Controlling Attention Dynamics for Factual Grounding in Neural Text Generation

Roseline Mary Rozario[1], Philip O.Ogunbona[1], Khin Than Win[1], Jie Yang[1]
[1]School of Computing and Information Technology, University of Wollongong, New South Wales, Australia
Email:rmr451@uowmail.edu.au, philipo@uow.edu.au, win@uow.edu.au, jiey@uow.edu.au

***Abstract*—Attention mechanisms govern internal information routing in neural encoder–decoder architectures, yet their dynamics are typically optimized only implicitly under likelihood-based objectives. In this work, we argue that unconstrained attention dynamics contribute to factual errors in neural text generation and that attention should be treated as a controllable architectural mechanism rather than as an emergent alignment artifact. We propose *Attention Bias Augmentation* (ABA), a lightweight intervention that injects a bounded, head-specific bias into decoder cross-attention logits to explicitly steer information flow toward factually salient source content. The bias is derived from a rule-based token-level salience signal, enabling domain-informed guidance without introducing additional supervision. ABA operates entirely within the attention computation and preserves architectural simplicity by avoiding auxiliary objectives and prediction heads while introducing only a minimal set of per-head scaling parameters. We evaluate ABA on echocardiography report summarization using the MIMIC-III dataset in a controlled setting and observe consistent improvements in entailment-based factuality metrics, including SummaC, while maintaining fluency and content coverage. These results demonstrate that the controlled modulation of attention dynamics provides an effective and scalable mechanism for improving factual grounding in neural text generation.**

***Index Terms*—attention control, encoder–decoder models, neural factual grounding, clinical text generation.**

I. Introduction

Attention mechanisms play a central role in neural encoder–decoder architectures [1] by governing how information is routed between source and target representations during generation. Despite this centrality, attention dynamics are typically optimized only implicitly under likelihood-based training objectives and are treated as an emergent alignment phenomenon rather than as explicitly controllable architectural components. As a result, the internal routing decisions that determine which source evidence influences generation remain weakly constrained, even in high-stakes applications where factual accuracy is critical.

In neural text generation, and particularly in structured domains such as clinical summarization, this lack of explicit control over attention dynamics can lead to systematic factual errors. Generated content may appear fluent and coherent while relying on source representations that are linguistically compatible but evidentially insufficient to support the stated facts. Such failures manifest as hallucinated findings, unsupported inferences, or distortions of clinical severity. These errors are not merely stylistic artifacts but reflect breakdowns in internal evidence selection, with potential downstream safety implications. Most existing approaches to improving factual consistency attribute these failures to limitations in training data, supervision, or model scale [2], [3] and address them through auxiliary objectives, reinforcement learning, external verification modules, or post-hoc correction pipelines. While effective in certain regimes, these methods treat factuality as an external constraint imposed on generation, often at the cost of increased architectural complexity, additional supervision signals, or inference-time overhead.

In this work, we adopt a complementary perspective grounded in neural network design. We argue that a significant source of factual inconsistency arises from unconstrained decoder cross-attention dynamics during generation. Rather than modifying training objectives or appending external verification stages, we treat attention as a controllable internal routing mechanism and intervene directly at the point where source representations compete for influence. By explicitly shaping attention dynamics, we aim to improve factual grounding by stabilizing internal evidence selection without altering the underlying model architecture or decoding procedure.

We introduce *Attention Bias Augmentation* (ABA), a lightweight architectural intervention that injects a bounded, head-specific bias into decoder cross-attention logits based on token-level salience. ABA operates entirely within the attention computation, requires no auxiliary losses or prediction heads, and preserves the original training objective and inference pipeline. By modifying attention logits prior to normalization, ABA provides explicit architectural control over internal information routing while leaving all other components of the network unchanged.

Our contributions are threefold:

- **Architectural reframing of attention:** We present a perspective in which factual inconsistency in neural text generation is attributed, in part, to misaligned decoder cross-attention dynamics, and argue that attention should be treated as a controllable architectural mechanism rather than an emergent byproduct of likelihood optimization.
- **Attention Bias Augmentation (ABA):** We propose a parameter-efficient attention-logit intervention that enables targeted control of internal information routing without modifying model depth, width, training objectives, or decoding strategies.
- **Empirical validation in neural text generation:** We demonstrate the effectiveness of ABA on echocardiogra-

phy report summarization from the MIMIC-III dataset, showing consistent improvements in entailment-based factuality metrics while preserving fluency and content coverage.

II. Related Work

A. Factual Consistency in Abstractive Summarization

Prior work on improving factual consistency in abstractive summarization has predominantly focused on training-level or post-generation interventions rather than modifications to the attention mechanism itself. A common strategy introduces auxiliary objectives that explicitly penalize hallucination or reward faithful generation. For example, contrastive and reinforcement-based approaches encourage models to distinguish factually consistent summaries from corrupted alternatives [2], [4]. Similarly, counterfactual data augmentation exposes models to synthetic factual errors during training to improve robustness [5]. While effective, these methods rely on additional supervision signals and increase training complexity.

Another line of research incorporates external evidence or verification modules to assess factual correctness. Entailment-based evaluators and question-answering frameworks are often used either as training rewards or as post-hoc filters [6], [7]. Although such approaches improve factual reliability, they introduce inference-time overhead and treat factuality as an external constraint rather than an internal property of the generation process.

B. Attention Behaviour and Salience Modeling

Early attempts to improve content selection through attention include coverage mechanisms that track previously attended regions to reduce repetition and omission [8]. Subsequent work introduced salience-aware summarization models that estimate token importance using auxiliary predictors and indirectly guide attention during decoding [7]. However, such methods rely on additional prediction heads or supervision signals and influence attention only indirectly.

Several studies have analyzed attention behavior in neural encoder–decoder architectures. Ferrando and Costa-juss [9] show that attention weights in Transformer models are often weakly aligned with informative source tokens. Voita et al. [10] further find that attention heads exhibit specialized roles, with only a small subset responsible for alignment and information flow. Structured attention networks [11] support this view by explicitly constraining attention distributions within neural architectures. Collectively, these findings suggest that attention exhibits structured behavior and explicitly guides, rather than arising purely as an emergent artifact of likelihood optimization. However, direct intervention at the attention-logit level remains largely unexplored in clinical summarization settings.

C. Positioning of ABA

Existing approaches to improve factual consistency in abstractive summarization typically operate outside the core generation mechanism, either by modifying training objectives, introducing auxiliary rewards, or appending external verification or correction modules. While effective, such methods treat factuality as an external constraint imposed on generation and often increase training complexity or inference-time cost. In contrast, Attention Bias Augmentation(ABA) intervenes directly within the decoder cross-attention mechanism, at the point where source evidence is selected during token generation.

The distinction is central to our approach. By reshaping attention logits rather than optimizing downstream factuality signals, ABA addresses factual inconsistency as an internal alignment problem rather than a data or supervision deficiency. Unlike instruction-tuned or large-scale language models, which often improve factuality through conservative generation or paraphrastic abstraction [12], [13]. ABA preserves the base model's capacity and decoding behavior while enforcing evidence-sensitive alignment. As such, ABA represents a complementary and orthogonal strategy, providing architectural control over internal information flow independent of model scale, external knowledge source, or task-specific rewards.

III. Problem Formulation

We therefore formulate factual consistency as a failure of evidence selection during decoding: at generation time, the decoder attends to source representations that are linguistically compatible but not evidentially sufficient to support the generated content.

Let $X = (x_1, \ldots, x_n)$ denote a source clinical document, and let $Y = (y_1, \ldots, y_m)$ denote the target summary. We consider a standard encoder–decoder Transformer parameterized by θ, which models the conditional distribution.

$$p_\theta(Y \mid X) = \prod_{t=1}^{m} p_\theta\left(y_t \mid y_{<t}, X\right). \tag{1}$$

The context vector used to generate y_t is computed as

$$c_t = \sum_{i=1}^{n} \alpha_{t,i} V_i, \tag{2}$$

where V_i denotes the encoder value representation corresponding to the i-th source token.

A. Evidence Selection Failure

We define *factual inconsistency* as a failure of evidence selection during encoding. Specifically, although the attention distribution $\boldsymbol{\alpha}_t$ may assign high probability mass to source tokens that are linguistically compatible with the predicted token y_t, these tokens may be evidentially insufficient to support the factual content being generated.

Formally let,

$$E_t \subseteq \{1, \ldots, n\} \tag{3}$$

denote the set of source token indices that constitute sufficient evidence for generating y_t. A factual error occurs when:

$$\sum_{i \in E_t} \alpha_{t,i} \;\ll\; \sum_{i \notin E_t} \alpha_{t,i}. \tag{4}$$

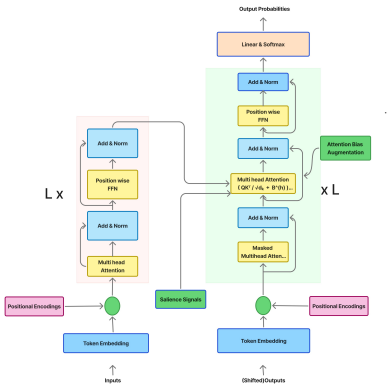

Fig. 1. Attention as a controllable mechanism for factual grounding. ABA injects a head-specific bias $B^{(h)}$ into decoder cross-attention logits before softmax, *modulating* attention allocation toward salient source tokens while preserving the standard BART architecture.

In this case, the decoder conditions its prediction on spurious or weakly related source content, leading to hallucinated, distorted, or unsupported facts in the generated summary.

B. Objective

The learning problem is therefore to estimate model parameters θ such that, for each decoding step t, the attention distribution $\boldsymbol{\alpha}_t$ is biased toward evidentially sufficient source tokens:

$$\boldsymbol{\alpha}_t \;\rightarrow\; \boldsymbol{\alpha}_t^*, \tag{5}$$

where

$$\sum_{i \in E_t} \alpha_{t,i}^* \text{ is maximized.} \tag{6}$$

This formulation reframes factual consistency as an attention-level optimization problem, rather than solely a decoding or post-hoc verification task as illustrated in Fig. 1.

IV. Method

A. Baseline Cross-Attention

Let $X = x_1, \ldots, x_n$ denote the source document and $Y = y_1, \ldots, y_m$ the generated summary. In a Transformer encoder–decoder architecture, the encoder maps the source sequence to a set of hidden representations $\mathbf{H} = \{\mathbf{h}_i\}_{i=1}^n$. During decoding, cross–attention enables the decoder to condition each generated token on relevant source representations.

At decoding timestep t, the decoder forms a query vector $\mathbf{q}_t$ from its current hidden state. Standard cross–attention computes attention scores between $\mathbf{q}_t$ and encoder keys $\mathbf{K} = \{\mathbf{k}_i\}$, where $\mathbf{k}_i = W_K \mathbf{h}_i$, using scaled dot-product attention. The resulting attention distribution is given by

$$\mathbf{A}_t = \text{softmax}\left(\frac{\mathbf{q}_t \mathbf{K}^\top}{\sqrt{d_k}}\right), \tag{7}$$

where d_k denotes the key dimensionality. The context vector used to predict the next token is then computed as a weighted sum of encoder values $\mathbf{V} = \{\mathbf{v}_i\}$, with $\mathbf{v}_i = W_V \mathbf{h}_i$.

While this formulation is effective for general sequence modeling, it treats attention as an emergent alignment mechanism optimized primarily for likelihood maximization. As a result, attention distributions may remain diffuse or weakly aligned with factually salient source evidence, particularly in long and structured clinical documents.

B. Attention Bias Augmentation

To address this limitation, we introduce Attention Bias Augmentation (ABA), which explicitly shapes decoder cross–attention by injecting a task–informed bias into the attention logits prior to normalization. The core idea is to influence evidence selection at the point where source representations compete for attention, rather than supervising outputs or introducing auxiliary objectives. For each attention head h, ABA defines a bounded bias vector $B(\mathbf{h}) \in \mathbb{R}^n$ over source positions. This bias is derived from a token-level salience signal $\hat{\mathbf{s}}$ that encodes factual or clinical importance and is computed independently of the attention mechanism.

C. Salience Signal Construction

The salience signal $\hat{\mathbf{s}}$ is computed using a deterministic, rule-based procedure, where each source token is assigned a score based on configurable domain heuristics. The computation integrates multiple features designed to reflect clinically relevant signal importance, including: (i) lexical and structural indicators (e.g., numbers with units, name:value patterns, acronyms), (ii) section-based priors emphasizing informative sections such as *Findings* or *Conclusion*, (iii) negation and uncertainty handling that down-weights tokens within uncertain or speculative contexts, and (iv) temporal anchors such as dates and time expressions.

The resulting $\hat{\mathbf{s}}$ serves as a domain-informed salience prior for guiding attention allocation, computed independently of model parameters and without introducing additional learnable components, thereby preserving the lightweight design of ABA. The bias for head h is defined as:

$$B^{(h)} = \text{clip}\left(\gamma \cdot \tanh(\alpha_h) \cdot \hat{s}, -\beta, \beta\right), \tag{8}$$

where α_h is a learnable head–specific scaling parameter, γ is a global scaling constant, and β bounds the magnitude of the bias to ensure numerical stability. The number of additional parameters introduced by α_h is negligible relative to the base model size. The $tanh$ nonlinearity further prevents unbounded amplification of salience signals. This formulation allows ABA to inject inductive bias into attention computation while maintaining the architectural simplicity, avoiding auxiliary objectives and prediction heads, and introducing only a minimal set of per-head scaling parameters.

D. Biased Cross-Attention Computation

With ABA, the cross-attention computation for head h at timestep t becomes:

$$A_t^{(h)} = \text{softmax}\left(\frac{Q_t K^\top}{\sqrt{d_k}} + B^{(h)}\right). \tag{9}$$

The bias term shifts attention mass toward factually salient source tokens prior to softmax normalization, thereby directly influencing which encoder representations contribute to the context vector used for token generation.

Importantly, ABA operates entirely within the existing attention computation. It does not introduce additional losses, supervision signals, or inference-time constraints, and preserves the original training objective and decoding procedure. By acting directly on attention logits, ABA provides explicit architectural control over evidence selection during generation.

V. Results

A. Results and Discussion

Tables I and II report performance on the ECHO dataset using surface-level metrics (ROUGE), semantic similarity (BERTScore), and source-based factuality metrics (SummaC and QuestEval). To isolate the effect of Attention Bias Augmentation (ABA), we primarily compare BioBART-base and BioBART-base+ABA, which share identical architectures, training data, and decoding settings, differing only in the attention-logit biasing mechanism.

B. Surface-level and Semantic Similarity Evaluation

We evaluate surface-level overlap using ROUGE [14] and semantic similarity using BERTScore [15]. On the ECHO test split, BioBART-base+ABA achieves the strongest overall performance across surface-level and semantic similarity metrics. As shown in Table II, ABA yields modest but consistent improvements in ROUGE, with gains in ROUGE-1 (+0.33), ROUGE-2 (+0.58), and ROUGE-L (+0.37) over BioBART-base, indicating improved lexical overlap with reference summaries. Beyond ROUGE, BERTScore further corroborates these improvements at the semantic level. BioBART-base+ABA achieves the highest BERTScore-F1 (91.24), reflecting improved semantic alignment between generated summaries and references. While gains in BERTScore Precision and Recall are small, the consistent improvement in F1 suggests that ABA enhances semantic faithfulness without sacrificing coverage or introducing spurious content.

These results indicate that attention biasing primarily improves content selection quality, leading to summaries that are not only lexically similar but also semantically closer to reference summaries, even when absolute surface-level gains remain modest.

C. Factual Consistency Evaluation

We evaluate source-based factual consistency using an entailment-based metric, SummaC [16], and a question-answering-based metric, QuestEval [17]. ABA consistently improves factual consistency relative to the base model, increasing from 72.39 to 72.74(+0.35). Paired significance testing on per-instance scores over the ECHO test set ($N = 1{,}624$) as in Table III shows that SummaC improvement is statistically significant (mean $\Delta = +0.0125$, 95% CI $[0.0054, 0.0197]$, $p < 0.001$ under both bootstrap and permutation tests). In contrast, the QuestEval improvement is not statistically significant ($\Delta = +0.0015$, $p \approx 0.33$), indicating that ABA primarily improves entailment-based faithfulness rather than factual coverage breadth. This asymmetry reflects differences in metric focus: SummaC captures entailment consistency, while QuestEval evaluates coverage via question-answering. Since ABA modulates token-level cross-attention, it improves alignment with source evidence, reducing contradictions and unsupported inferences without explicitly encouraging broader content coverage. Thus ABA strengthens faithfulness, while completeness may require complementary mechanisms.

Interpreting metric gains: Although absolute improvements are modest, they are systematic and expected given the saturation behavior of source-based metrics once outputs remain broadly entailed. As ABA does not alter the training objective, decoding strategy, or model capacity, its effect is isolated to attention modulation. Consequently, aggregate gains underestimate its impact. Record-level analysis shows that ABA corrects high-impact factual errors, including unsupported inferences, omission of clinically salient findings, and severity distortion. These errors arise from diffuse cross-attention over irrelevant content. By biasing attention toward salient tokens, ABA stabilizes evidence selection during generation, reducing clinically meaningful failures while preserving fluency and coverage. This supports attention as a controllable mechanism for factual grounding in neural text generation.

D. Comparison with instruction–tuned models

Instruction–tuned and text–to–text models FLAN-T5-large and T5-base achieve higher absolute SummaC and QuestEval scores despite weaker or less stable ROUGE performance. This behavior is consistent with their tendency to generate conservative, paraphrastic summaries that remain broadly entailed by the source, which is favored by source-based factuality metrics but penalized by reference-overlap measures such as ROUGE. In clinical summarization, where reference impressions may omit or compress salient details, this divergence between source-based and reference-based evaluation is expected.

E. Implications of attention-level intervention

ABA achieves statistically reliable improvements in entailment-based factual consistency without increasing model capacity or introducing auxiliary objectives. This supports the interpretation that factual errors in clinical summarization are not solely attributable to limited model scale or supervision, but also to misaligned decoder cross-attention during generation. By directly biasing attention logits toward clinically salient source tokens, ABA provides a lightweight architectural mechanism for improving factual grounding while preserving standard training and inference pipelines. Although the overall metric gains are small, consistent statistical improvements and corrections of clinically important errors indicate that misaligned attention contributes to factual inconsistency.

TABLE I
ROUGE SCORES ON THE ECHO EVALUATION SPLIT

Model	Split	R-1	R-2	R-L
Evaluation Split				
Flan-T5-large	Eval	53.78	41.55	49.85
T5-base	Eval	54.58	41.55	49.85
BART-large	Eval	55.36	40.39	54.59
BioBART-base	Eval	55.97	40.95	55.26
BioBART-base + ABA	Eval	55.94	40.83	55.14

TABLE II
PERFORMANCE ON THE ECHO TEST SPLIT: ROUGE, BERTSCORE, AND SOURCE-BASED FACTUALITY METRICS

Model	R-1	R-2	R-L	B-P	B-R	B-F1	SummaC	QuestEval
Flan-T5-large	53.85	40.16	40.88	91.99	89.99	90.93	75.70	77.25
T5-base	54.14	40.66	49.18	92.04	90.11	91.02	75.27	77.41
BART-large	54.72	39.32	53.91	91.46	90.75	91.07	72.62	75.73
BioBART-base	55.22	39.84	54.31	91.67	90.80	91.19	72.39	75.80
BioBART-base + ABA	**55.55**	**40.42**	**54.68**	**91.83**	**90.74**	**91.24**	**72.74**	**76.28**

TABLE III
PAIRED SIGNIFICANCE TESTING ON THE ECHO TEST SET ($N = 1{,}624$) COMPARING BIOBART-BASE AND ABA. Δ DENOTES THE MEAN PER-INSTANCE DIFFERENCE (ABA $-$ BIOBART-BASE).

Metric	Δ	95% CI	p (boot)	p (perm)
SummaC	$+0.0125$	$[0.0054,\ 0.0197]$	< 0.001	< 0.001
QuestEval	$+0.0015$	$[-0.0016,\ 0.0046]$	0.33	0.34

F. Qualitative Error Analysis

Sample 1: Severity Distortion (Record ID: 1321)

Source (excerpt). *Hyperdynamic left ventricle with small cavity size... **mild diastolic dysfunction**... **moderate pulmonary hypertension**.*

BioBART-base (error). *"Hyperdynamic left ventricle with small cavity... signijicant diastolic dysfunction and moderate pulmonary hypertension."*

Error Type. Severity distortion: exaggerates the degree of diastolic dysfunction beyond what is supported by the source.

ABA (improved). *"Hyperdynamic left ventricle with small cavity size... diastolic dysfunction and moderate pulmonary artery systolic hypertension."*

Metric evidence. SummaC: $0.36 \rightarrow 0.97$ ($\Delta = +0.61$).

Sample 2: Salient Content Omission (Record ID: 1464)
Source (excerpt): *Findings: left ventricle with normal wall thickness and a small, hyperdynamic left ventricle.*
BioBART-base (error): *The left ventricle is normal.*
Error Type: *Omission of clinically salient functional descriptor.*
ABA (improved): *Small, hyperdynamic left ventricle.*
Metric evidence: SummaC: $0.29 \rightarrow 0.96$ ($\Delta = +0.67$).

Sample 3: Unsupported Inference / Hallucination (Record ID: 1438)
Source (excerpt): *Findings: new vegetation on the posterior leaflet of the mitral valve.*
BioBART-base (error): *Abnormal findings are noted on echocardiography.*
Error Type: *Entity omission and loss of specificity.*
ABA (improved): *New vegetation on the posterior leaflet of the mitral valve.*
Metric evidence: SummaC: $0.21 \rightarrow 0.87$ ($\Delta = +0.66$).

VI. CONCLUSION

This work demonstrates that factual grounding in neural text generation can be improved by explicitly controlling attention dynamics within encoder–decoder architectures, rather than treating attention as a passive byproduct of likelihood-based training. We introduced Attention Bias Augmentation (ABA), a lightweight architectural intervention that injects a bounded, head-specific bias into decoder cross-attention logits to guide internal information routing toward factually salient source content.

Crucially, ABA operates entirely within the attention computation and preserves the original model architecture, training objective, and decoding procedure. As a result, differences in factual behavior can be causally attributed to the controlled modulation of attention dynamics rather than to changes in model capacity, supervision, or inference strategy. Experiments on echocardiography report summarization using the MIMIC-III dataset show consistent improvements in

entailment-based factuality metrics while maintaining fluency and content coverage.

While our evaluation is conducted on a structured clinical domain, it may limit direct generalization to more heterogeneous text settings, ABA itself is not tied to any specific dataset. The proposed framework operates at the level of attention computation and can accommodate alternative salience signals, including learned or task-adaptive formulations. We therefore view this work as a controlled demonstration of attention-level intervention for factual grounding, with broader applicability to diverse generation tasks.

These findings support the view that misaligned attention dynamics contribute directly to factual errors in neural text generation and that attention can serve as an explicit architectural control mechanism for stabilizing evidence selection during generation. By exposing attention modulation as a controllable degree of freedom, this work provides a principled and scalable approach to improving factual grounding in high-stakes text generation tasks. Finally, this study focuses on isolating the effect of ABA within a single backbone (BioBART-base) to enable controlled comparison. While this design allows clear attribution of gains to attention biasing, evaluating ABA across additional model architectures and conducting more extensive ablations over its components remain important directions for future work.

ACKNOWLEDGMENT

We gratefully acknowledge the support of University of Wollongong, Australia through the International Postgraduate Tuition Award(IPTA), as well as its academic support and research environment, which enabled this research. Portions of this manuscript were edited for language clarity using ChatGPT, a large language model developed by OpenAI.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[2] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada, 2018.

[3] W. Kryscinski, B. McCann, C. Xiong, and R. Socher, "Evaluating the factual consistency of abstractive text summarization," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 9332–9346.

[4] I.-c. Chern, Z. Wang, S. Das, B. Sharma, P. Liu, and G. Neubig, "Improving factuality of abstractive summarization via contrastive reward learning," in *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP)*, 2023, pp. 55–60.

[5] Z. Ling, Y. Xie, C. Dong, and Y. Shen, "Enhancing factual consistency in text summarization via counterfactual debiasing," in *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, 2025, pp. 7912–7924.

[6] A. Pagnoni, V. Balachandran, and Y. Tsvetkov, "Understanding factuality in abstractive summarization with FRANK: A benchmark for factuality metrics," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2021, pp. 4812–4829.

[7] Z. Cao, F. Wei, W. Li, and S. Li, "Faithful to the original: Fact-aware neural abstractive summarization," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

[8] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2017, pp. 1073–1083.

[9] J. Ferrando and M. R. Costa-juss, "Attention weights in transformer NMT fail aligning words between sequences but largely explain model predictions," in *Findings of the Association for Computational Linguistics: EMNLP*, 2021, pp. 434–443.

[10] E. Voita, D. Talbot, F. Moiseev, R. Sennrich, and I. Titov, "Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019, pp. 5797–5808.

[11] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, "Structured attention networks," in *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

[12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[13] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.

[14] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proceedings of the ACL Workshop on Text Summarization Branches Out*, 2004, pp. 74–81.

[15] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with BERT," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[16] P. Laban, J. Hsu, and M. Hearst, "Summac: Re-visiting nli-based models for inconsistency detection in summarization," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2021, pp. 163–177.

[17] T. Scialom, P.-A. Dray, S. Lamprier, B. Piwowarski, and J. Staiano, "Questeval: Summarization asks for fact-based evaluation," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 6598–6609.