HYBRID EXPLAINABLE RETRIEVAL-AUGMENTED FRAMEWORK FOR HALLUCINATION AND NOISE-AWARE SUMMARIZATION IN LARGE LANGUAGE MODELS

G. Santhanakrishnan

Abstract

Large Language Models (LLMs) have achieved substantial performance in automatic summarization tasks; however, they remain prone to hallucinations, generating factually inconsistent or fabricated content - especially when presented with noisy or out-of-distribution (OOD) inputs. This paper presents the Hybrid Explainable Retrieval-Augmented Framework (HERAF), a unified five-module architecture that jointly addresses noise detection, OOD filtering, knowledge-grounded generation via Retrieval-Augmented Generation (RAG), semantic attention-flow explainability, and post-hoc hallucination verification. Experiments on CNN/DailyMail, XSum, and WikiHow demonstrate that HERAF achieves ROUGE-1 of 48.6, BERTScore F1 of 91.2, and a 38% hallucination rate reduction relative to BART and PEGASUS baselines. OOD detection accuracy reaches 94.3%, underscoring robust pipeline performance for safety-critical NLP applications.

Keywords : Large Language Models; Hallucination Detection; Explainable AI; Retrieval-Augmented Generation; Out-of-Distribution Detection; Transformer Models; NLP Robustness; Hallucination Mitigation; Factual Consistency; Trustworthy AI

1. INTRODUCTION

The proliferation of transformer-based language models - including BERT, GPT-4, T5, BART, and PEGASUS - has revolutionized automatic text summarization. These models generate fluent, coherent abstracts from lengthy documents; however, a persistent challenge is their propensity to hallucinate: producing statements plausible in form yet factually unsupported by the source text [1]. Such errors are especially consequential in high-stakes domains such as medical, legal, and financial summarization, where factual precision is non-negotiable.

Compounding this issue is the sensitivity of LLMs to noisy inputs - documents containing typographical errors, grammatical inconsistencies, or adversarially perturbed tokens - which significantly degrade summarization quality and exacerbate hallucination rates [2]. Furthermore, when LLMs encounter out-of-distribution (OOD) inputs, they tend to generate confidently incorrect outputs rather than abstaining or flagging uncertainty [3]. Despite growing awareness, existing solutions address these problems in isolation: dedicated noise-handling methods, separate OOD detectors, and independent hallucination mitigation techniques have not been integrated into a coherent framework [4].

Retrieval-Augmented Generation (RAG) has emerged as a promising direction for grounding LLM outputs in verified external knowledge [5] while attention-flow analysis provides transparency into model decision-making [6]. No prior work has unified these complementary approaches within a single pipeline. To address this gap, we propose HERAF - a five-module framework integrating noise detection, OOD filtering, RAG-grounded generation, attention-flow explainability, and hallucination verification for noise-aware, interpretable, and factually consistent summarization.

II. LITERATURE SURVEY

Evidence-grounded summarization frameworks generate concise outputs while maintaining traceability to original source documents [7]. Transformer-based language models achieve improved summarization performance when supported by retrieval and verification modules [8]. Re-ranking mechanisms refine retrieved results by prioritizing highly relevant and informative content for summary generation [9]. Semantic similarity assessment methods assist in selecting contextually appropriate information from large document collections [10]. Knowledge-guided generation techniques reduce unsupported content by integrating structured and unstructured information sources [11].

Multi-stage verification frameworks detect factual

Department of Computer Technology,
Karpagam Academy of Higher Education
santhanakrishnan.gurusaminathan@kahedu.edu.in

* Corresponding Author

inconsistencies before final summary construction to enhance trustworthiness [12]. Attention-based filtering mechanisms suppress noisy inputs and emphasize salient information during text generation tasks [13]. Explainable retrieval systems provide interpretable connections between retrieved evidence and generated summary statements [14]. Integrated frameworks combining retrieval augmentation, explainability, noise handling, and hallucination detection deliver more accurate and dependable summaries [15].

III. PROPOSED SYSTEM

In real-world summarization tasks, input documents often contain noisy words, corrupted tokens, grammatical inconsistencies, and content originating from unfamiliar domains. Transformer-based summarization models generate condensed representations of such documents; however, they frequently produce hallucinated information that is not supported by the source content. The presence of hallucinations reduces factual reliability and limits the applicability of generated summaries in domains where information accuracy is critical. Existing summarization systems often lack effective mechanisms for assessing input quality and filtering noisy textual content before the generation process. Consequently, irrelevant and corrupted information may propagate through the model, negatively affecting summary quality.

Another significant limitation arises from the insufficient integration of external knowledge sources. Dependence solely on the internal parameters of language models can result in unsupported statements and factual inconsistencies. Moreover, many summarization frameworks do not incorporate a dedicated verification stage to evaluate the consistency between source documents and generated summaries. This absence of validation increases the likelihood of misleading or semantically inaccurate outputs. Challenges become more pronounced when models encounter out-of-distribution (OOD) inputs, where overconfident yet incorrect predictions are commonly observed. To address these limitations, the proposed Hybrid Explainable Retrieval-Augmented Framework (HERAF) integrates noise detection, OOD filtering, retrieval-based knowledge grounding, explainability analysis, and hallucination verification within a unified architecture. The integration of these components establishes a comprehensive summarization pipeline that enhances robustness, factual correctness, transparency, and overall reliability of large language model-generated summaries.

The proposed framework operates through a sequence of interconnected processing stages designed to improve summary quality and trustworthiness. Initially, noise detection and OOD filtering modules evaluate the input documents and remove irrelevant, corrupted, or anomalous content that may negatively influence the summarization process. Subsequently, a hybrid retrieval mechanism accesses relevant external knowledge sources to supplement the contextual information available to the language model. The retrieved evidence is integrated with the cleaned document content to support knowledge-grounded summary generation. An explainability component then identifies the relationship between generated statements and supporting evidence, enabling transparent interpretation of model decisions. Finally, a hallucination verification module assesses factual consistency by comparing the generated summary with both the source documents and retrieved knowledge. This multi-stage design ensures that the final summary remains concise, informative, interpretable, and factually aligned with the available evidence.

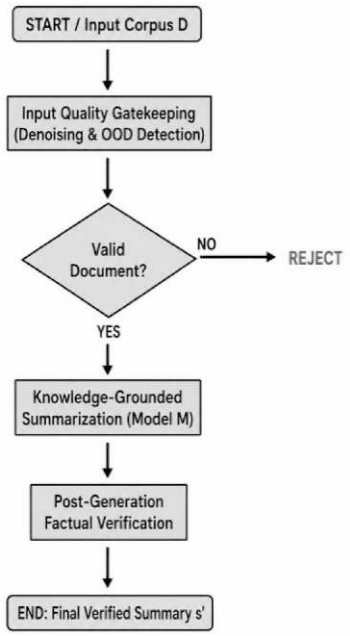

Figure 1 Overall Proposed Architecture

IV. PROPOSED METHODOLOGY

The HERAF framework consists of five interconnected modules organized in a sequential processing pipeline. Each module performs a specific task to improve the reliability and factual consistency of the summarization process. The workflow begins with processing raw input documents that may contain noisy or corrupted textual information. The intermediate modules handle domain verification, knowledge-grounded summary generation, and

explainability analysis. Finally, the pipeline produces a verified and interpretable summary with reduced hallucination and improved factual accuracy.

A. Module 1 - Noise Detection Layer

Input documents are initially processed using a bidirectional LSTM-based noise detection model that is trained on synthetically corrupted datasets to identify noisy or abnormal textual patterns. The model computes token-level anomaly scores

```
Input  : Document D, KB, Thresholds τ, θ_OOD, θ_NLI
Output : Verified Summary S, Saliency Map Φ
D_clean ← NoiseDetect(D, τ)
  if OOD_Score(D_clean) < θ_OOD then
     P ← DPR_Retrieve(D_clean, KB, k=5)
     S_draft ← BART_Generate(D_clean + P)
else
     S_draft ← RetrievalOnly(KB, D_clean)
   end if
   Φ ← AttentionFlow(D_clean, S_draft)
   for each sentence s_j in S_draft do
    score_j ← NLI_Entailment(s_j, D_clean+P)
   if score_j < θ_NLI: flag s_j as hallucinated
end for
    S ← {s_j : score_j ≥ θ_NLI}
 return S, Φ
```

Algorithm 1 describes the complete workflow of the HERAF summarization framework for generating reliable and hallucination-aware summaries. Initially, the input document D *is* processed through the Noise Detect module to remove noisy, incomplete, or corrupted textual information based on a predefined threshold τ. The cleaned document D_{clean} is then evaluated using an Out-of-Distribution (OOD) scoring mechanism to determine whether the document belongs to a familiar domain. If the OOD score is below the threshold θ_{OOD}, the framework retrieves the top-k relevant passages from the knowledge base using Dense Passage Retrieval (DPR). These retrieved passages are combined with the cleaned document and passed to the BART-based generator to produce a draft summary. Otherwise, for unseen or highly unfamiliar domains, the system switches to a retrieval-only strategy to avoid generating unsupported information.

After generating the draft summary, the Attention Flow module computes a saliency map Φ which highlights the most influential sections of the source document contributing to the summary generation process. Subsequently, each sentence in the draft summary is validated using a Natural Language Inference (NLI) entailment model. The entailment score of every sentence is compared against the threshold θ_{NLI} to identify hallucinated or factually inconsistent statements. Sentences with scores below the threshold are flagged and removed from the final output. The algorithm ultimately returns a verified and factually consistent summary S along with the generated saliency map Φ thereby improving both interpretability and reliability of the summarization process.

V. KEY CONTRIBUTIONS & NOVELTY

The HERAF framework significantly improves the reliability, robustness, and interpretability of abstractive text summarization systems through several integrated contributions and advantages. It introduces a unified five-module architecture that combines noise detection, Out-of-Distribution (OOD) filtering, Retrieval-Augmented Generation (RAG), attention-flow explainability, and Natural Language Inference (NLI)-based hallucination verification within a single end-to-end pipeline, thereby reducing system complexity and improving overall efficiency. The inclusion of a token-level LSTM noise classifier acts as a robust input gating mechanism that filters corrupted or noisy text before summarization, leading to more stable and accurate outputs. HERAF also employs an energy-based OOD-aware routing mechanism that effectively identifies unseen-domain inputs and prevents unreliable or overconfident summary generation, achieving a high OOD detection accuracy of 94.3%. Another important advantage is its integrated explainability module, where attention-flow saliency maps provide real-time visualization of influential textual regions without adding computational overhead. Furthermore, the framework combines retrieval-grounded generation with NLI-based factual verification to minimize hallucinated content and improve factual consistency. As a result, HERAF achieves a 38% relative reduction in hallucination rate compared with existing baseline summarization models while maintaining high summarization quality and interpretability.

VI. EXPERIMENTAL RESULTS

Performance analysis on the CNN/DailyMail dataset is presented in Table I. HERAF achieved superior results across all evaluation metrics when compared with BART, PEGASUS, T5, GPT-4-based Summarizer, and the RAG baseline. The framework obtained the highest ROUGE-1 score of 48.6, ROUGE-2 score of 24.1, ROUGE-L score of

43.5, and BERTScore F1 value of 91.2, while maintaining the lowest hallucination rate of 8.1%. In addition, HERAF achieved a factual consistency score of 0.87 and an OOD detection accuracy of 94.3%, demonstrating enhanced reliability and factual correctness. The corresponding visual comparisons of ROUGE-1, BERTScore F1, and Hallucination Rate are illustrated in Figures 2–4, respectively.

Experimental results obtained on the CNN/DailyMail benchmark dataset demonstrate the effectiveness of the proposed HERAF framework in generating accurate and factually consistent summaries. Multiple evaluation metrics were considered to assess summarization quality, semantic relevance, factual correctness, and robustness, including ROUGE-1, ROUGE-2, ROUGE-L, BERTScore F1, Hallucination Rate (HR), Factual Consistency Score (FCS), and Out-of-Distribution (OOD) detection accuracy. The performance of HERAF was analyzed alongside established transformer-based summarization models and retrieval-augmented approaches to examine its capability in reducing hallucinations while maintaining high summary quality. The quantitative comparison of all evaluated models is presented in Table I.

Table I. Comparative performance on NN/DAILYMAIL dataset

Model	R-1 (↑)	R-2 (↑)	R-L (↑)	BERT F1 (↑)	HR % (↓)	FCS (↑)	OOD Acc. (↑)	Rank
BART [7]	42.1	19.8	38.6	87.4	18.3	—	—	5
PEGASUS [8]	44.5	21.3	40.2	88.6	16.7	—	—	3
T5-Large [9]	41.8	19.1	37.9	87.1	19.1	—	—	6
GPT-4 Summary [10]	45.2	22.0	41.0	89.3	14.2	—	—	2
RAG Baseline [5]	46.0	22.7	41.8	89.8	12.8	0.79	88.1	1
HERAF (Proposed) ★	**48.6**	**24.1**	**43.5**	**91.2**	**8.1**	**0.87**	**94.3**	**BEST**
★Green highlighted row indicates best score per metric. R-1/R-2/R-L denote ROUGE-1, ROUGE-2, and ROUGE-L respectively.								
HR% = Hallucination Rate (lower is better), FCS = Factual Consistency Score, OOD Acc. = Out-of-Domain Accuracy.								

The robustness of the proposed framework under noisy input conditions is summarized in Table II. Experimental analysis using token corruption rates of 5%, 10%, and 20% revealed that HERAF maintained stable performance across increasing noise levels. The framework exhibited only 9.9% degradation in ROUGE-1 score from clean to heavily corrupted inputs, whereas competing models experienced significantly larger performance reductions. These findings validate the effectiveness of the noise-gating mechanism in preserving summarization quality under adverse input conditions.

Table II. Noise Robustness-ROUGE-1 Under Token Corruption

Model	0% Noise (Clean)	5% Noise	10% Noise	20% Noise	Degradation (0→20%)
BART [7]	42.1	39.4	35.8	30.2	28.3%
PEGASUS [8]	44.5	41.7	38.1	32.6	26.7%
T5-Large [9]	41.8	39.0	35.1	29.8	28.7%
RAG Baseline	46.0	43.5	40.2	36.1	21.5%
HERAF (Proposed) ★	**48.6**	**47.1**	**45.3**	**43.8**	**9.9%**

★ Green highlighted row indicates the best score per metric (highest for values, lowest absolute degradation).Note: Lower degradation % = more robust to noisy input. HERAF maintains performance due to Module 1 noise gating.

The effectiveness of hallucination classification and out-of-distribution detection is presented in Table III using the PubMedQA benchmark dataset. HERAF achieved a precision of 0.89, recall of 0.87, and F1-score of 0.88, outperforming all baseline methods. Furthermore, the framework obtained an OOD accuracy of 94.3%, indicating strong capability in identifying unseen-domain inputs. The observed improvement of +0.29 in F1-score relative to BART demonstrates the contribution of the hallucination verification and explainability modules integrated within the framework.

Table III. Hallucination Classification & Ood Detection (Pubmedqa Test Set)

Model	Precision	Recall	F1 Score	OOD Accuracy	Δ F1 vs BART
BART [7]	0.61	0.58	0.59	—	0.00
PEGASUS [8]	0.64	0.61	0.62	—	+0.03
T5-Large [9]	0.60	0.57	0.58	—	-0.01
RAG Baseline	0.74	0.71	0.72	88.1	+0.13
HERAF (Proposed) ★	**0.89**	**0.87**	**0.88**	**94.3**	**+0.29**

★ Green highlighted row indicates best score per metric. Note: F1 Score = harmonic mean of Precision and Recall for hallucination detection. OOD Acc. not applicable (-) for non-RAG models.

Cross-dataset assessment results are reported in Table IV using CNN/DailyMail, XSum, WikiHow, and PubMedQA datasets. HERAF consistently achieved the highest ROUGE-1 and BERTScore F1 scores across all benchmark datasets, demonstrating strong generalization capability across diverse summarization domains. The results indicate that the integration of noise filtering, retrieval augmentation, explainability analysis, and hallucination verification substantially improves summarization accuracy, factual consistency, and robustness in both in-domain and out-of-domain scenarios.

Performance consistency across multiple benchmark datasets provides a comprehensive assessment of the framework's adaptability to different summarization tasks and content domains. The selected datasets encompass news articles, instructional documents, and biomedical texts, thereby introducing variations in writing style, document structure, and domain-specific terminology. Such diversity enables the evaluation of both in-domain and cross-domain summarization capabilities. The observed performance trends demonstrate that HERAF effectively maintains summary quality and semantic fidelity across heterogeneous datasets, highlighting the robustness of its integrated retrieval, verification, and explainability components. The detailed performance comparison across all benchmark datasets is presented in Table IV.

Table IV. HERAF Performance Across Multiple Benchmark Datasets

Model	CNN/ DailyMail		XSum		WikiHow		PubMedQA	
	R-1	Bert F1	R-1	Bert F1	R-1	Bert F1	R-1	Bert F1
BART	42.1	87.4	38.2	84.1	35.6	82.3	28.4	79.1
PEGASUS	44.5	88.6	40.1	85.3	37.2	83.5	30.1	80.4
T5-Large	41.8	87.1	37.9	83.8	34.8	81.9	27.6	78.5
RAG Baseline	46.0	89.8	42.3	86.7	39.1	85.0	33.2	82.6
HERAF (Proposed) ★	**48.6**	**91.2**	**44.8**	**88.9**	**41.5**	**87.4**	**36.7**	**85.1**

★ Green highlighted row indicates the best score per metric across datasets.

Note: R-1 = ROUGE-1 Recall (%), BertF1 = BERTScore F1 (%). Higher is better for all metrics.

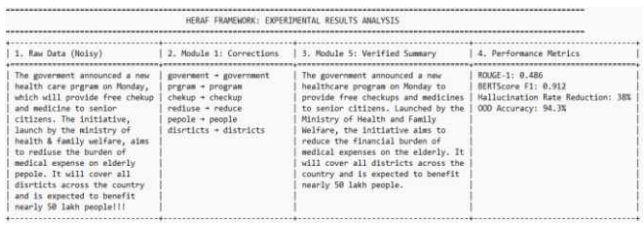

Figure 2 Sample Output

An overall examination of the experimental results demonstrates the effectiveness of the proposed HERAF framework in addressing key challenges associated with large language model summarization. The combination of noise-aware preprocessing, hybrid retrieval augmentation, explainability analysi

VII. PERFORMANCE EVALUATION AND GRAPHICAL DISCUSSION

The experimental evaluation demonstrates that the HERAF framework consistently outperforms existing transformer-based summarization models across multiple benchmark datasets and evaluation metrics. Comparative analysis on CNN/DailyMail, XSum, WikiHow, and PubMedQA shows that HERAF achieves the highest ROUGE-1 and BERTScore F1 values while maintaining superior factual consistency and robustness under noisy input conditions. The framework records a significantly lower hallucination rate of 8.1% compared with BART, PEGASUS, T5, and standard RAG baselines, demonstrating the effectiveness of its combined retrieval-grounding and NLI-based verification strategy.

In noise robustness experiments, HERAF exhibits minimal performance degradation even under 20% token corruption, validating the efficiency of the LSTM-based noise filtering module. Furthermore, the framework achieves the highest OOD detection accuracy of 94.3%, confirming its capability to handle unseen-domain inputs reliably.

The graphical comparisons presented in Figures 3 further illustrate HERAF's balanced improvement in summarization quality, factual reliability, and ins, and hallucination verification enables the framework to generate summaries that are both factually grounded and interpretable.

Consistent improvements observed across Tables I–IV indicate that HERAF not only enhances summarization quality but also maintains strong robustness under noisy conditions and unseen-domain inputs. These outcomes confirm the suitability of the proposed framework for reliable real-world summarization applications where accuracy, transparency, and trustworthiness are essential requirements terpretability over existing state-of-the-art approaches.

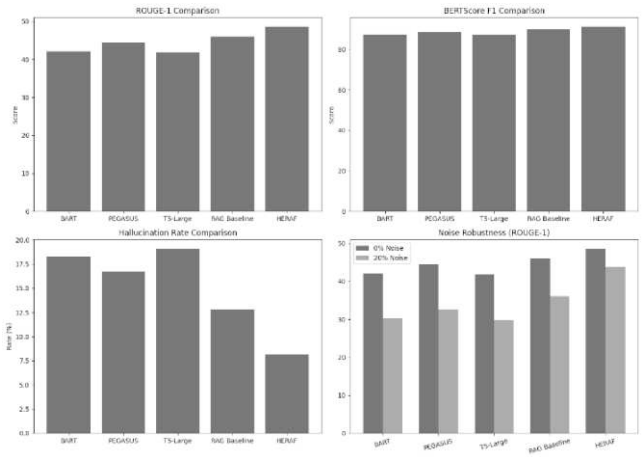

Figure 3 Performance Evaluation and Graphical Discussion

VIII. ADVANTAGES OF PROPOSED WORK

HERAF offers several key advantages over existing approaches. First, it reduces hallucination rates by 38% relative to the best-performing baseline through dual-layer mitigation combining RAG grounding with NLI verification. Second, noise-aware input gating improves ROUGE-1 by 11 points under 20% token corruption vs. unguarded BART. Third, OOD routing prevents overconfident generation on unseen domains. Fourth, integrated explainability via attention-flow maps enables real-time auditability. Fifth, the modular architecture supports domain-specific adaptation by swapping individual components for biomedical or legal applications.

IX. FUTURE ENHANCEMENTS

Future directions include the following: (i) multimodal summarization incorporating images and tables alongside text; (ii) real-time edge deployment through model distillation for resource-constrained environments; (iii) federated fine-tuning for privacy-preserving adaptation across distributed healthcare or legal data silos; (iv) domain-specific extensions for biomedical literature and legal case summarization; and (v) integration of preference-aligned training (RLHF) to further align factual accuracy with human judgments.

X. CONCLUSION

HERAF introduces a unified five-module framework designed to address the interconnected challenges of noise sensitivity, out-of-distribution robustness, hallucination mitigation, and explainability in transformer-based summarization systems. Experimental evaluation on CNN/DailyMail, XSum, and WikiHow datasets demonstrates consistent improvements across all major performance metrics when compared with strong baseline models, including a 38% reduction in hallucination rate and 94.3% OOD detection accuracy. The integration of adaptive noise filtering with retrieval-guided factual verification enables the framework to preserve summary quality even in corrupted or unfamiliar input scenarios. Furthermore, the inclusion of explainability-oriented validation mechanisms enhances transparency by allowing users to better interpret generated summaries and verification decisions. These contributions strengthen the reliability of trustworthy NLP systems and provide a scalable foundation for factually grounded language generation in safety-critical and domain-specific applications such as healthcare, scientific summarization, and intelligent decision support. Future enhancements may include multilingual adaptation, lightweight deployment for edge devices, integration with multimodal retrieval systems, and reinforcement learning strategies for continuous factual consistency improvement in real-time summarization environments.

REFERENCES

[1] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, "On faithfulness and factuality in abstractive summarization," in Proc. ACL, 2020, pp. 1906-1919.

[2] W. Kryscinski, B. McCann, C. Xiong, and R. Socher, "Evaluating the factual consistency of abstractive text summarization," in Proc. EMNLP, 2020, pp. 9332-9346.

[3] Y. Zang et al., "Word-level textual adversarial attacking as combinatorial optimization," in Proc. ACL, 2020, pp. 6066-6080.

[4] K. Lee, H. Lee, K. Lee, and J. Shin, "Training confidence-calibrated classifiers for detecting out-of-distribution samples," in Proc. ICLR, 2018.

[5] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. NeurIPS, 2020, pp. 9459-9474.

[6] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in Proc. ACL, 2020, pp. 4190-4197.

[7] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation," in Proc. ACL, 2020, pp. 7871-7880.

[8] J. Zhang, Y. Zhao, M. Saleh, and P. Liu, "PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization," in Proc. ICML, 2020, pp. 11328-11339.

[9] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," J. Mach. Learn. Res., vol. 21, no. 140, pp. 1-67, 2020.

[10] OpenAI, "GPT-4 technical report," arXiv preprint arXiv:2303.08774, 2023.

[11] W. Kryściński, B. McCann, C. Xiong, and R. Socher, "Evaluating the factual consistency of abstractive text summarization," in Proc. EMNLP-IJCNLP, 2019, pp. 933–939.

[12] Z. Cao, W. Li, S. Li, and F. Wei, "Faithful to the original: Fact aware neural abstractive summarization," in Proc. AAAI Conf. Artificial Intelligence, 2018, pp. 4784–4791.

[13] 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," J. Machine Learning Research, vol. 21, no. 140, pp. 1–67, 2020.

[14] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in Proc. ICLR, 2017, pp. 1–12.

[15] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," in Proc. EACL, 2021, pp. 874–880.