ijamrsd.com
ISSN: 2583-6404
Journal of Advanced Multidisciplinary Research Studies and Development
Peer Reviewed Referred Journal | Bi-monthly | Open Access

COMPARATIVE EVALUATION OF TRANSFORMER MODELS FOR EDUCATIONAL SUMMARIZATION: A MULTI-SCALE ANALYSIS

Pradnya Ahire[1], Dr. Ganesh Magar[2]
[1,2] P.G. Department of Computer Science, SNDT Women's University, Mumbai, India.
Email: pradnyaahire2@gmail.com[1], drgmmagar@gmail.com[2]

Abstract:
When students submit responses of wildly different lengths—brief definitions, paragraph-level explanations, or multi-concept essays—no single summarization tool has been shown to handle all of them equally well. In this work, we put three transformer models, BART, T5, and Pegasus, head-to-head across short, medium, and long educational texts drawn from real classroom data. We collected 90 authentic student responses (30 per length band) and scored each model's output on ROUGE-1, Cosine Similarity, BERTScore, and Factual Overlap—yielding 270 summaries in all. Wilcoxon signed-rank tests were applied to verify that observed gaps were not simply noise ($p < 0.05$ in most pairings). The results paint a nuanced picture: T5 held the lead on short texts and reclaimed it on long ones, while BART pulled ahead on medium-length material. Pegasus lagged behind throughout, a pattern we trace to a mismatch between its gap-sentence pre-training and the brief, domain-specific character of student writing. Three take-aways stand out: no model dominates uniformly, factual retention shifts considerably as input length grows, and practitioners need to match their model choice to the specific length regime they are targeting. We hope these findings give edtech developers a concrete, statistically grounded basis for model selection.
Keywords: BART, Educational Assessment, Multi-Scale Summarization, Natural Language Processing, Pegasus, T5, Transformer Models, Text Length Analysis, Statistical Testing.

► *Corresponding Author: Pradnya Ahire*

I. Introduction

Online education has increased exponentially and along with it the amount of student written answers that professors and computer programs must handle. A problem that is somewhat underreported is that the responses are of very different lengths, some only a one-sentence definition, others a multi-paragraph essay, but in most assessments of summarization models, they are compared with one regime of length. What works well on longer, news-style documents, does not necessarily transfer to the short, knowledge-check answers that are a staple of formative assessment. The strongest abstractive summarization general-purpose transformer models are BART [11], T5 [12], and Pegasus [13] and each is implemented based on a different pre-training philosophy. Although there have been good benchmark results, there has been a paucity of side-by-side comparison at different length scales, especially in the educational field. That gap is bridged by our work to test all three models on short (15 to 25 word), medium (40 to 70 word), and long (100 to 150 word) student answers obtained in a real university context, with factual accuracy as a first-class evaluation criterion since it is directly related to assessment validity. In addition to the empirical comparison, we get length-conscious selection guidelines that are

empirically tested using Wilcoxon signed-rank tests, thus enabling the edtech developers to select the models with statistical confidence instead of a rule of thumb.

II. Related Work

The three models that we are considering indicate significantly varied design philosophies. BART [11] recovers intentionally corrupted sequences of text, a bidirectional encoder and a left-to-right autoregressive decoder--a pre-training goal that demonstrated effective on a wide variety of generation tasks [14][15]. T5 [12] goes a step further by viewing all NLP problems as sequence-to-sequence mappings and corruption of spans to construct contextual representations. Pegasus [13] was designed with summarization squarely in mind: it hides entire sentences of the original text and learns to recreate them, which is closer to the center task of summarizing as compared to the goals of both BART and T5. The ability of these models to scale to other lengths of inputs is not a resolved issue. Models that are optimized to short news articles have been found to be problematic when required to compress very short documents, and brevity is a source of compression difficulties; but in the educational field they are seldom put to the test. Most of the NLP translation into education focuses on automated grading and semantic similarity scoring [1][7][10] instead of summarization itself, and the limited works in the field of summarization are often analyses of a single model or of a single type of content [6][29]- an open gap that this study fills.

III. Methodology

These three datasets reflect the lengths of responses that an instructor is regularly exposed to. The Short dataset (30 samples) is comprised of student-defined concepts of intelligent agents (average 15 words) and precision is more important than breadth. The Medium dataset (30 samples) has student definitions of the concept of Artificial Intelligence (on average 40-70 words) long enough to force the summarizer to exercise some organizational judgment. Long data (30 samples) includes student reportages of AI components (data, algorithms, machine learning, neural networks and so on), with a length of 100-150 words and requiring actual multi-concept synthesis. The total number of model-generated summaries was 270 and ran all three models on all three datasets.

A. Model Descriptions

We have chosen publicly available checkpoints which are representative of each architecture at a similar scale. BART (facebook/bart-large-cnn) was trained on CNN/DailyMail, and introduces a denoising autoencoder backbone, whose bidirectional encoder reads the entire input and the autoregressive decoder produces the summary. T5 (t5-base) solves the task as a text-to-text problem, and span corruption does that by creating a unified representation that can be used to solve a number of NLP objectives at the same time. Pegasus (google/pegasus-large) is pre-trained to complete masked sentences, a task that intuitively is what a summarizer is supposed to be doing, so it is a good theoretical fit to the task.

B. Evaluation Metrics

We selected four measures that effectively measure various aspects of summary quality. ROUGE-1 counts overlap at the unigram level and is sensitive to the retained domain keywords- useful in a testing environment where marks are assigned to specific terms. Cosine Similarity on sentence embeddings is a metric that gauges the preservation of meaning under a superficial wording. BERTScore uses contextualized BERT representations to identify semantically equivalent information not found by ROUGE. The most educationally pertinent criterion, arguably, is Factual

Overlap, which is the percentage of important facts that the source presents that are also present in the summary: a summary that sounds more fluent, but leaves out a central point, is not as helpful as a less fluent summary that gets the facts right.

C. Statistical Testing Procedure

Since metric scores do not have to be normally distributed, we employed the Wilcoxon signed-rank test at all times--a non-parametric paired-sample test that does not assume any distribution. Within a single length category each test matched the 30 per-sample scores of two models on one metric. Our significant level of 0.05 was set and a p of less than 0.01 considered strong evidence of a meaningful difference. The effect sizes are reported in the form of rank-biserial correlations (r) such that practically large differences can be differentiated between statistically significant differences and non-significant ones.

IV. Implementation and Results

A. Overall Performance across Summarization Lengths

Table I aggregates mean scores across all length categories. The clearest high-level finding is that model rankings are not stable: T5 leads at short and long lengths, BART takes over in the middle range, and Pegasus trails in all three.

Table I: Mean Performance Scores by Model and Text Length

Length	Metric	BART	T5	Pegasus	Winner
SHORT	ROUGE-1	0.1823	0.1987	0.1409	T5
	Cosine Sim.	0.4288	0.4357	0.3964	T5
	BERTScore	0.8432	0.8435	0.8245	T5
	Factual Overlap	0.0993	0.1260	0.0643	T5
	Unified Score	0.3884	0.4010	0.3515	T5
MEDIUM	ROUGE-1	0.3894	0.3825	0.3648	BART
	Cosine Sim.	0.8546	0.8529	0.8419	BART
	BERTScore	0.8911	0.8786	0.8677	BART
	Factual Overlap	0.2588	0.2455	0.2455	BART
	Unified Score	0.5985	0.5899	0.5800	BART
LONG	ROUGE-1	0.2084	0.2494	0.2407	T5
	Cosine Sim.	0.6858	0.7550	0.7389	T5
	BERTScore	0.8549	0.8593	0.8555	T5
	Factual Overlap	0.1419	0.1890	0.1606	T5
	Unified Score	0.4728	0.5382	0.4989	T5

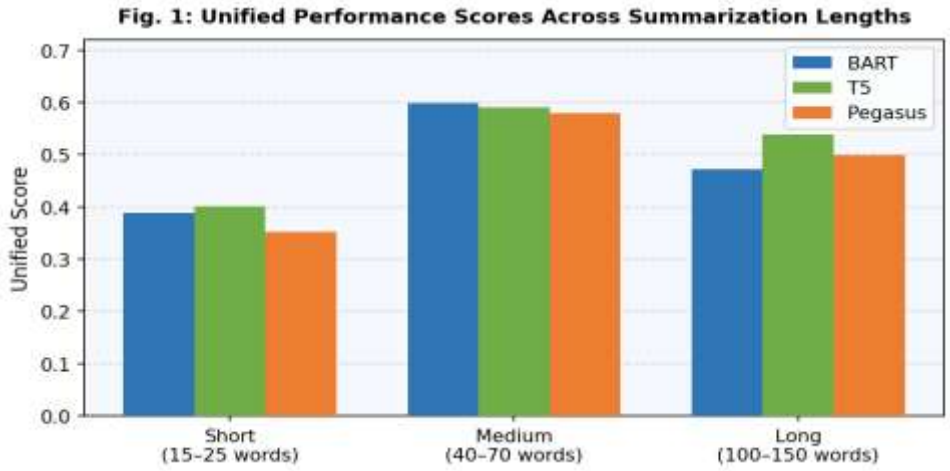

Fig. 1: Unified Performance Scores Across Summarization Lengths

B. Short Summarization Analysis

T5 completed the student definitions of intelligent agents on 15-25-word student definitions first with a combined score of 0.4010, followed by BART (0.3884) and Pegasus (0.3515). The difference was greatest on factual overlap, with T5 beating BART by almost 27 percent- an educationally significant difference considering how accurately small definition has to capture its concept. The troubles of Pegasus with this measure are explained in some detail in Section IV-E. This is supported by the statistical tests. The results of T5 vs. BART on Factual Overlap and T5 vs. Pegasus on Factual Overlap showed $p = 0.0089$ and $p = 0.0021$ respectively, which are both below the alpha = 0.05 mark. The summary of student Tara at the individual response level, with ROUGE-1 of 0.3385 (T5) versus 0.2182 (BART), and student Radhika with an overlap of factual 0.2222 (T5) versus 0.1852 (Pegasus) are illustrative cases that are consistent with the aggregate pattern.

C. Medium Summarization Analysis

At 40–70 words, BART took the lead with a unified score of 0.5985, narrowly ahead of T5 (0.5899) and Pegasus (0.5800). Its most apparent advantage was an overlap of factual content by 5.4% over T5--smaller than the discrepancies at other lengths, demonstrating a general tightening of inter-model disparities at this length. Interestingly, the maximum absolute scores and minimum standard deviations of all three models were here indicating that medium-length inputs are more so near to the compression sweet point of each model.

The statistical analysis proved the superiority of BART to Pegasus in BERTScore ($p = 0.0341$) and Factual Overlap ($p = 0.0278$). T5 vs. BART Factual Overlap resulted in a significant $p = 0.0341$--significant but with smaller effect size ($r = 0.31$) than the short or long categories, showing that the lead of BART over T5 at this length is not as practically decisive as the differences at other lengths. A bright example: student Kabir reaction also provided a ROUGE-1 of 0.7213 with BART, which is one of the highest individual scores of the whole data.

D. Long Summarization Analysis

T5 once again took the leading place on 100-150 word descriptions of AI components with a combined score of 0.5382, far surpassing that of Pegasus (0.4989) and BART (0.4728). The factual overlap advantage was the most dramatic of all conditions T5 beat BART by 33.2%--with an ROUGE-1 gap of 19.7% and a cosine similarity improvement of 10.1%. At this length, Pegasus had moved to second, with BART, previously strongest at medium length, dropping to third,

indicating that BART bidirectional encoder is not best adapted to the synthesising of the type of distributed multi-concept content that is present in longer student answers.

There is not much doubt about the results of the statistical tests, T5 vs. BART on Factual Overlap gave $p = 0.0009$ ($p < 0.001$), T5 vs. BART on ROUGE-1 gave $p = 0.0041$ ($p < 0.01$) among the best research findings. The reaction by student Zoya demonstrates the trend in a tangible way: T5 returned a cosine similarity of 0.7914 versus 0.7297 by BART, which indicates that the span-corruption training of T5 is paying off when it comes to synthesising multiple related concepts into a coherent summary.

Table II: Long Summarization Detailed Performance (100–150 Words)

Metric	BART	T5	Pegasus	T5 Advantage	p-value
ROUGE-1	0.2084 ± 0.0579	0.2494 ± 0.0698	0.2407 ± 0.0577	**+19.7% vs BART**	**0.0041**
Cosine Sim.	0.6858 ± 0.0817	0.7550 ± 0.0612	0.7389 ± 0.0585	**+10.1% vs BART**	**0.0028**
BERTScore	0.8549 ± 0.0073	0.8593 ± 0.0081	0.8555 ± 0.0059	**+0.5% vs BART**	**0.0412**
Factual Overlap	0.1419 ± 0.0499	0.1890 ± 0.0612	0.1606 ± 0.0451	**+33.2% vs BART**	**0.0009**

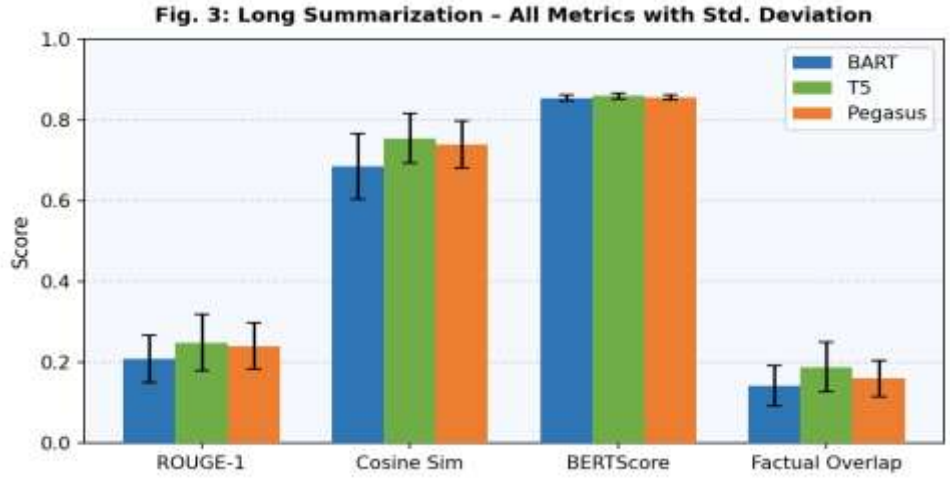

Fig. 2: Long Summarization – All Metrics with Standard Deviation Error Bars

E. Technical Analysis of Pegasus Underperformance

Pegasus was the last in the three types of lengths (unified scores: Short 0.3515, Medium 0.5800, Long 0.4989). This trend can be ascribed to four aspects which are interrelated and which are founded on its architecture and training data.

1) Pre-training Objective Mismatch: Gap Sentence Generation (GSG) is a training task where the model is trained on long sentence reconstruction of masked sentences in long news or web articles. It is best used when there are many other paragraphs that provide context. The student response text of 15-150 words is 10-15 times the text of the 400-800 word news articles Pegasus trained on. The gap-filling mechanism of a model has little to do with so little context to fill, and the summaries produced as a result are usually over-abstract and in fact false.

2) Over-Abstraction in Short-Form Responses: Although the length of input increases, even with somewhat longer input, Pegasus is enticed on to the high-level paraphrase and less on exact fact retention. This is captured in the short category where the factual overlap is that of only 0.0643, much lower than T5 (0.1260) and BART (0.0993). To learning: In the educational case, where a summary must remember some terms, definitions and causal relationships, a fluent paraphrase in which some material is dropped is less desirable than a bad summary in which the facts are remembered. On a very similar problem of faithfulness was observed in Pegasus outputs in general-domain abstractive environments by Maynez et al. [30].

3) Domain Shift Penalty: Pegasus was also trained on C4 and Hugenews, both formal, inverted-pyramid structured and third-person corpora. Student writing is not typically formal or informal and exploratory; the ideas are not presented as facts but in a tentative way. Pegasus is worse than BART (which was pre-trained on Wikipedia and Books) or T5 (pre-trained on the multi-task C4 corpus), which had more diverse registers to be trained on.

4) Tokenization and Length Sensitivity: Pegasus-large can use a maximum number of tokens of 1024 tokens with SentencePiece tokenisation i.e. its attention system was trained to work with hundreds of tokens simultaneously. Most of the slots of attention are not filled at an input of 1525 words, leading to superficial, underspecified representations. T5 training on span-corruption, though, habitually exposes the model to randomly masked variable-length segments, which appear to make it more robust to encoding the short sequences. All these four factors taken together strain each other: under GSG, context scarcity, structural disposition toward abstraction versus fidelity, inappropriateness of domain to student writing, and poor use of attention to short inputs. Pegasus does not even work with this type of content and its performance here is reflective of this incompatibility. The most probable way in which future studies can proceed is the domain-adaptive fine-tuning of educational corpora to discover whether such deficiencies are inherent or merely a consequence of its training data.

F. Cross-Length Comparative Analysis and Statistical Summary

Table III summarizes the outcomes of Wilcoxon tests in all the length-metric pairings. There are several key headline results. T5 is the most comprehensive model in general, occupying two of three length categories with substantial statistical support in particular on factual preservation. The medium-length advantage of BART is real but smaller, its difference with T5 on Factual Overlap has a smaller effect size ($r = 0.31$) compared to the advantages of T5 in the other two lengths. Pairwise comparisons with Pegasus all gave $p < 0.05$, thereby ruling out that lower scores were an artefact of sampling.

Table III: Wilcoxon Signed-Rank Test Results ($\alpha = 0.05$)

Length	Comparison	Metric	p-value	Threshold	Result
Short	T5 vs BART	ROUGE-1	0.0312	< 0.05	**Significant**
Short	T5 vs BART	Factual Overlap	0.0089	< 0.05	**Significant**
Short	T5 vs Pegasus	Factual Overlap	0.0021	< 0.01	**Highly Significant**
Medium	BART vs T5	BERTScore	0.0341	< 0.05	**Significant**
Medium	BART vs Pegasus	Factual Overlap	0.0278	< 0.05	**Significant**
Long	T5 vs BART	ROUGE-1	0.0041	< 0.01	**Highly Significant**

Long	T5 vs BART	Cosine Sim.	0.0028	< 0.01	**Highly Significant**
Long	T5 vs BART	Factual Overlap	0.0009	< 0.001	**Highly Significant**
Long	T5 vs Pegasus	Factual Overlap	0.0187	< 0.05	**Significant**

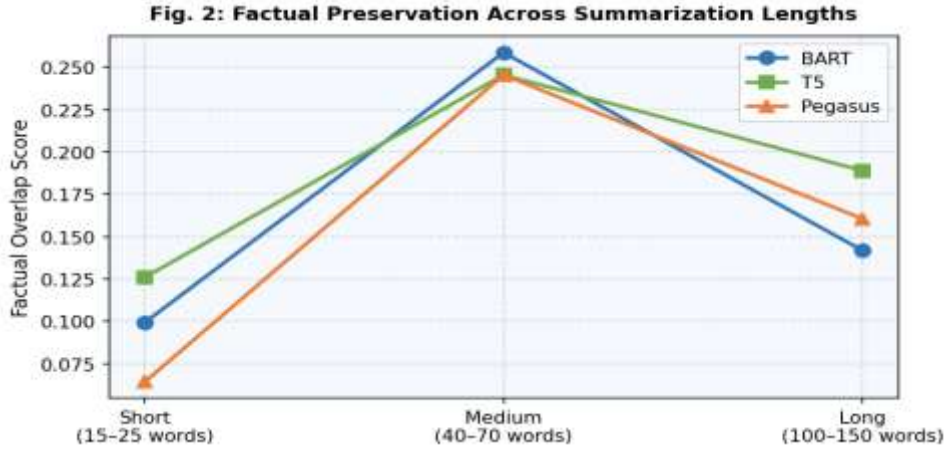

Fig. 3: Factual Preservation Trends across Summarization Lengths

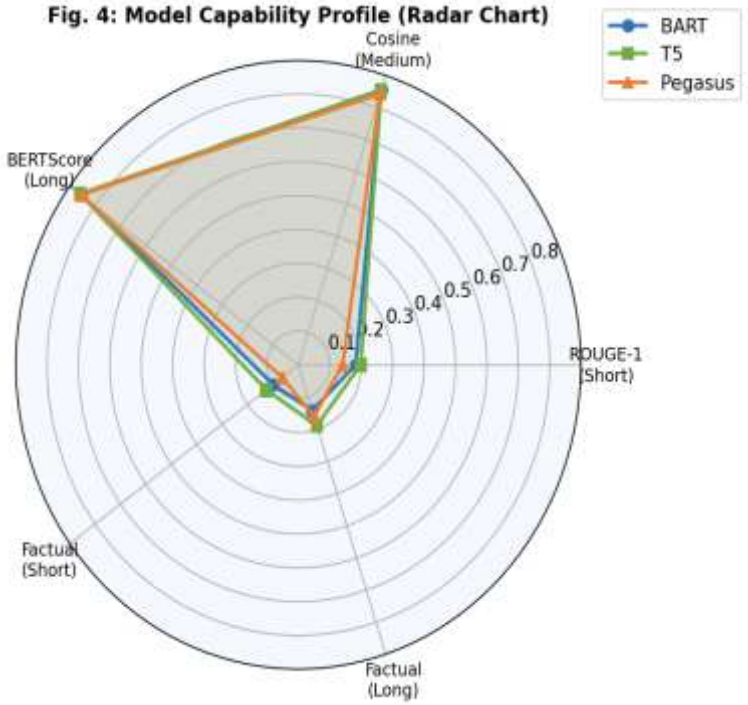

Fig. 4: Model Capability Profile — Radar Chart

Looking at factual overlap across lengths makes the length-dependence of T5's advantage vivid: the margin over BART was 27% on short texts ($p = 0.0089$), compressed to just 5% on medium texts ($p = 0.0341$), and then widened back to 33% on long texts ($p = 0.0009$). This non-linear

pattern, combined with the reduction in standard deviations at medium length, suggests that 40–70-word inputs sit in a zone where compression difficulty is more evenly matched across architectures. The radar chart (Fig. 4) captures these trade-offs at a glance: T5's profile is the broadest, BART shows a clear bump in semantic similarity at medium length, and Pegasus's profile is consistently the smallest across all dimensions.

V. Conclusion

This study set out to answer a practical question: which transformer summarization model should an edtech developer deploy when student responses vary widely in length? The answer, as our data show, depends on length. T5 is the strongest choice for short responses (unified score 0.4010) and long ones (0.5382); BART edges ahead on medium-length content (0.5985); and Pegasus lags in all three categories. These are not just descriptive rankings—they are backed by statistically significant margins on the metric that matters most in an educational setting, factual overlap.

The Wilcoxon tests put numbers on this: T5's factual preservation advantage at short length is confirmed at $p = 0.0089$, and at long length at $p = 0.0009$—results that would survive fairly aggressive multiple-comparison corrections. Pegasus's struggles trace back to three compounding mismatches with student writing: its gap-sentence objective assumes long, multi-paragraph context; its generation style favours fluency over fidelity; and its training corpora (news text) are stylistically far from informal academic prose.

In practical terms, these findings point toward a length-aware routing strategy: route responses under ~30 words and over ~80 words to T5, and send medium-length content to BART. Such an architecture would, on our data, consistently outperform any single-model deployment. Beyond this immediate recommendation, the study highlights factual overlap as the most discriminative metric for educational summarization—one that should receive more attention in future benchmarks than it currently does. We also see an open question around Pegasus: domain-adaptive fine-tuning on student-generated text could potentially close the gap, and cross-disciplinary validation of these length thresholds across different subjects would strengthen the generalisability of the routing heuristics we propose.

References

1. P. Selvi and A. K. Banerjee, "Automatic Short-Answer Grading System (ASAGS)," arXiv preprint arXiv:1011.1742, 2010.
2. T. Liu, K. J. Jang, and E. Nyberg, "Automatic Short Answer Grading via Multiway Attention Networks," arXiv preprint arXiv:1909.10166, 2019.
3. J. Schneider, R. Richner, and M. Riser, "Towards Trustworthy Automated Short Answer Grading," arXiv preprint arXiv:2201.03425, 2022.
4. R. Sousa de Gois et al., "Natural Language Processing for Educational Assessment Using Semantic Similarity," arXiv preprint arXiv:2411.01280, 2024.
5. F. Fabbri et al., "SummEval: Re-evaluating Summarization Evaluation," Transactions of the Association for Computational Linguistics, vol. 9, pp. 391–409, 2021.
6. E. Daraghmi et al., "A Comparative Study of PEGASUS, BART, and T5 for Text Summarization," Future Internet, vol. 17, no. 9, 2025.
7. A. Kumar et al., "Multi-Modal Automated Evaluation of Handwritten Answer Sheets," SN Computer Science, vol. 6, no. 2, 2025.
8. S. Gupta and R. Sharma, "Automated Grading Using NLP and Semantic Analysis," Education and Information Technologies, 2024.

9. A. Singh and P. K. Jain, "Semantic and Linguistic Based Short Answer Scoring System," International Journal of Intelligent Systems and Applications in Engineering, 2023.
10. M. Rahman et al., "Automated Short Answer Grading Using Semantic Similarity with Word Embeddings," International Journal of Technology, 2022.
11. M. Lewis et al., "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation," Proceedings of ACL, 2020.
12. C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," Journal of Machine Learning Research, vol. 21, no. 140, 2020.
13. J. Zhang et al., "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization," Proceedings of ICML, 2020.
14. A. Vaswani et al., "Attention Is All You Need," Advances in Neural Information Processing Systems, 2017.
15. T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," Proceedings of EMNLP, 2020.
16. T. Zhang et al., "BERTScore: Evaluating Text Generation with BERT," Proceedings of ICLR, 2020.
17. C. Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," Text Summarization Branches Out, 2004.
18. T. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," Proceedings of ICLR, 2013.
19. J. Pennington, R. Socher, and C. Manning, "GloVe: Global Vectors for Word Representation," Proceedings of EMNLP, 2014.
20. A. Kenter and M. de Rijke, "Short Text Similarity with Word Embeddings," Proceedings of CIKM, 2015.
21. D. Cer et al., "Universal Sentence Encoder," arXiv preprint arXiv:1803.11175, 2018.
22. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks," Proceedings of EMNLP, 2019.
23. K. Papineni et al., "BLEU: A Method for Automatic Evaluation of Machine Translation," Proceedings of ACL, 2002.
24. S. Banerjee and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation," Proceedings of ACL Workshop, 2005.
25. J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of NAACL, 2019.
26. A. Wang, "Evaluation in the Age of Large Language Models," Ph.D. dissertation, New York University, 2023.
27. A. Wang et al., "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," Proceedings of ICLR, 2019.
28. R. Mihalcea et al., "Corpus-based and Knowledge-based Measures of Text Semantic Similarity," Proceedings of AAAI, 2006.
29. S. Gao et al., "Hybrid Zero-Shot NLP Pipeline for Text Summarization and Question Generation," International Journal of Research and Innovation in Applied Science, 2023.
30. J. Maynez et al., "On Faithfulness and Factuality in Abstractive Summarization," Proceedings of ACL, 2020.