MONSOON - 2024
M24.CS7.501 - ADVANCED NLP

Combining Semantic Uncertainty & Hallucination Detection in Natural Language Generation

Research Report

By Team: 35
Arush Sachdeva - 2023121008
Rayaan Khan - 2021101120
Yash Shivhare - 2021101105

Presentation Link | Repository Link

Content

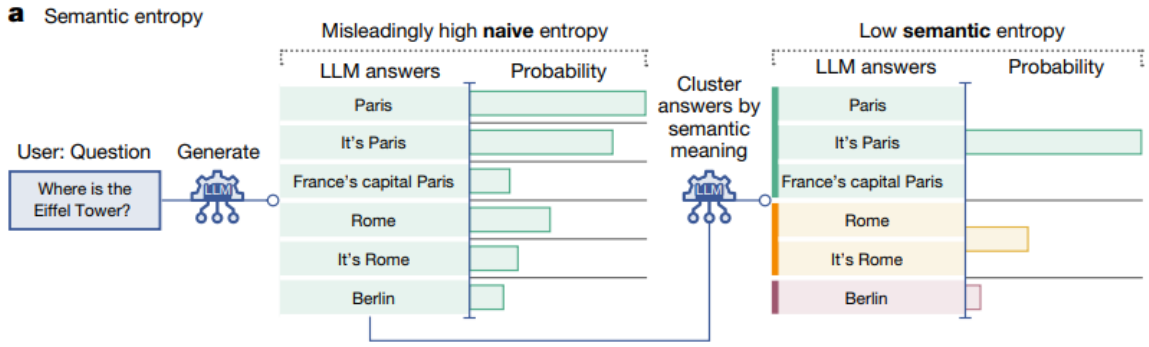
Abstract

Natural language generation (NLG) systems, including large language models, often encounter challenges such as semantic ambiguity and hallucination. This research investigates the connection between semantic uncertainty and hallucinations in NLG, demonstrating that(Our Initial Hypothesis) ***higher semantic uncertainty corresponds to higher hallucination rates***. By employing semantic entropy as a novel metric for uncertainty estimation, we refine existing methods and establish meaningful insights into model behavior. Using question-answering datasets (SQuAD, CoQA) and robust models (RoBERTa, Muwa-1.3B), we integrate state-of-the-art tools like bi-directional entailment clustering and SelfcheckBERT to assess correctness. Our findings emphasize semantic entropy's utility in predicting hallucinations and evaluating model reliability, particularly in **larger models.**

1. Introduction

Natural language generation (NLG) is increasingly pivotal across applications like question answering (QA), summarization, and conversational AI. However, challenges like uncertainty in outputs and hallucinations (confidently incorrect outputs) undermine trustworthiness. Traditional uncertainty metrics, such as lexical similarity and probability-based measures, fail to capture semantic nuances critical for real-world deployment.

If I ask you – What is your name? You might answer - X, It's X or My name is X. Semantically all of these three things mean the same. Thus, it is expected from anyone who understands the meaning to tell each of the statement confidence.



Excitingly, the model does not provide same confidence value for all the three values as expected. This is the genesis of the problem.

It is then hypothesised that if the model is unable to understand the semantic meaning of sentences, it is highly likely that it hallucinates.

This study introduces ***semantic entropy***, a measure of uncertainty that incorporates semantic equivalence, and explores its relationship with hallucinations. Our hypothesis, ***"Higher semantic uncertainty implies higher hallucination rates***,"(**NOTE:** *Our findings demonstrate a correlation, not a causal relationship. While it's plausible that high uncertainty contributes*

to hallucinations we don't claim that it necessarily cause them. Other factors may contribute to hallucinations.) is evaluated across diverse datasets and model architectures.

2. Related Work

2.1 Research

Recent advancements in natural language processing (NLP) have extensively explored uncertainty estimation and hallucination detection in large language models (LLMs). While individual approaches provide critical insights, our study leverages these works collectively to uncover deeper relationships between semantic uncertainty and hallucination rates. Below, we summarize key contributions from relevant works and outline how they inform our integrated methodology.

---

1. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation

This paper introduces **semantic entropy** as a novel measure to capture uncertainty in natural language generation (NLG). Traditional uncertainty metrics often rely on token-level probabilities and fail to account for semantically equivalent outputs. The authors propose clustering semantically equivalent sentences using **bi-directional entailment** and calculating entropy over the meaning clusters.

The method employs **Monte Carlo sampling** to approximate the entropy of the meaning space, enabling the detection of hallucinations and improving model reliability. Importantly, this approach is unsupervised and task-agnostic, making it applicable across diverse NLG tasks. Their findings show that semantic entropy outperforms traditional methods like lexical overlap metrics, particularly in free-form QA tasks, where models often confabulate.

Our Perspective: This paper forms the backbone of our study by providing a robust framework to measure uncertainty. We build on their methodology to explore how semantic entropy correlates with hallucinations in QA systems.

---

2. Detecting Hallucinations in Large Language Models Using Semantic Entropy

This paper extends the use of semantic entropy to directly detect hallucinations in LLM outputs. The focus is on understanding how **semantic variability** impacts the reliability of free-form text generation. By shifting the focus to meaning rather than token-level probabilities, the proposed method detects hallucinations without requiring domain-specific training data.

Semantic entropy emerges as a powerful unsupervised tool, offering insights into confabulations—outputs that appear fluent but lack factual grounding. This methodology ensures greater trustworthiness in model outputs across diverse applications.

Our Perspective: The work validates the effectiveness of semantic entropy for hallucination detection. We incorporate their insights to investigate how hallucination patterns correlate with high semantic uncertainty in free-form QA tasks.

---

3. Looking for a Needle in a Haystack: A Comprehensive Study of Hallucinations in Neural Machine Translation

This paper focuses on **sequence log probability** as a metric for detecting hallucinations in neural machine translation (NMT). The authors demonstrate that low sequence log probabilities strongly correlate with hallucinated outputs—translations that are incorrect or irrelevant to the source text. Unlike reference-based methods (e.g., COMET-QE), which depend on high-quality ground truth data, sequence log probability offers a robust, reference-free approach for real-time hallucination detection.

The study further highlights the computational efficiency of sequence log probability compared to other heuristics like attention alignment scores.

Our Perspective: We extend the application of sequence log probability to QA systems, comparing its effectiveness with semantic entropy for hallucination detection. This enables a comprehensive understanding of hallucination patterns across tasks.

---

4. CMU OAQA at TREC 2017 LiveQA: A Neural Dual Entailment Approach for Question Paraphrase Identification

This paper introduces a **bi-directional entailment algorithm** for detecting semantic equivalence between questions. Using a neural dual entailment approach, the model encodes premise and hypothesis questions and compares their semantic similarity in both directions. This ensures robust paraphrase detection by aligning phrase vectors from both questions.

The dual entailment approach improves paraphrase identification, particularly for ambiguous or complex question pairs, and serves as the foundation for clustering semantically equivalent outputs.

Our Perspective: We adopt this algorithm to cluster model-generated answers in QA tasks. By identifying semantic equivalence through bi-directional entailment, we derive clusters that form the basis for calculating semantic entropy.

2.2 Integrating The Above Research

While each of these papers provides standalone contributions to uncertainty estimation, hallucination detection, and semantic clustering, our work seeks to **combine these methodologies into a unified framework**. Specifically:

1. **Semantic Entropy for Uncertainty Estimation**:
    - Building on Kuhn et al.'s semantic entropy, we quantify uncertainty based on meaning clusters rather than token probabilities.

    - We extend its application to study how semantic uncertainty correlates with hallucinations.
2. **Sequence Log Probability for Confidence Measurement**:
    - We incorporate log probabilities to evaluate the model's confidence and correlate it with semantic entropy.
    - This allows us to explore how overconfidence aligns with hallucinations.
3. **Bi-Directional Entailment for Semantic Clustering**:
    - Adopting CMU's dual entailment approach, we ensure that clustering captures semantic equivalence across generated outputs.
4. **Hallucination Detection Across Metrics**:
    - By integrating semantic entropy, sequence log probability, and correctness scores (via SelfCheckBERT), we provide a multi-faceted analysis of hallucinations.

2.3 Our Contribution

The related works collectively provide the theoretical and methodological groundwork for our study. However, while previous studies often focus on individual metrics or tasks, our research uniquely investigates **how semantic uncertainty relates to hallucination rates**. This holistic approach bridges gaps between uncertainty quantification and hallucination detection, paving the way for more reliable NLP systems.

3. Methodology

3.1. Datasets

- **SQuAD**: A large-scale QA dataset with high-quality, factual queries.
- **CoQA**: A conversational QA dataset with multi-turn question-answer pairs.

3.2. Models

- **RoBERTa (base, fine-tuned on QA)** and **Muwa-1.3B**: Chosen for their performance on generative QA tasks.(Although we ran multiple models during our interim report,we analysed that these 2 models turned out to be the most appropriate for our analysis due to difference in their sizes(so that we can also relate the changes in pattern with the model sizes) also had to do multiple hyper parameterizations to find the most appropriate parameters).
- **DeBERTa (NLI)**: Used for bi-directional entailment in clustering.
- **SelfcheckBERT**: Evaluates correctness of generated answers based on semantic equivalence.

3.3. Semantic Uncertainty Estimation

3.3.1. Answer Generation

Models generate 20 answers per question using multinomial sampling. Sampling temperature is optimized (e.g., CoQA: 0.5).

3.3.2. Clustering Using Bi-directional Entailment

Sentences (s) and (s') are semantically equivalent if (E(s, s')) and (E(s', s)), that is we take both as premise-hypothesis and vice-versa and see how much they entail to each other.

Steps:

- Concatenate question-context pairs.
- Use DeBERTa (NLI) to classify entailment between (s) and (s').
- Group answers into semantic clusters.

3.3.3. Semantic Entropy Calculation

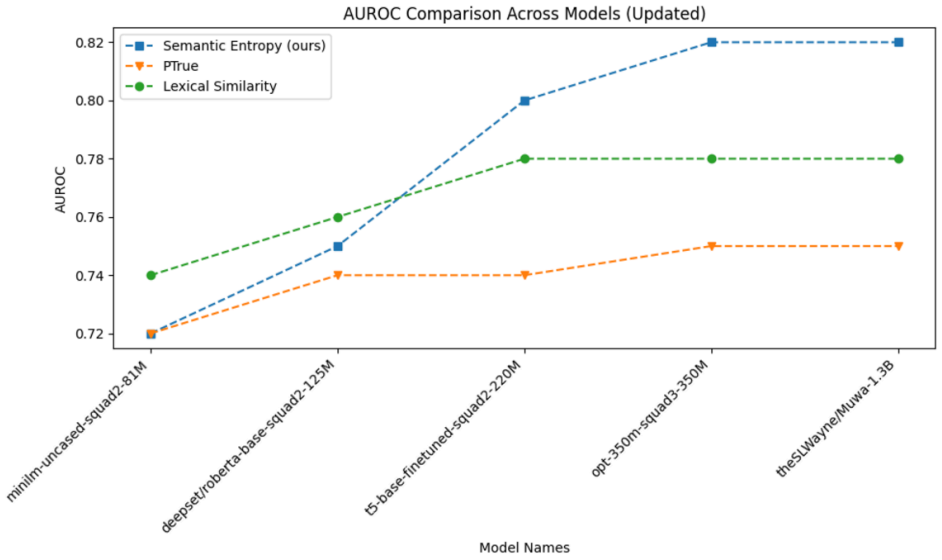
- Likelihood of each semantic class (c):

$$SE(x) = -\sum_{c} p(c \mid x) \log p(c \mid x) = -\sum_{c} \left(\left(\sum_{\mathbf{s} \in c} p(\mathbf{s} \mid x) \right) \log \left[\sum_{\mathbf{s} \in c} p(\mathbf{s} \mid x) \right] \right)$$

- Monte Carlo integration approximates this entropy.

$$SE(x) \approx -|C|^{-1} \sum_{i=1}^{|C|} \log p(C_i \mid x).$$

3.3.4. AUROC Analysis



(Already Discussed in the Interim Report)

4. Hallucination Detection

4.1 Metrics

Sequence Log Probability*:* Measures confidence of start and end logits for generated outputs.

Sequence log-probability (Seq-Logprob). For a trained model $P(y|x, \theta)$ and a generated translation y, Seq-Logprob (i.e., model confidence) is the *length-normalised* sequence log-probability:

$$\frac{1}{L}\sum_{k=1}^{L}\log P(y_k \mid y_{<k}, x, \theta). \qquad (1)$$

Definition: Sequence log probability measures the likelihood of a sequence of tokens (words) being generated by a language model given a specific input context. It is calculated using the probabilities assigned to each token in the sequence, based on the model's internal state.

White-Box Methods: These methods require access to the model's internal states, specifically the logits (raw output scores before applying softmax) for each token. By using these logits, one can compute the probabilities for each token in the output sequence conditional on the input.

4.2 Application in Hallucination Detection

Identifying Hallucinations

- **Low Log Probability**: If a generated output has a low log probability, it suggests that the model finds that output unlikely given the input context. This can indicate potential hallucination, especially if the output is factually incorrect.
- **Max Log Probability**: The maximum log probability among tokens can also be used to assess hallucination risk. If the least likely token (as determined by max log probability) is significantly low, it raises flags about the reliability of that output.

4.3 Hallucination Analysis

Hallucinations in language models can be defined as high-confidence, incorrect answers that the model generates and presents as factual. This phenomenon poses significant challenges in ensuring the reliability of AI-generated content.

4.4 Research Backing

SAC3: **Reliable Hallucination Detection in Black-Box Language Models via Semantic-aware Cross-check Consistency:**
This paper discusses how language models can produce high-confidence incorrect answers, which are indicative of hallucinations. It highlights that these models can consistently generate incorrect responses to specific questions while maintaining high confidence, thus reinforcing the idea that such outputs represent a clear case of hallucination.

On Large Language Models' Hallucination with Regard to Known Facts:
The authors investigate the dynamics of output token probabilities and demonstrate that when models generate incorrect outputs despite having access to correct knowledge, it often reflects high confidence in those hallucinated responses. Their findings suggest that understanding these dynamics is crucial for detecting when models are hallucinating.

Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models:
This research introduces methods for detecting hallucinations based on the internal states of LLMs, emphasizing that high-confidence outputs often correlate with factual inaccuracies. The study underlines the need for real-time detection mechanisms to address this issue effectively

4.5 Reasoning

High Confidence Misleading Users: High-confidence outputs can mislead users into believing that the information is accurate, which is a critical risk associated with hallucinations.

Detection Challenges: As highlighted in these studies, traditional methods may struggle to identify these high-confidence incorrect answers, necessitating advanced detection frameworks that account for both confidence levels and factual accuracy.
In summary, defining hallucinations as high-confidence incorrect answers is well-supported by recent research, which underscores the importance of developing robust detection mechanisms to mitigate the risks associated with such outputs in language models.

4.6 Correlation Between Hallucination Rates and Semantic Entropy

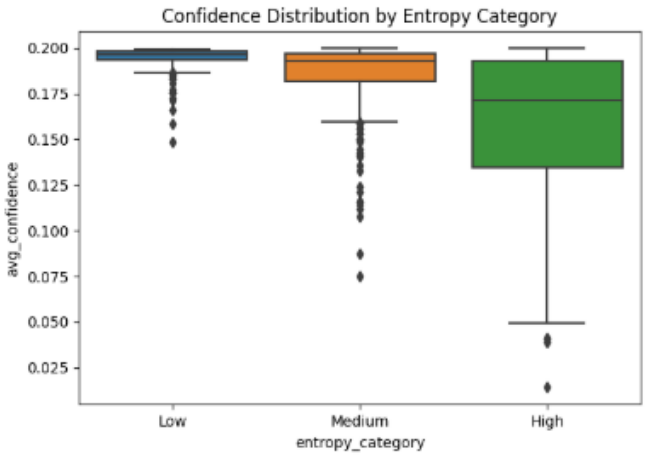
In our analysis, we investigated the relationship between **hallucination rates** and **semantic entropy** across different models and questions to understand how the uncertainty represented by entropy correlates with the likelihood of hallucinations. The following patterns and observations were noted:

1. **High Semantic Entropy Correlates with Hallucination**:
    - Models or questions exhibiting **higher semantic entropy** often showed **higher hallucination rates**. This indicates that when semantic entropy is high, the model is generating highly confident outputs that deviate from grounded knowledge, leading to hallucinations.

    - This aligns with the interpretation of semantic entropy as a measure of uncertainty in the model's understanding of the input.
2. **Random Patterns for Other Entropy Combinations**:
    - For other cases with lower or mid-range semantic entropy, the relationship with hallucination rates appeared more **random**. This suggests that low entropy is not always indicative of correct answers, and hallucinations can still occur even when the model exhibits lower uncertainty.
    - These patterns underline that while semantic entropy provides valuable signals for hallucination prediction, it may not be sufficient in isolation for complete assessment.
3. **Insights from the Observed Trends**:
    - The relationship between semantic entropy and hallucination rates emphasizes the importance of **calibrating model uncertainty**. A model with well-calibrated entropy values could potentially provide early warnings for hallucinations.
    - However, the lack of strong patterns in certain ranges indicates the need to combine semantic entropy with other metrics (e.g., predictive entropy, lexical similarity) for robust detection.
4. **Conclusion**:
    - While high semantic entropy generally signals hallucination, the **absence of systematic patterns in other cases** highlights the complexity of the relationship between entropy measures and hallucination rates. This reinforces the idea that hallucinations are multi-faceted and influenced by a combination of factors beyond just entropy.

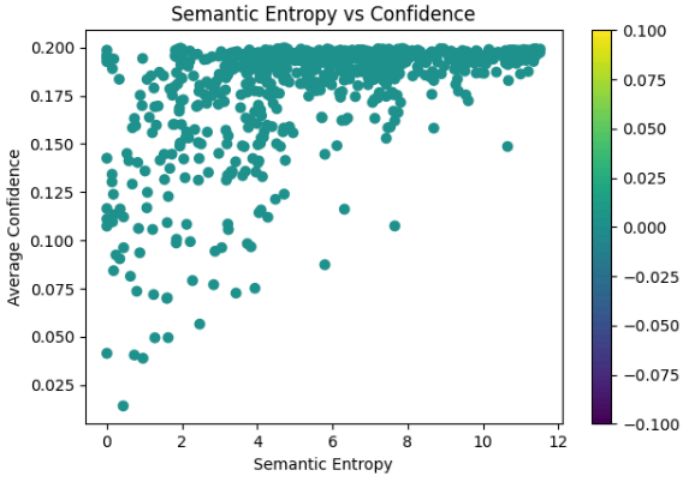
5. Analysis

Hypothesis 1: *High Entropy → Incorrect Answer + High Confidence*
Supported by the scatter plot (points with high entropy and high confidence, mostly incorrect).
Hypothesis 2: *High Entropy → Randomness in Correct Answer + Confidence*
Partially supported by the boxplot (high entropy spread) and AUROC for semantic entropy.
Hypothesis 3: *Low Entropy → Randomness in Correct/Incorrect Answer + Confidence*
Supported by both the scatter plot (low entropy points) and boxplot (low category shows variability)



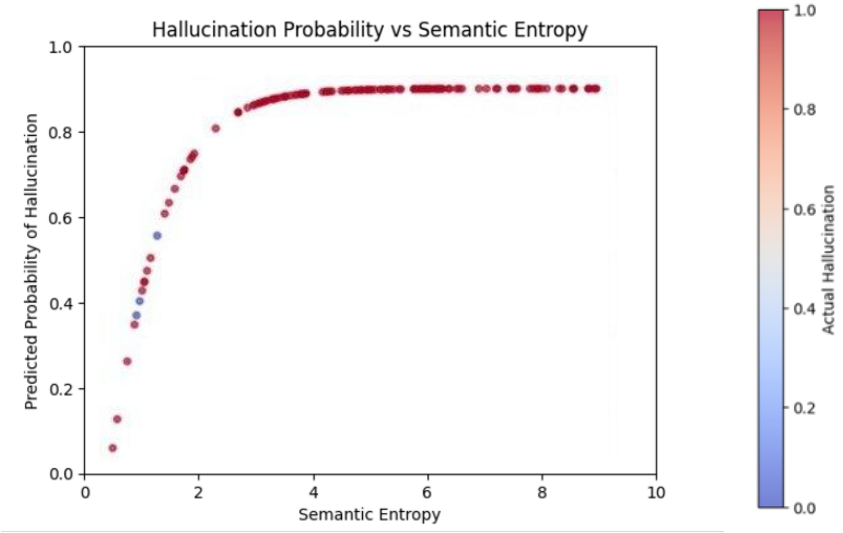
Inference: Confidence is distributed across three entropy categories: Low, Medium, and High.
Insights:

- **High Entropy:** The median confidence in the "High" category is generally higher, and the spread indicates both high and low confidence cases (some outliers). This aligns with Hypothesis 2: *High entropy can be random but can also exhibit high confidence even for correct answers.*
- **Low Entropy:** Confidence is more consistent in the "Low" category but still allows for variability in predictions. This aligns with Hypothesis 3.



Inference: This scatter plot shows how the average confidence of model predictions varies with semantic entropy. The color bar represents the proportion of correct answers.
Insights:

- **High Entropy:** Observations on the right (higher entropy) with high confidence (y-axis near 0.2) and incorrect answers (purple color) support your first hypothesis: *High entropy implies incorrect answers with high confidence.*
- **Low Entropy:** Points with low entropy (left side) exhibit varied confidence levels and correct/incorrect answers (color variability), supporting the third hypothesis: *Low entropy scenarios can be random.*



Inferences:

- **Low Entropy:** Points with entropy closer to 0 show a significant spread in predicted hallucination probabilities and actual hallucination values, implying some randomness in the system's outputs.
- **High Entropy:** Points with high entropy consistently have high predicted hallucination probabilities, which aligns with the expectation that the model is likely incorrect in such cases.

Relevance to Hypotheses:

1. **High entropy + incorrect answers:** The clustering of red points on the right indicates that high entropy is often associated with incorrect or hallucinated predictions, supporting this hypothesis.
2. **High entropy + correct answers (random behavior):** A lack of blue points in high-entropy regions suggests this behavior may not be frequent or significant.
3. **Low entropy + random behavior:** The clustering of points at the top-left corner (low entropy, high hallucination probability) contradicts the randomness hypothesis.

Model Size Impact

Larger models (e.g., Muwa-1.3B) exhibit higher semantic entropy, reinforcing its scalability (Semantic entropy scales better with model size and makes better use of increasing numbers of samples than baselines.)

6. Discussion

6.1. Insights into Semantic Entropy

- Semantic entropy better captures uncertainty by factoring in meaning rather than token-level likelihoods.

- Replacing ROUGE-L with SelfcheckBERT enhances reliability.

6.2. Hallucination Implications

Semantic uncertainty provides a predictive signal for hallucinations but requires additional context to assess completeness.

6.3. Practical Challenges

1. Computational Overhead: Bi-directional entailment clustering is computationally expensive.
2. Generalization: Performance on other NLG tasks (e.g., summarization) requires validation.

7. Conclusion

This study establishes a direct relationship between semantic uncertainty and hallucinations in NLG. Semantic entropy emerges as a robust metric for uncertainty estimation, outperforming traditional baselines. Moreover, the use of semantic uncertainty for hallucination detection is promising but requires further exploration across diverse tasks and model families.

8. Future Work

- Extend analysis to other NLG tasks (e.g., summarization, machine translation).
- Investigate approximate clustering methods to reduce computational cost.
- Explore hybrid metrics combining semantic entropy with hallucination-specific measures like sequence log probability.

9. References

1. Kuhn, L., Gal, Y., & Farquhar, S. (2023). *Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation.* ICLR.
2. Guerreiro, N. M., Voita, E., & Martins, A. F. T. (2023). *Looking for a Needle in a Haystack: A Comprehensive Study of Hallucinations in Neural Machine Translation.*
3. *Detecting Hallucinations in Large Language Models Using Semantic Entropy.* Nature.
4. Rao, J., He, H., Lin, J., & Sun, M. (2017). *CMU OAQA at TREC 2017 LiveQA: A Neural Dual Entailment Approach for Question Paraphrase Identification.*
5. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* NAACL-HLT.

6. Nogueira, R., & Cho, K. (2019). *Passage Re-ranking with BERT.* arXiv preprint arXiv:1901.04085.
7. Raffel, C., Shazeer, N., Roberts, A., et al. (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer (T5).* JMLR.
8. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). *Attention is All You Need.* NeurIPS.