7B model contain hallucinations in 31 cases (out of 100) compared to 14 with SparseGPT applied. In comparison, the results for Mistral 7B also suggest that 10 (out of 100) summaries from Mistral 7B pruned with SparseGPT contain hallucinations, compared to 12 summaries generated using the original model (i.e., a smaller difference compared to Llama-2 7B).

This aligns well with our initial expectations and HRR results (Table 5), as Mistral 7B benefits less from pruning in terms of hallucination risk compared to Llama-2 7B. For example, considering SummaC$_{ZS}$ for SummEval, Llama-2 7B pruned with SparseGPT approximately halves the hallucination risk (0.49) compared to 0.79 with Mistral 7B. From analyzing human evaluation results, we found that the large difference between pruned and original Llama-2 7B is predominantly driven by major factual errors (discussed in Section 6).

Original Models Omit and Repeat Slightly Less. With substantial (0.61–0.80) agreement between participants, the results agree that both original models had no repetitions in their summaries and omitted less important information compared to pruned model summaries (e.g., nine instances with Mistral 7B compared to 13 with its pruned version with SparseGPT).

Comparing how well the summaries semantically align with the source document, the results show a preference towards the original models (with moderate agreement; 0.40–0.60). For example, 28 (out of 100) summaries of the original Llama-2 7B were selected as more aligned compared to 21 summaries when pruned with SparseGPT.

5 Impact of Pruning Sparsity on Hallucination Risk

To better understand previous observations and test our hypothesis (i.e., sparsity likely encourages models to focus more on the source document during generation), we analyze hallucination risk across different sparsity levels. We additionally track the lexical overlap (using ROUGE-1/2/L) and semantic overlap (using BERTScore) between the generated summary and the source document. Our hypothesis is: *If lexical overlap positively correlates with sparsity levels, it suggests that pruned models may rely more on the source document for generation.*

Figure 2 shows the summarization performance ratio (ROUGE-1/2/L and BERTScore; ratio computed as pruned over original) and HRR (↓) for five LLMs and two pruning methods, across increasing levels of unstructured sparsity (10% to 50%). We only consider unstructured sparsity, since the 2:4 semi-structured pattern enforces a fixed sparsity level of 50%. The ratio for each metric is averaged across datasets for brevity, with error bars indicating standard deviation. For summarization performance, a ratio higher than 1.0 indicate that the pruned model performs better than the original, whereas a HRR lower than 1.0 indicates that summaries from the pruned model have a lower hallucination risk.

Hallucination Risk Reduces as Sparsity Increases. Results consistently show that hallucination risk reduces as sparsity levels increase, across all models and pruning methods. For example, with Llama-2 13B and Wanda, SummaC$_{ZS}$ HRR reduces from 0.98 at 10% sparsity, to 0.90 at 30% to finally 0.73 at 50%. Moreover, OPT-IML 30B displays a remarkably linear improvement (i.e., with SparseGPT the HRR is 1.00 at 10% sparsity, 0.95 at 30% and 0.90 at 50%, for all hallucination risk metrics). These findings suggest that *increasing sparsity to moderate levels (up to 50%) does indeed appear to reduce hallucination risk in generated summaries*.

Semantic and Lexical Overlaps Differ. Observing the lexical (ROUGE) and semantic

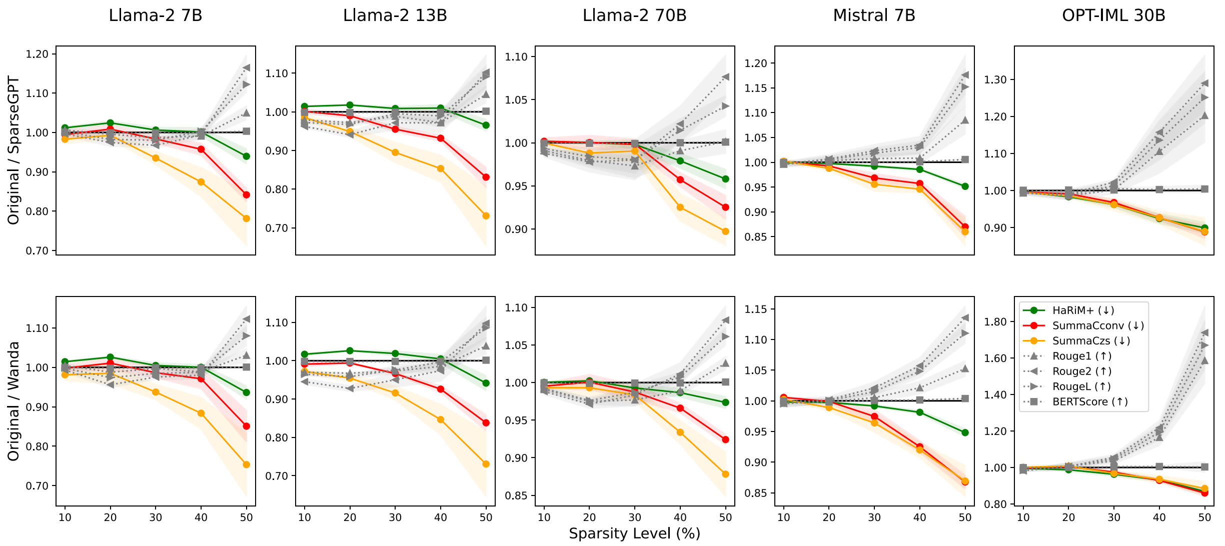


Figure 2: Ratio between a pruned model and the original across five sparsity levels, three hallucination risk metrics (lines with circled markers; lower means pruned is better) and four summary generation performance metrics (gray dotted lines; higher means pruned is better). The ratio for each metric is averaged across all datasets, with error bars indicating standard deviation.

(BERTScore) similarity ratios between document and generated summary across sparsity levels, the outcomes are mixed. In almost all cases for both pruning methods, BERTScore results remain comparable to the original model (close to 1.0) up to 50% sparsity, with minimal deviation across datasets. This shows that summaries from pruned models are as semantically similar to the source document as those from original models, across all sparsity levels.

However, there is a stark contrast with ROUGE-1/2/L. For Llama-2 models, ROUGE-based ratios appear to decrease until 30% sparsity, then increase substantially and peak above 1.0 (the original model baseline) at 50% sparsity. For Mistral 7B and OPT-IML 30B, we observe that ROUGE-based ratios increase above 1.0 (higher than original) from a lower sparsity (20%). As summaries from pruned models remain as semantically similar to the source document as those from original models, their *higher lexical overlap with the source document indicates that pruned models focus more on the input document to generate a summary.*

Higher Lexical Overlap, Lower Hallucination Risk. Surprisingly, we observe an inversely proportional relationship between ROUGE-based ratios and HRRs. We hypothesize that a higher lexical overlap with the source document is a possible reason for the lower hallucination risk. To assess this, we calculate Pearson's correlation coefficient, averaged across sparsity levels between all HRR and ROUGE-based metrics (Table 7, significant correlations in **bold**; $p < 0.05$).

	ROUGE-1/2/L	
Model	SparseGPT	Wanda
Llama-2 7B	−0.69 / **−0.89** / **−0.90**	−0.45 / **−0.86** / −0.79
Llama-2 13B	−0.70 / −0.77 / −0.84	−0.72 / −0.78 / **−0.85**
Llama-2 70B	−0.39 / **−0.86** / **−0.84**	−0.69 / **−0.86** / **−0.86**
Mistral 7B	**−0.91** / **−0.97** / **−0.97**	−0.88 / **−0.96** / **−0.97**
OPT-IML 30B	−0.70 / **−0.93** / **−0.89**	**−0.93** / **−0.94** / **−0.93**

Table 7: Averaged Pearson's correlation coefficient (r) between hallucination risk and ROUGE-based metrics (calculated between the generated summaries and the source documents) across sparsity levels. **Bold** values indicate significant correlations ($p < 0.05$).

We note a strong significant inverse correlation (Pearson's $r < -0.8$) for both pruning methods for ROUGE-2/L across almost all models (excluding Llama-2 13B) and $r < -0.4$ for ROUGE-1. This suggests that *a higher lexical overlap could be responsible for the reduced hallucination risk, while increasing sparsity appears responsible for an increasing lexical overlap*. In particular, we find an almost perfect negative relationship between ROUGE-based ratios and HRRs (−0.97 with SparseGPT) for Mistral 7B. This corroborates findings from the study by Durmus et al. (2020), which shows that summaries with a higher

Source Document	Generated Summary
FactCC #205 What do we have for the contestant on ‘‘The Price Is Right’’? A brand-new car! Whoops. [. . .] model Manuela Arbelaez accidentally revealed the correct answer [. . .]	*Hallucination (Q1)* The contestant on ‘‘The Price Is Right’’ accidentally revealed the correct answer to a guessing game for a new Hyundai Sonata [. . .]
SummEval #86 [. . .] Patrick Kluivert has made a winning start in World Cup qualifying as coach [. . .] The Caribbean island team advanced to the second qualifying round in CONCACAF for the 2018 tournament [. . .]	*Omission (Q2)* Patrick Kluivert [. . .] has successfully led Curacao to the second qualifying round in CONCACAF for the 2018 tournament. [. . .]
SummEval #33 [. . .] auction at Sotheby’s on Wednesday. While the 13 letters were expected to bring in anywhere from $300,000 to $600,000, [. . .] estimated to bring in anywhere from $300,000 to $600,000, but [. . .]	*Repetition (Q3)* [. . .] The 13 letters were expected to bring in anywhere from $300,000 to $600,000, but received no bids. [. . .] The letters were estimated to be worth between $300,000 and $600,000. [. . .]
FactCC #136 Deion Sanders is such a dad. The NFL legend called out Deion Sanders Jr. on Twitter for saying he only eats ‘‘hood doughnuts.’’ In response, the elder Sanders [. . .] reminded his son he has a trust fund, a condo and his own clothing line called ‘‘Well Off’’ [. . .]	*Alignment (Q4)* Deion Sanders Jr. is a Southern Methodist University wide receiver and aspiring entrepreneur who loves ‘‘hood doughnuts’’ and has a trust fund, condo, and clothing line called ‘‘Well Off.’’ [. . .]

Table 8: Examples of errors in generated summaries corresponding to the human evaluation questions and the context from the source document.

lexical similarity to the source document are less likely to contain hallucinations.

6 Qualitative Analysis

Following the human evaluation (see Sections 3.7 and 4.4), we review specific cases, highlighting issues with the summaries generated by pruned models in Table 8.

Hallucinations. Our analysis of the human evaluation task results suggests that hallucinations in the summaries from both Llama-2 7B and Mistral 7B are either: (a) additional information not supported by the source document, or (b) modified or misplaced information from the source document (e.g., FactCC #205).

Omissions. Omission is a category where we found a few instances of disagreement between the participants. In general, participants agree in clear cases like SummEval #86 (e.g., *‘‘2018 tournament’’* should be *‘‘2018 World Cup’’*). Comparatively in disagreements, omitted information is more nuanced and difficult to detect, such as important details from the source document (e.g., missing dates).

Repetitions. Interestingly, we find that summaries containing repetitions occur when the source document also contains repeating information (e.g., the price range ‘‘*$300,000 to $600,000*’’ duplicated in SummEval #33).

Alignment. The generated summaries that are less aligned to the source document do not necessarily contain any hallucinations, omissions, or repetitions. However, we found that they do not entirely convey the original meaning of the source document. For example in FactCC #136, the source describes *Deion Sanders Jr.* being publicly scolded by his father for downplaying his wealthy lifestyle. However, this particular piece of information is not conveyed in the generated summary.

7 Conclusion

We conducted an extensive study to assess the hallucination risk of LLMs after pruning. We experimented with two state-of-the-art pruning methods applied to five instruction-tuned LLMs. We measured the hallucination risk using three established automatic metrics, in addition to a human evaluation. Our results show that as models are pruned to moderately high sparsity levels, the risk of generating hallucinating content decreases. Our analysis suggests that pruned models tend to generate summaries that have a greater lexical overlap with the source document, offering a possible explanation for the lower hallucination risk.

In future work, we plan to explore the relationship between hallucination risk and model quantization (Dettmers et al., 2022; Frantar et al., 2023) and also expand to tasks such as open-book question answering (Ciosici et al., 2021) and machine translation (Guzmán et al., 2019). Finally, an interesting direction is to investigate the relationship between hallucination risk and explanation faithfulness (Chrysostomou and Aletras, 2022; Zhao and Aletras, 2023).

Acknowledgments

We would like to thank the anonymous reviewers and action editor for their invaluable feedback. MW is supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation grant EP/S023062/1. ZZ and NA are supported by EPSRC grant EP/V055712/1, part of the European Commission CHIST-ERA programme, call 2019 XAI: Explainable Machine Learning-based Artificial Intelligence. NA is also supported by EPSRC grant EP/Y009800/1, part of the RAI UK Keystone projects.

References

Griffin Adams, Alex Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From sparse to dense: GPT-4 summarization with chain of density prompting. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 68–74, Singapore. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2023.newsum-1.7`

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The Falcon series of open language models. *arXiv preprint*, arXiv:2311.16867.

Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. 2020. What is the state of neural network pruning? In *Proceedings of Machine Learning and Systems*, volume 2, pages 129–146.

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. `https://doi.org/10.18653/v1/2020.emnlp-main.506`

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact-aware neural abstractive summarization. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press. `https://doi.org/10.1609/aaai.v32i1.11912`

Xiuying Chen, Mingzhe Li, Xin Gao, and Xiangliang Zhang. 2022. Towards improving faithfulness in abstractive summarization. In *Advances in Neural Information Processing Systems*, volume 35, pages 24516–24528. Curran Associates, Inc.

Prafulla Kumar Choubey, Alex Fabbri, Jesse Vig, Chien-Sheng Wu, Wenhao Liu, and Nazneen Rajani. 2023. CaPE: Contrastive parameter ensembling for reducing hallucination in abstractive summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10755–10773, Toronto, Canada. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2023.findings-acl.685`

George Chrysostomou and Nikolaos Aletras. 2022. An empirical study on explanations in out-of-domain settings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6920–6938, Dublin, Ireland. Association for Computational Linguistics. `https://doi.org/10.18653/v1/2022.acl-long.477`

Manuel Ciosici, Joe Cecil, Dong-Ho Lee, Alex Hedges, Marjorie Freedman, and Ralph Weischedel. 2021. Perhaps PTLMs should go to school – a task to assess open book and closed book QA. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6104–6111, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.493

Trevor Cohn and Mirella Lapata. 2008. Sentence compression beyond word deletion. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 137–144, Manchester, UK. Coling 2008 Organizing Committee. https://doi.org/10.3115/1599081.1599099

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*.

Daniel Deutsch, Tania Bedrax-Weiss, and Dan Roth. 2021. Towards question-answering as an automatic metric for evaluating the content quality of a summary. *Transactions of the Association for Computational Linguistics*, 9:774–789. https://doi.org/10.1162/tacl_a_00397

Chen Dun, Mirian Hipolito Garcia, Guoqing Zheng, Ahmed Hassan Awadallah, Anastasios Kyrillidis, and Robert Sim. 2023. Sweeping heterogeneity with smart mops: Mixture of prompts for LLM task adaptation. *arXiv preprint*, arXiv:2310.02842.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.454

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031. https://doi.org/10.1162/tacl_a_00410

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409. https://doi.org/10.1162/tacl_a_00373

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1213

Constanza Fierro and Anders Søgaard. 2022. Factual consistency of multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3046–3052, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.240

Elias Frantar and Dan Alistarh. 2022. Optimal brain compression: A framework for accurate post-training quantization and pruning. In *Advances in Neural Information Processing Systems*, volume 35, pages 4475–4488. Curran Associates, Inc.

Elias Frantar and Dan Alistarh. 2023. SparseGPT: Massive language models can be accurately pruned in one-shot. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10323–10337. PMLR.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*.

Prakhar Ganesh, Yao Chen, Xin Lou, Mohammad Ali Khan, Yin Yang, Hassan Sajjad, Preslav Nakov, Deming Chen, and Marianne Winslett. 2021. Compressing large-scale transformer-based models: A case study on BERT. *Transactions of the Association for Computational Linguistics*, 9:1061–1080. https://doi.org/10.1162/tacl_a_00413

Nuno M. Guerreiro, Pierre Colombo, Pablo Piantanida, and André Martins. 2023. Optimal transport for unsupervised hallucination detection in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13766–13784, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.770

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. 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 6098–6111, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1632

Masafumi Hagiwara. 1994. A simple and effective method for removal of hidden units and weights. *Neurocomputing*, 6(2):207–218. Backpropagation, Part IV. https://doi.org/10.1016/0925-2312(94)90055-8

Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Adib Hasan, Ileana Rugina, and Alex Wang. 2024. Pruning for protection: Increasing jailbreak resistance in aligned LLMs without fine-tuning. *arXiv preprint*, arXiv:2401.10862.

B. Hassibi, D. G. Stork, and G. J. Wolff. 1993. Optimal brain surgeon and general network pruning. In *IEEE International Conference on Neural Networks*, pages 293–299 vol. 1. https://doi.org/10.1109/ICNN.1993.298572

Dandan Huang, Leyang Cui, Sen Yang, Guangsheng Bao, Kun Wang, Jun Xie, and Yue Zhang. 2020. What have we achieved on text summarization? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 446–469, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.33

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2023. OPT-IML: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint*, arXiv:2212.12017.

Ajay Kumar Jaiswal, Zhe Gan, Xianzhi Du, Bowen Zhang, Zhangyang Wang, and Yinfei Yang. 2024. Compressing LLMs: The truth is rarely pure and never simple. In *The Twelfth International Conference on Learning Representations*.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.123

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint*, arXiv:2310.06825.

Daniel King, Zejiang Shen, Nishant Subramani, Daniel S. Weld, Iz Beltagy, and Doug Downey. 2022. Don't say what you don't know: Improving the consistency of abstractive summarization by constraining beam search. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 555–571, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.gem-1.51

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.750

Philippe Laban, Wojciech Kryscinski, Divyansh Agarwal, Alexander Fabbri, Caiming Xiong, Shafiq Joty, and Chien-Sheng Wu. 2023. SummEdits: Measuring LLM ability at factual reasoning through the lens of summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9662–9676, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.600

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. https://doi.org/10.1162/tacl_a_00453

Mateusz Lango and Ondrej Dusek. 2023. Critic-driven decoding for mitigating hallucinations in data-to-text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2853–2862, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.172

Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan TN. 2023. Building real-world meeting summarization systems using large language models: A practical perspective. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 343–352, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-industry.33

Yann LeCun, John Denker, and Sara Solla. 1989. Optimal brain damage. In *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann.

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.

Hui Lin and Vincent Ng. 2019. Abstractive summarization: A survey of the state of the art. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9815–9822. https://doi.org/10.1609/aaai.v33i01.33019815

Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. LLM-Pruner: On the structural pruning of large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Laura Manor and Junyi Jessy Li. 2019. Plain English summarization of contracts. In *Proceedings of the Natural Legal Language Processing Workshop 2019*, pages 1–11, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-2201

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.173

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

Kirill Milintsevich and Navneet Agarwal. 2023. Calvados at MEDIQA-chat 2023: Improving clinical note generation with multi-task instruction finetuning. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 529–535, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.clinicalnlp-1.56

Asit Mishra, Jorge Albericio Latorre, Jeff Pool, Darko Stosic, Dusan Stosic, Ganesh Venkatesh, Chong Yu, and Paulius Micikevicius. 2021. Accelerating sparse deep neural networks. *arXiv preprint*, arXiv:2104.08378.

Markus Nagel, Rana Ali Amjad, Mart Van Baalen, Christos Louizos, and Tijmen Blankevoort. 2020. Up or down? Adaptive rounding for post-training quantization. In *Proceedings of* *the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7197–7206. PMLR.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics. https://doi.org/10.18653/v1/K16-1028

Satya Sai Srinath Namburi, Makesh Sreedhar, Srinath Srinivasan, and Frederic Sala. 2023. The cost of compression: Investigating the impact of compression on parametric knowledge in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5255–5273, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.349

Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Conditional generation with a question-answering blueprint. *Transactions of the Association for Computational Linguistics*, 11:974–996. https://doi.org/10.1162/tacl_a_00583

Neural Magic. 2021. DeepSparse.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya

Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, C. J. Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 technical report. *arXiv preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1250

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.92

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1437

Horacio Saggion and Thierry Poibeau. 2013. *Automatic Text Summarization: Past, Present and Future*, pages 3–21. Springer Berlin Heidelberg, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-28569-1_1

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green AI. *Communications of the ACM*, 63(12):54–63. https://doi.org/10.1145/3381831

Seonil (Simon) Son, Junsoo Park, Jeong-in Hwang, Junghwa Lee, Hyungjong Noh, and Yeonsoo Lee. 2022. HaRiM$^+$: Evaluating summary quality with hallucination risk. 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 895–924, Online only. Association for Computational Linguistics.

Mingjie Sun, Zhuang Liu, Anna Bair, and J. Zico Kolter. 2024. A simple and effective pruning approach for large language models. In *The*

Twelfth International Conference on Learning Representations.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023a. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.650

Yuting Tang, Ratish Puduppully, Zhengyuan Liu, and Nancy Chen. 2023b. In-context learning of large language models for controlled dialogue summarization: A holistic benchmark and empirical analysis. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 56–67, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.newsum-1.6

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*, arXiv:2307.09288.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint*, arXiv:1506.05869.

Byron C. Wallace, Sayantan Saha, Frank Soboczenski, and Iain J. Marshall. 2021. Generating (factual?) narrative summaries of RCTs: Experiments with neural multi-document summarization. *AMIA Summits on Translational Science Proceedings*, 2021:605.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.450

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 214–229, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3531146.3533088

Miles Williams and Nikolaos Aletras. 2023. How does calibration data affect the post-training pruning and quantization of large language models? *arXiv preprint*, arXiv:2311.09755.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yijun Xiao and William Yang Wang. 2021. On hallucination and predictive uncertainty in conditional language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2734–2744, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.236

Guangxuan Xu and Qingyuan Hu. 2022. Can model compression improve NLP fairness. *arXiv preprint*, arXiv:2201.08542.

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. 2023. Understanding and detecting hallucinations in neural machine translation via model introspection. *Transactions of the Association for Computational Linguistics*, 11:546–564. https://doi.org/10.1162/tacl_a_00563

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. 2024. Benchmarking Large Language Models for News Summarization. *Transactions of the Association for Computational Linguistics*, 12:39–57. https://doi.org/10.1162/tacl_a_00632

Zhixue Zhao and Nikolaos Aletras. 2023. Incorporating attribution importance for improving faithfulness metrics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4732–4745, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.261

Zhixue Zhao, George Chrysostomou, Kalina Bontcheva, and Nikolaos Aletras. 2022a. On the impact of temporal concept drift on model explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4039–4054, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-emnlp.298

Zheng Zhao, Shay B. Cohen, and Bonnie Webber. 2020. Reducing quantity hallucinations in abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2237–2249, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.203

Zhixue Zhao and Boxuan Shan. 2024. Reagent: A model-agnostic feature attribution method for generative language models. *arXiv preprint*, arXiv:2402.00794. https://doi.org/10.22541/au.170709121.16176681/v1

Zhixue Zhao, Ziqi Zhang, and Frank Hopfgartner. 2022b. Utilizing subjectivity level to mitigate identity term bias in toxic comments classification. *Online Social Networks and Media*, 29:100205. https://doi.org/10.1016/j.osnem.2022.100205

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.120