A COMPREHENSIVE COMPARISON OF TEXT SUMMARIZATION PERFORMANCE: A MULTI-FACETED EVALUATION OF LARG LANGUAGE MODELS WITH PRACTICAL CONSIDERATIONS

Anantharaman Janakiraman and Behnaz Ghoraani

Department of Electrical Engineering and Computer Science, Florida Atlantic University

ABSTRACT

Text summarization is crucial for mitigating information overload across domains. This research evaluates summarization performance across 17 large language models using seven diverse datasets at three output lengths (50, 100, 150 tokens). We employ a novel multi-dimensional framework assessing factual consistency, semantic similarity, lexical overlap, and human-like evaluation while considering both quality and efficiency factors. Key findings reveal significant differences between models, with specific models excelling in factual accuracy (deepseek-v3), human-like quality (claude-3-5-sonnet), processing efficiency (gemini-1.5-flash), and cost effectiveness (gemini-1.5-flash). Performance varies dramatically by dataset, with models struggling on technical domains but performing well on conversational content. We identified a critical tension between factual consistency (best at 50 tokens) and perceived quality (best at 150 tokens).

KEYWORDS

Text Summarization, Large Language Models, Multi-dimensional Evaluation, Evaluation Metrics, Model Comparison

1. INTRODUCTION

Text summarization systems condense information in source documents while preserving key content, enabling users to understand essential information without reading entire documents. Despite recent progress in large language models (LLMs), comprehensive evaluation frameworks accounting for multiple quality dimensions alongside practical deployment concerns remain underdeveloped.

Traditional evaluation methods in text summarization have relied on lexical overlap metrics (e.g.,ROUGE [1]), which cannot fully capture semantic equivalence, factual consistency, and other critical dimensions of summary quality. Our research addresses these limitations by proposing a balanced multidimensional evaluation framework that assigns appropriate weights to factual consistency (35%), semantic similarity (25%), lexical overlap (15%), and human-like evaluation (25%), while incorporating a 70/30 quality efficiency split to assess practical deployment considerations.

David C. Wyld et al. (Eds): CSML, AISCA, DNLP, SOEA, NET, BDHI, SIPO – 2026
pp. 72-91, 2026. CS & IT - CSCP 2026 DOI: 10.5121/csit.2026.160207

This research makes several key contributions: (1) conducting a comprehensive comparative analysis of 17 state-of-the-art LLMs using a balanced multidimensional framework; (2) offering evidence-based insights into quality-efficiency relationships; (3) providing detailed recommendations for model selection across different use cases; and (4) establishing a replicable evaluation methodology that better aligns with real-world requirements.

2. BACKGROUND AND RELATED WORK

Text summarization approaches include extractive methods (selecting important sentences verbatim) and abstractive approaches (generating new text conveying essential information). Summarization can target single or multiple documents, create generic or query-focused outputs, and produce indicative or informative summaries.

Evaluating summarization systems presents inherent challenges due to subjectivity in defining "good" summaries. Traditional approaches include: Automated Reference-based Metrics like ROUGE [1] (lexical overlap) and BERTScore [2] (semantic similarity); Reference-Free Evaluation methods like SummaC [3] (factual consistency); LLM-Based Evaluation using models like GPT-4 or Claude; Human Evaluation assessing relevance, coherence, and readability; and Efficiency and Deployment Metrics considering processing time, computational requirements, and operational costs.

Most prior comparative studies focus on individual quality dimensions without considering the multifaceted nature of summarization quality or practical deployment factors. Our study bridges this gap by integrating multiple quality dimensions with efficiency metrics, enabling an assessment more representative of real-world requirements where model selection must balance quality and practical constraints.

3. METHODOLOGY

Our research employs a systematic approach to evaluate text summarization capabilities across a diverse range of models, datasets, and output lengths. This section details our experimental setup, including model selection, datasets, evaluation metrics, and procedural details.

3.1. Models

We evaluate 17 models representing a diverse range of architectures, capabilities, and accessibility. Table 1 provides an overview of the models included in our evaluation.

This selection encompasses a range of model sizes, architectures, and training approaches, allowing us to identify performance patterns across different model families and scales.

3.2. Multi-Dimensional Evaluation Framework

Text summarization quality cannot be adequately assessed through a single metric or under a single condition. Our research employs a comprehensive multi-dimensional evaluation framework (Figure 1) that systematically evaluates LLM performance across three key dimensions: quality, efficiency, and content. By analyzing these dimensions simultaneously, we can identify complex trade-offs and interactions that would not be apparent from single-dimension evaluations, such as how factual consistency varies with summary length across different domains, or how efficiency considerations influence model selection for specific use cases.

3.3. Datasets

To ensure comprehensive evaluation across diverse domains and summarization challenges, we selected Table 1: Overview of evaluated models and their access methods

Model Family	Model Name	Type	Access Method
Anthropic	claude-3-5-haiku	Commercial	API
	claude-3-5-sonnet	Commercial	API
	claude-3-opus	Commercial	API
Google	gemini-1.5-flash	Commercial	API
	gemini-1.5-pro	Commercial	API
	gemini-2.0-flash	Commercial	API
DeepSeek	deepseek-v3	Commercial	API
OpenAI	gpt-3.5-turbo gpt-4-turbo gpt-4o gpt-4o-mini	Commercial Commercial Commercial Commercial	API API API API
	o1	Commercial	API
	o1-mini	Commercial	API
Open-source	deepseek-7b falcon-7b llama-3.2-3b	Open-source Open-source Open-source	Local inference Local inference Local inference
	mistral-7b	Open-source	Local inference

Seven datasets representing different text types, styles, and complexity levels.

Table 2 provides an overview of these datasets, which span news (CNN/Daily Mail, XSum), technical documentation (BigPatent), legal texts (BillSum), scientific literature (PubMed), conversational dialogues (SAMSum), and instructional content (WikiHow).

Table 2: Detailed characteristics of datasets used in the evaluation

Dataset	Domain	Characteristics
CNN/Daily Mail	News (journalism)	Multi-paragraph news articles covering a wide range of topics
XSum	News (BBC)	News articles spanning various topics with diverse writing styles
BigPatent	Technical (patent documentation)	Technical documents with specialized terminology and complex structures
BillSum	Legal (U.S. congressional bills)	Legal and legislative language with domain-specific terminology
PubMed	Scientific (biomedical literature)	Scientific articles with specialized medical terminology
SAMSum	Conversational (dialogues)	Messenger-like conversations between multiple participants
WikiHow	Instructional (how-to guides)	Step-by-step procedures across various topics with instructional intent

For each dataset, we randomly sampled 30 documents to ensure feasible evaluation while maintaining representative coverage across document lengths and complexity levels. For particularly long source documents (BigPatent and PubMed), we truncated inputs to 4,096 tokens to accommodate model context window limitations while preserving essential content.

These diverse datasets enable us to evaluate multiple performance dimensions: domain adaptation (handling specialized terminology), text length handling (from brief dialogues to lengthy technical documents), abstractive capacity (generating novel phrasings rather than verbatim extraction), and information density (effectiveness across varying compression requirements).

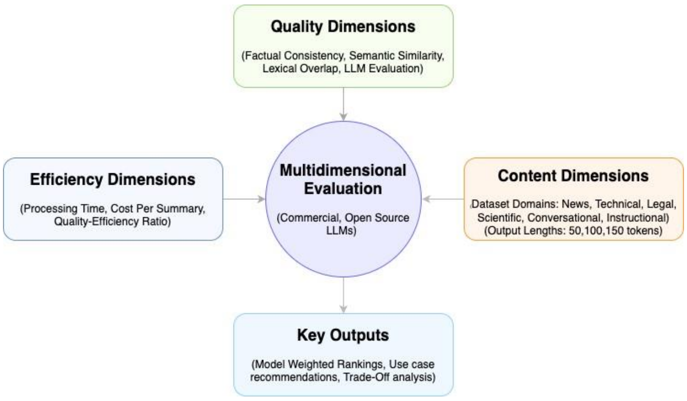

Quality Dimensions
(Factual Consistency, Semantic Similarity, Lexical Overlap, LLM Evaluation)
Efficiency Dimensions
(Processing Time, Cost Per Summary, Quality-Efficiency Ratio)
Multidimensional Evaluation
(Commercial, Open Source LLMs)
Content Dimensions
Dataset Domains: News, Technical, Legal, Scientific, Conversational, Instructional)
(Output Lengths: 50,100,150 tokens)
Key Outputs
(Model Weighted Rankings, Use case recommendations, Trade-Off analysis)

Figure 1: Multi-dimensional evaluation framework for assessing large language models on text summarization

3.4. Evaluation Metrics

Our evaluation framework employs a balanced multi-dimensional approach capturing different aspects of summarization quality and practical deployment factors (Table 3). For quality assessment, we use ROUGE (lexical overlap), BERTScore (semantic similarity), SummaC (factual consistency), and LLM based evaluation (human-like assessment). For efficiency, we measure processing time and cost per summary.

Table 3: Quality evaluation metrics with balanced weights

Metric Category	Representative Measures	Weight
Factual Consistency	SummaC scores	35%
Semantic Similarity	BERTScore F1	25%
Lexical Overlap	ROUGE-1, ROUGE-2, ROUGE-L	15%
Human-like Quality	LLM-based evaluation	25%

Our ranking methodology combines component rankings (quality metrics, factual consistency, humanlike evaluation, efficiency, cost) into a final score using weighted ranks: quality rank (30%), factual consistency rank (25%), human-like evaluation rank (20%), efficiency rank (15%), and cost efficiency rank (10%).

3.5. Experimental Setup

For each model-dataset pair, we generated summaries at three output lengths (50, 100, 150 tokens), processing 30 examples per configuration using consistent minimal prompts with temperature=0.1. API based models used their respective service providers, while open-source models were evaluated on NVIDIA A100 GPUs. For each summary, we computed all metrics, collected timing information, and tracked token usage to calculate costs.

3.6. Ranking Methodology

To produce comprehensive rankings balancing quality and efficiency,we implemented a two-levelweighting scheme as shown in Table 4. First, we ranked models separately on different components (quality metrics, factual consistency, human-like evaluation, efficiency, and cost). Second, we combined these component rankings into a final score using our balanced weights: quality rank (30%), factual consistency rank (25%), human-like evaluation rank (20%), efficiency rank (15%), and cost efficiency rank (10%).

Additionally, for applications requiring both quality and efficiency, we applied a quality-efficiency tradeoff analysis with a 70/30 split (70%weighted toward combined quality metrics and 30% toward combined efficiency metrics). This balanced approach identifies models excelling in specific dimensions while recognizing those providing the best overall value across multiple evaluation criteria, reflecting the practical reality that most applications prioritize output quality while considering computational and financial constraints.

Table 4: Ranking components with balanced weights

Ranking Component	Weight	Description
Quality Rank	30%	Combined quality metrics rank
Factual Consistency Rank	25%	Specific emphasis on factual consistency
Human-like Evaluation Rank	20%	LLM-based evaluation importance
Efficiency Rank	15%	Processing time considerations
Cost Efficiency Rank	10%	Budget impact for production systems

3.7. Implementation Details

The evaluation pipeline was implemented in Python, utilizing HuggingFace'stransformers and datasets libraries for model access and data loading, SummaC implementation for factual consistency evaluation, BERTScore and ROUGE implementations for semantic and lexical evaluation, API clients for commercial models (OpenAI, Anthropic, Google, DeepSeek), and PyTorch for local model inference. Our implementation allows for extension to additional models, datasets, and metrics as they become available.

4. RESULTS

This section presents the comprehensive evaluation results of 17 large language models across seven diverse datasets and three output lengths. We organize our findings to address multiple dimensions of summarization performance: overall model rankings using our balanced weighting scheme, specialized analyses of factual consistency and human-like quality assessment, the critical impact of output length on different quality metrics, domain-specific performance patterns, efficiency considerations including processing time and cost, andquality-efficiencytrade-offsrelevanttopracticaldeploymentdecisions. Together, these results provide a multifaceted view of model capabilities and limitations across different use cases and constraints.

4.1. Overall Performance

Based on our balanced evaluation framework that prioritizes quality (30%), factual consistency (25%), human-like evaluation (20%), efficiency (15%), and cost (10%), we find significant differences in performance across the 17 evaluated models. As shown in Table 5, gpt-3.5-turbo emerges as the top overall performer with a weighted rank of 3.25, excelling in both quality metrics and factual consistency while maintaining good efficiency. Notably, deepseek-v3 achieves the second position despite ranking lower on efficiency, primarily due to its exceptional performance in factual consistency. Figure 2 shows the performance of all evaluated models.

4.2. Factual Consistency

Table 5: Top 5 models by weighted rank across all metrics. Lower ranks are better (1 = best). The weighted rank combines the component ranks according to our consistent weighting system.

Model	**Quality**	**Factual**	**Efficiency**	**Cost-Effect.**	**Weighted**
	Rank	**Rank**	**Rank**	**Rank**	**Rank**
gpt-3.5-turbo	2	2	3	5	3.25
deepseek-v3	1	1	13	4	4.30
claude-3-5-haiku	4	9	6	6	5.55
gpt-4o-mini	6	5	4	3	5.55
gemini-2.0-flash	5	14	2	2	6.30

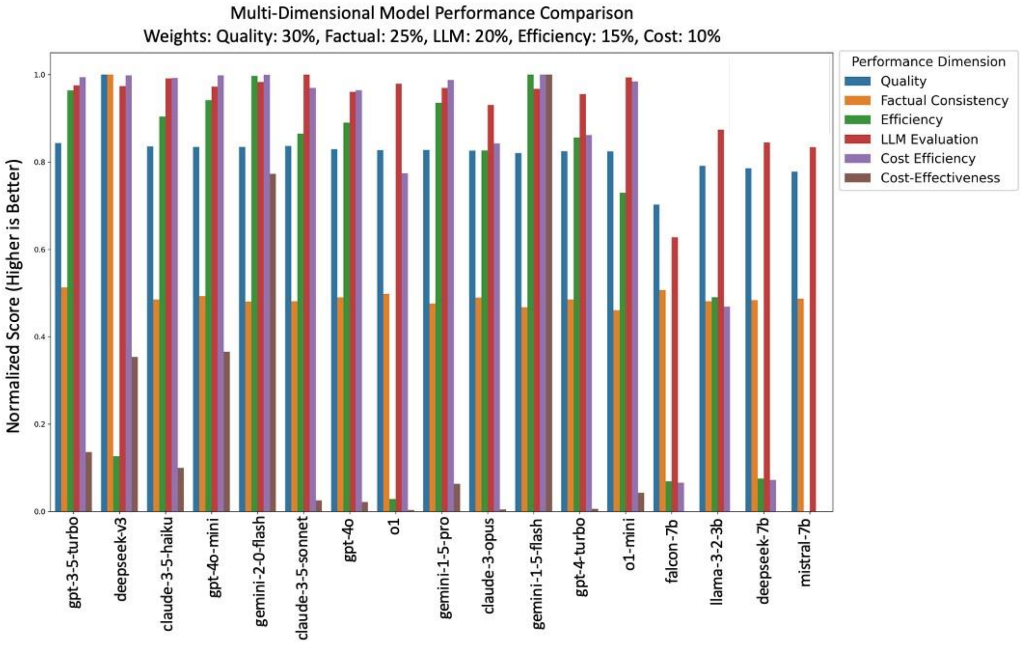


Figure 2: Comparative analysis of 17 models across all evaluation dimensions, normalized to the 0-1 range for fair comparison.

Factual consistency, weighted at 35% in our quality metrics framework, shows particularly interesting patterns. As shown in Table 6, deepseek-v3 dramatically outperforms all other models with a SummaC score of 0.6823, which is 94.9% higher than then ext best model(gpt-3.5-turboat0.3501). This exceptional performance in factual consistency suggests that deepseek-v3 has been specifically optimized to avoid generating content that contradicts or misrepresents the source material.

The remaining top models for factual consistency show relatively similar scores, clustering in the 0.33-
0.35 range, with both commercial API models (gpt-3.5-turbo, gpt-4o-mini, o1) and open-source models (falcon-7b) represented. This suggests that both proprietary and open-source approaches can achieve comparable levels of factual reliability.

4.3. Human-like Evaluation

For human-like quality assessment, measured through our LLM-based evaluation on a 1-5 scale and weighted at 25% in our quality metrics, Anthropic's Claude models demonstrate superior performance. As shown in Table 7, claude-3-5-sonnet achieves the highest score (4.75/5.0), followed closely by o1-mini and claude-3-5-haiku. These results indicate that these models excel at generating summaries that human evaluators would likely judge as high-quality in terms of relevance, coherence, factual accuracy, and conciseness.

Table 6: Models with highest factual consistency

Model	SummaC Score
deepseek-v3	0.6823
gpt-3.5-turbo	0.3501
falcon-7b	0.3460
o1	0.3399
gpt-4o-mini	0.3363

Table 7: Models with highest LLM-based evaluation scores (scale: 1-5)

Model	LLM Score
claude-3-5-sonnet	4.75
o1-mini	4.72
claude-3-5-haiku	4.70
gemini-2.0-flash	4.66
o1	4.65

4.4. Effect of Output Length

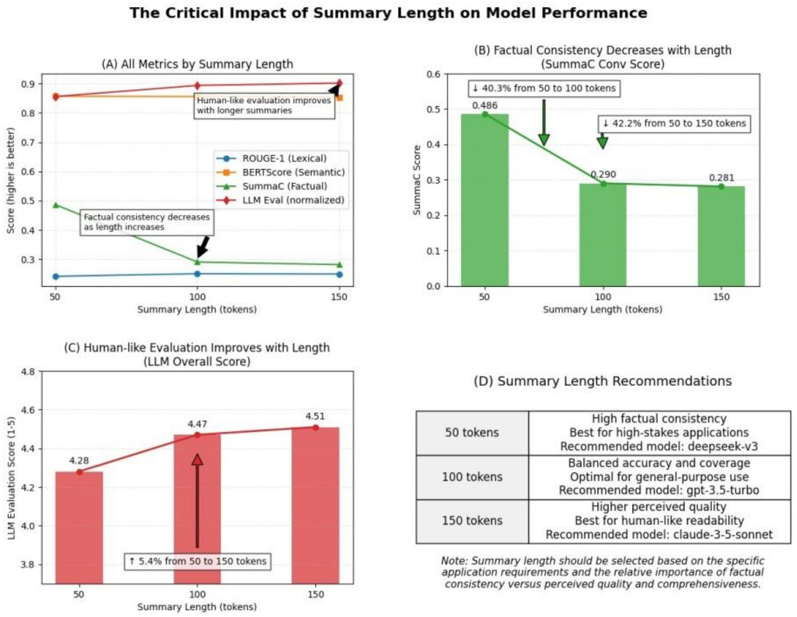
Output length analysis reveals critical trade-offs across quality dimensions (Table 8, Figure 3). ROUGE-1 F1 peaks at 100 tokens (0.250), while SummaC scores are dramatically higher at 50 tokens (0.486) compared to longer summaries (100: 0.290, 150: 0.281). Conversely, LLM ratings increase with length, peaking at 150 tokens (4.51). This reveals a fundamental tension: human-like evaluators prefer longer summaries, while factual consistency is substantially better with shorter outputs. BERTScore shows minimal variation but slightly decreases as length increases. This inverse relationship between factual accuracy and perceived quality has significant implications for applications where different quality dimensions may have varying importance.

Table 8: Average quality metrics by token length

Tokens	ROUGE-1	BERT	SummaC	LLM
50	0.241	0.857	0.486	4.28
100	0.250	0.856	0.290	4.47
150	0.249	0.854	0.281	4.51

4.5. Performance by Dataset

Performance varies substantially across datasets (Table 9). SAMSum (conversational dialogues) leads with top scores in BERTScore (0.883), SummaC (0.531), and LLM evaluation (4.77), indicating models excel at summarizing dialogues. Conversely, technical domains—BigPatent and PubMed—yield notably poor factual consistency scores (0.093 and 0.102), showing models struggle with specialized terminology. XSum combines the lowest ROUGE-1 (0.160) with strong factual consistency (0.480), reflecting its abstractive style that diverges lexically while maintaining factual alignment. BillSum shows strong lexical overlap (0.309 ROUGE-1) but lower human-like quality (4.18), suggesting content is captured but readability suffers. These domain-specific patterns highlight the importance of considering dataset characteristics when evaluating performance and selecting models for specific applications.



50 tokens	High factual consistency Best for high-stakes applications Recommended model: deepseek-v3
100 tokens	Balanced accuracy and coverage Optimal for general-purpose use Recommended model: gpt-3.5-turbo
150 tokens	Higher perceived quality Best for human-like readability Recommended model: claude-3-5-sonnet

Note: Summary length should be selected based on the specific application requirements and the relative importance of factual consistency versus perceived quality and comprehensiveness.

Figure 3: Impact of Summary Length on Performance Metrics: This comprehensive visualization shows how different quality metrics vary with summary length (50, 100, and 150 tokens).

Table 9: Average quality metrics by dataset

Dataset	ROUGE-1	BERT	SummaC	LLM
BigPatent	0.288	0.853	0.093	4.34
BillSum	0.309	0.853	0.439	4.18
CNN/DailyMail	0.242	0.864	0.426	4.57
PubMed	0.225	0.833	0.102	3.93
SAMSum	0.307	0.883	0.531	4.77
WikiHow	0.195	0.846	0.397	4.62
XSum	0.160	0.857	0.480	4.55

4.6. Efficiency Analysis

Processing efficiency, critical for real-world applications, shows significant variations across models (Table 10). Google’s Gemini models dominate the efficiency rankings, with gemini-1.5-flash leading at 1.08 seconds average processing time per summary. The efficiency advantage extends to cost metrics as well, with gemini-1.5-flash being the most cost-effective at approximately $0.000124 per summary.

When examining cost-effectiveness (Table 11), which considers the ratio of quality to cost, the Gemini models again demonstrate exceptional performance. Notably, deepseek-v3 appears in the top tier despite its lower processing efficiency, highlighting how its exceptional factual consistency provides value that offsets its higher computational requirements.

4.7. Quality-Efficiency Trade-offs

For real-world deployments, the balance between quality and efficiency is essential. Using a 70/30 quality

Table 10: Top 5 models by processing time

Model	Time (s)	Cost ($)
gemini-1.5-flash	1.08	0.00012
gemini-2.0-flash	1.14	0.00016
gpt-3.5-turbo	1.90	0.00108
gpt-4o-mini	2.41	0.00034
gemini-1.5-pro	2.55	0.00205

Table 11: Top 5 models by cost-effectiveness (Quality/Cost ratio)

Model	Value Score
gemini-1.5-flash	9696
gemini-2.0-flash	7493
gpt-4o-mini	3545
deepseek-v3	3432
gpt-3.5-turbo	1322

Efficiency split (Table12), deepseek-v3 emerges as the model with the best over all balance (0.738), followed closely by gpt-3.5-turbo (0.720) and gemini-2.0-flash (0.718). This balanced perspective provides a more holistic view of model capabilities that better reflects practical deployment considerations.

Table 12: Models with best balance of quality (70%) and efficiency (30%)

Model	Balance Score
deepseek-v3	0.738
gpt-3.5-turbo	0.720
gemini-2.0-flash	0.718
gpt-4o-mini	0.714
claude-3-5-haiku	0.712

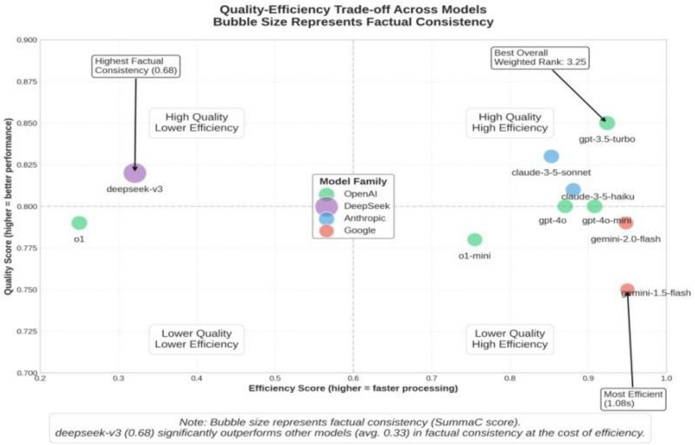
This balanced perspective, illustrated in Figure 4, provides a more holistic view of model capabilities that better reflects practical deployment considerations. The visualization highlights the trade-offs between different dimensions, with some models prioritizing quality at the expense of efficiency (deepseek-v3) and others emphasizing speed with competitive quality (gemini models).

5. Discussion

This section interprets our experimental findings and explores their implications for both research and practice. We examine patterns of model specialization that emerge across different domains and quality dimensions, analyze the complex relationships between summary length and various quality metrics, discuss the significant challenges models face in technical and specialized domains, and provide evidence based recommendations for selecting appropriate models and configurations for specific use cases. These insights bridge the gap between quantitative evaluation results and practical decision-making in real world summarization applications.

5.1. Model Specialization

Our analysis reveals distinct patterns of model specialization across dimensions and domains. While most models perform best on the SAMSum dataset (dialogue summarization), deepseek-v3 shows particular



Quality-Efficiency Trade-off Across Models
Bubble Size Represents Factual Consistency

Note: Bubble size represents factual consistency (SummaC score).
deepseek-v3 (0.68) significantly outperforms other models (avg. 0.33) in factual consistency at the cost of efficiency.

Figure 4: Quality-Efficiency Trade-off: Bubble size represents factual consistency score; position shows the balance between quality (y-axis) and efficiency (x-axis) metrics.

strength on BigPatent, suggesting domain-specific optimization for technical content. This specialization indicates that models may be trained or optimized differently depending on their intended use cases.

The exceptional factual consistency of deepseek-v3 is particularly noteworthy. With a SummaC score of 0.6823, nearly double that of other models, it represents a significant advancement in faithful summarization. This performance comes at the cost of efficiency, however, as deepseek-v3 ranks 13th in processing time (21.03 seconds per summary). This trade-off exemplifies the tension between quality and efficiency that practitioners must navigate.

5.2. Length-Quality Relationships

Our findings reveal critical trade-offs across summary lengths. Factual consistency decreases dramatically as length increases (0.486 at 50 tokens to 0.281 at 150tokens), while human-like evaluation scores show the opposite trend (4.28 at 50 tokens to 4.51 at 150 tokens). ROUGE peaks at 100 tokens (0.250), suggesting a middle ground for content coverage. This inverse relationship between factual accuracy and perceived quality requires application-specific calibration: shorter summaries when factual accuracy is paramount and longer summaries when perceived quality matters more. This represents a fundamental challenge in summarization system design, requiring deliberate length choices based on specific use case priorities.

5.3. Domain Challenges

The substantial variation in performance across datasets highlights domain-specific challenges in text summarization. The poor factual consistency on technical domains (Big Patent: 0.093, PubMed: 0.102) compared to conversational content (SAMSum: 0.531) reveals a significant gap in models' ability to maintain factual accuracy when handling specialized terminology and complex concepts. This domain gap suggests a need for domain-specific fine-tuning and caution when deploying general-purpose models for technical domains, where factual reliability is especially critical. Evaluating models on diverse datasets is essential for understanding their true capabilities rather than relying on performance metrics from a single domain that may not generalize.

5.4. Practical Recommendations

Based on our evaluation, we recommend specific models for different use cases: deepseek-v3 with 50- token outputs for high-stakes applications requiring factual consistency (SummaC: 0.6823); claude-3-5-sonnet with 150-token outputs for human-like quality (LLM score: 4.75/5.0); gpt-3.5-turbo with 100-token outputs for balanced general-purpose summarization (weighted rank: 3.25); gemini-1.5-flash (1.08s) or gemini-2.0-flash (1.14s) for resource-constrained applications; gemini-1.5-flash for optimal cost-effectiveness (score: 9695.99); and domain-specific model selection guided by targeted performance metrics, with deepseek-v3 showing advantages for technical content.

6. CONCLUSION

This study makes significant contributions by introducing a comprehensive evaluation framework integrating automated metrics and LLM-based evaluations across multiple models, datasets, and summary lengths. Our innovative approach provides a more nuanced assessment by emphasizing factual consistency, semantic similarity, readability, and efficiency. We identify key strengths of specific models and provide actionable insights tailored to real-world applications. Future work should build upon our findings by enhancing factual consistency metrics, exploring domain-adaptation techniques, evaluating newer models, and investigating hybrid human-model systems.

7. LIMITATIONS

While our evaluation provides valuable insights, several factors should be considered when interpreting the results. Automated metrics, such as ROUGE scores, present inherent limitations in capturing valid yet lexically diverse summaries, and their reliance on human-written references may not reflect the full scope of valid summarization possibilities. Additionally, our sample size was relatively modest (30 examples per model-dataset-token length combination), suggesting that larger-scale evaluations might yield different trends or more robust findings. Human-like quality assessments, conducted using LLM-based evaluations with a limited subset (10 samples per configuration), serve as proxies and might not precisely mirror actual human judgments. Performance variations across different domains further indicate that our results may not generalize uniformly, underscoring the potential need for domain-adapted models. The weighting scheme adopted (35% factual consistency, 25% semantic similarity, 15% lexical overlap, 25% human-like evaluation) offers a balanced but potentially subjective representation of evaluation criteria. Finally, the rapid pace of model development implies that newer models released after this evaluation may exhibit improved capabilities, necessitating continual reassessment.

ACKNOWLEDGEMENTS

In accordance with ACL’s policy on AI writing assistance, we acknowledge the use of Claude 3.7 Sonnet to help with language efficiency and formatting to meet page constraints while preserving the original content and findings of this research.

REFERENCES

[1] Lin,C.Y.(2004)"ROUGE: A Package for Automatic Evaluation of Summaries", Text Summarization Branches Out.

[2] Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q. &Artzi, Y. (2020) "BERTScore: Evaluating Text Generation with BERT", International Conference on Learning Representations.

[3] Laban, P., Schnabel, T., Bennett, P.N. & Hearst, M.A. (2022) "SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization", Transactions of the Association for Computational Linguistics, Vol. 10, pp1504-1520.

[4] Shi, T., Keneshloo, Y., Ramakrishnan, N., & Reddy, C.K. (2021) "Neural Abstractive Text Summarization with Sequence-to-Sequence Models", ACM/IMS Transactions on Data Science, Vol. 2, No. 1, pp. 1-37.

[5] Laban, P., Hsi, A., Canny, J., & Hearst, M.A. (2022) "Factual Consistency Evaluation for Text Summarization via Counterfactual Estimation", Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 323-336.

[6] Fan, C., Wang, Y., Liu, Y., Suo, Q., Wang, X., Tian, W., Tang, Z., Zheng, Y., Yin, D., Wang, W., Cui, R., Chen, S.N., & Ng, H.T. (2023) "A Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity", arXiv preprint arXiv:2310.07521.

[7] Laban, P., Dziri, N., Ghazvininejad, M., &Radev, D.(2022)"SummaC:Re-Visiting NLI- based Models for Inconsistency Detection in Summarization", Transactions of the Association for Computational Linguistics, Vol. 10, pp. 163-177.

[8] Lin, S., Hilton, J., & Evans, O. (2022) "TruthfulQA: Measuring How Models Mimic Human Falsehoods", Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Vol. 1, pp. 3214-3252.

[9] Bender, E.M., Gebru, T., McMillan-Major, A., & Shmitchell, S.(2021)"On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?", Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pp. 610-623.

[10] Chen, S., Xu, Y., Zhang, J., & Zhou, Y. (2024) "Empowering LLM summarization: The impact of integrating human feedback for model fine-tuning", Information Fusion, Vol. 103, p. 102172.

[11] Zhang, H., Yu, P.S., & Zhang, J. (2024) "Text Summarization Comparative Analysis", arXiv preprint arXiv:2406.11289.

[12] Nguyen, H., Chen, H., Pobbathi, L., & Ding, J. (2024) "A Comparative Study of Quality Evaluation Methods for Text Summarization", arXiv preprint arXiv:2407.00747v1, University of North Texas, Denton, TX.

[13] Koh, H.,Ju, J., Liu, M., & Pan, S. (2022) "An Empirical Survey on Long Document Summarization: Datasets, Models and Metrics", arXiv preprint arXiv:2207.00939.

APPENDIX

A. Full Ranking Table

Table 13 provides the comprehensive ranking of all evaluated models across multiple dimensions, including quality, factual consistency, efficiency, and cost metrics.

B. Detailed Quality Metrics

Table 14 presents the raw scores for all quality metrics across models, providing a more granular view of performance.

C. Model Weighted Rankings

Figure 5 illustrates that weighted ranking combines all evaluation components according to our consisTable 13: Full model rankings across quality dimensions, efficiency, and cost-effectiveness. Lower rank numbers indicate better performance (1 = best). tent weighting system to provide an overall assessment of model performance.

Model	Quality Rank	Factual Rank	Efficiency Rank	Cost-Effect. Rank	Weighted Rank
gpt-3.5-turbo	2	2	3	5	3.25
deepseek-v3	1	1	13	4	4.30
claude-3-5-haiku	4	9	6	6	5.55
gpt-4o-mini	6	5	4	3	5.55
gemini-2.0-flash	5	14	2	2	6.30
gemini-1.5-flash	10	16	1	1	8.45
gpt-4o	3	7	7	14	8.80
o1-mini	8	6	11	9	9.00
llama-3.2-3b	12	10	12	7	10.65
gpt-4-turbo	7	11	9	15	11.30
gemini-1.5-pro	9	17	5	12	11.55
claude-3-5-sonnet	11	12	8	11	11.70
o1	14	4	16	13	11.95
claude-3-opus	13	13	10	16	13.25
falcon-7b	15	3	15	17	13.40
mistral-7b	16	8	17	10	14.25
deepseek-7b	17	15	14	8	14.70

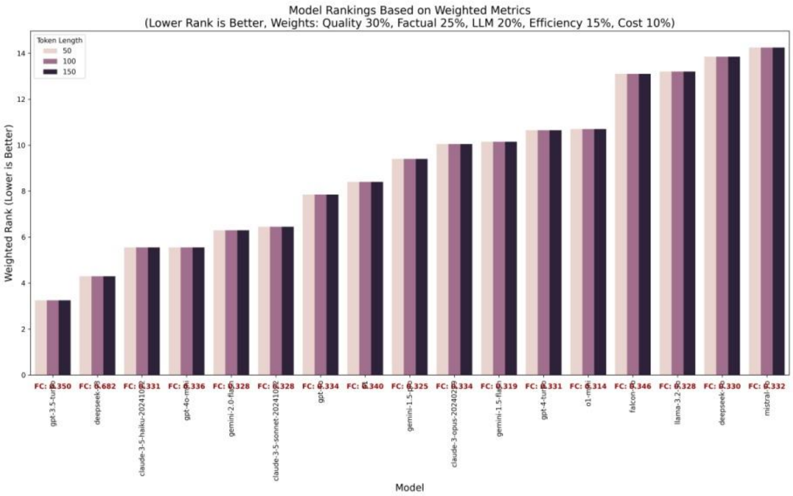


Figure 5: Models ranked by weighted score (lower is better)

D. Factual Consistency Rankings

Figure 6 illustrates the ranking of models by factual consistency scores, highlighting the exceptional performance of deepseek-v3.

E. LLM Evaluation Rankings

Figure 7 shows models ranked by human-like LLM evaluation scores, providing insight into perceived quality.

Table 14: Detailed quality metrics across all models

Model	ROUGE-1 F1	ROUGE-2 F1	BERTScore F1	SummaC	LLM Score	Time (s)
gpt-3.5-turbo	0.251	0.089	0.860	0.350	4.58	1.90
deepseek-v3	0.255	0.092	0.863	0.682	4.55	21.03
claude-3-5-haiku	0.249	0.088	0.858	0.318	4.70	3.27
gpt-4o-mini	0.247	0.087	0.856	0.336	4.60	2.41
gemini-2.0-flash	0.248	0.087	0.857	0.295	4.66	1.14
gemini-1.5-flash	0.242	0.083	0.853	0.289	4.54	1.08
gpt-4o	0.250	0.089	0.859	0.325	4.63	3.60
o1-mini	0.246	0.085	0.855	0.334	4.72	7.26
llama-3.2-3b	0.241	0.082	0.852	0.317	4.45	12.72
gpt-4-turbo	0.247	0.086	0.856	0.310	4.62	4.37
gemini-1.5-pro	0.245	0.084	0.854	0.284	4.57	2.55
claude-3-5-sonnet	0.241	0.083	0.852	0.307	4.75	4.16
o1	0.240	0.081	0.851	0.340	4.65	23.28
claude-3-opus	0.241	0.082	0.852	0.303	4.60	5.04
falcon-7b	0.239	0.080	0.848	0.346	4.42	22.35
mistral-7b	0.238	0.079	0.847	0.322	4.36	23.93
deepseek-7b	0.237	0.078	0.846	0.292	4.32	22.20

F. Cost-Quality Tradeoff Multidimensional Visualization

Figure 8 shows the multidimensional trade-off between quality aspects and cost across models, helping to identify the most cost-effective options for different budget constraints.

G. Cost-Effectiveness Rankings

Figure 9 presents models ranked by cost-effectiveness, identifying options that provide the best value for money.

H. Dataset-Specific Optimal Model Configuration

Table 15 identifies the best-performing model and token length configuration for each dataset, providing guidance for domain-specific applications.

Table 15: Optimal model and token length configuration by dataset

Dataset	Best Model	Optimal Length	Token
CNN/DailyMail	gpt-3.5-turbo	100	
XSum	deepseek-v3	50	
BigPatent	deepseek-v3	50	
BillSum	gpt-3.5-turbo	100	
PubMed	deepseek-v3	50	
SAMSum	claude-3-5-haiku	100	
WikiHow	gemini-2.0-flash	100	

I. Quality-Efficiency Trade-off by Use Case

Table 16 summarizes recommended model configurations for different use cases, balancing quality and efficiency requirements.

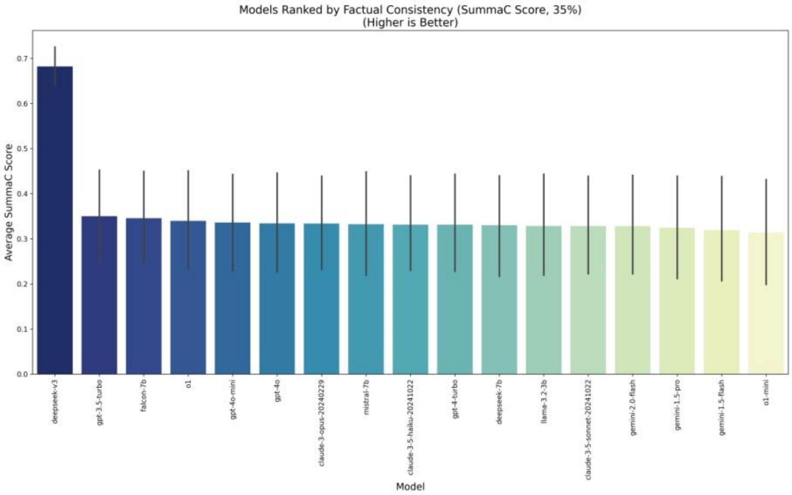


Figure 6: Models ranked by factual consistency

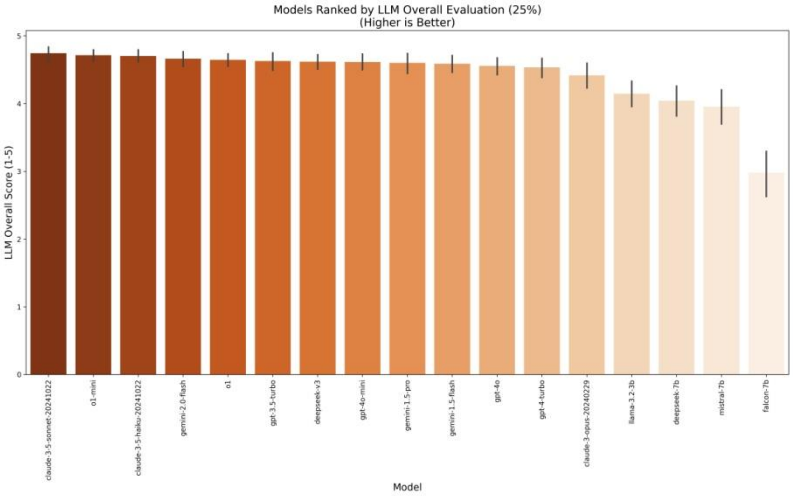


Figure 7: Models ranked by LLM evaluation scores

J. Dataset-Specific Performance Visualizations

Figures 10-16 illustrate the performance of models across different datasets, highlighting the significant performance variations by domain. These visualizations show how models perform differently on various text types and styles.

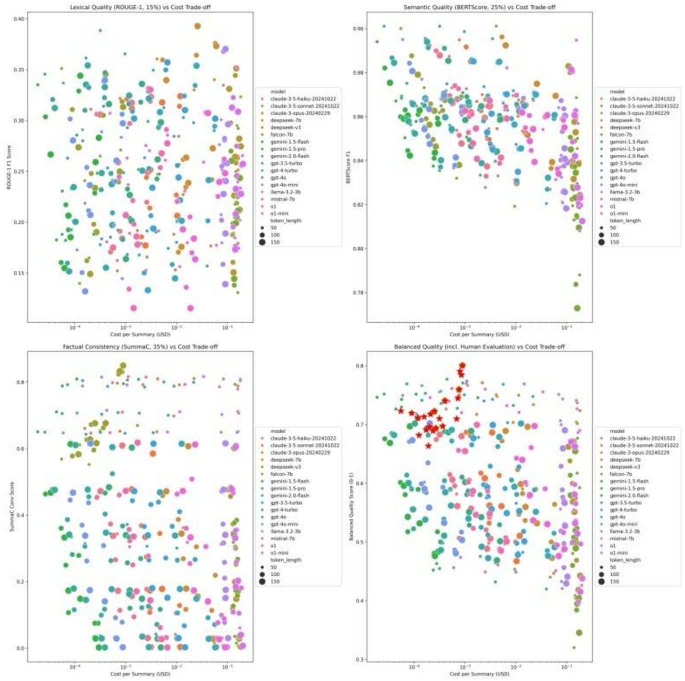


Figure 8: Multidimensional cost-quality trade-off across models

Table 16: Recommended model configurations by use case

Use Case	**Recommended Model**	**Token Length**
High-stakes	deepseek-v3	50
Human-like quality	claude-3-5-sonnet	150
General-purpose	gpt-3.5-turbo	100
Resource-constrained	gemini-1.5-flash	50-100
Cost-effective	gemini-1.5-flash	100

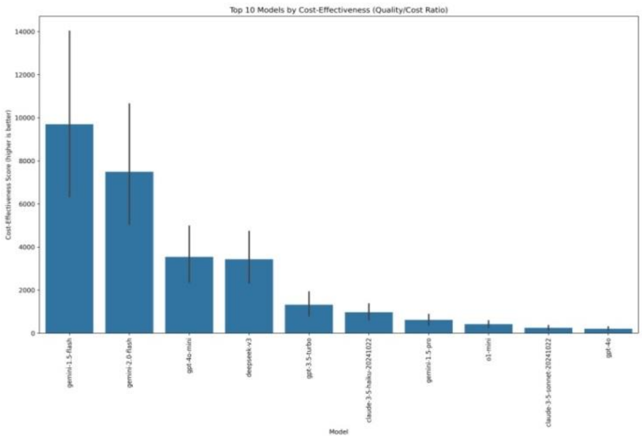


Figure 9: Models ranked by cost-effectiveness (quality per dollar)

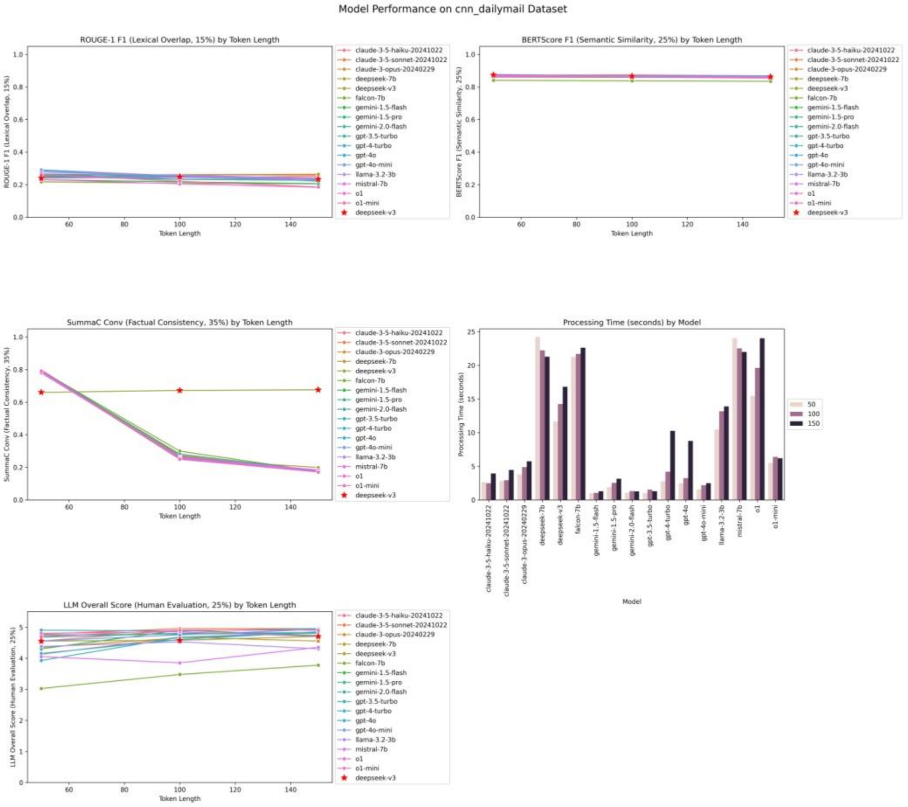


Figure 10: CNN/DailyMail dataset performance across metrics and models

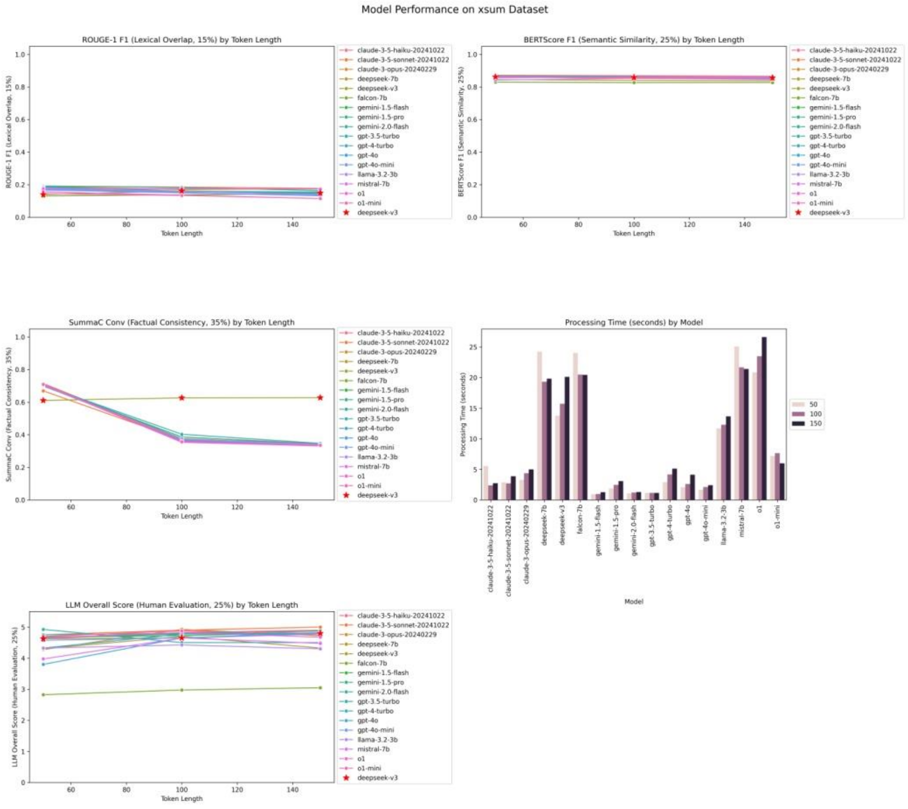


Figure 11: XSum dataset performance across metrics and models

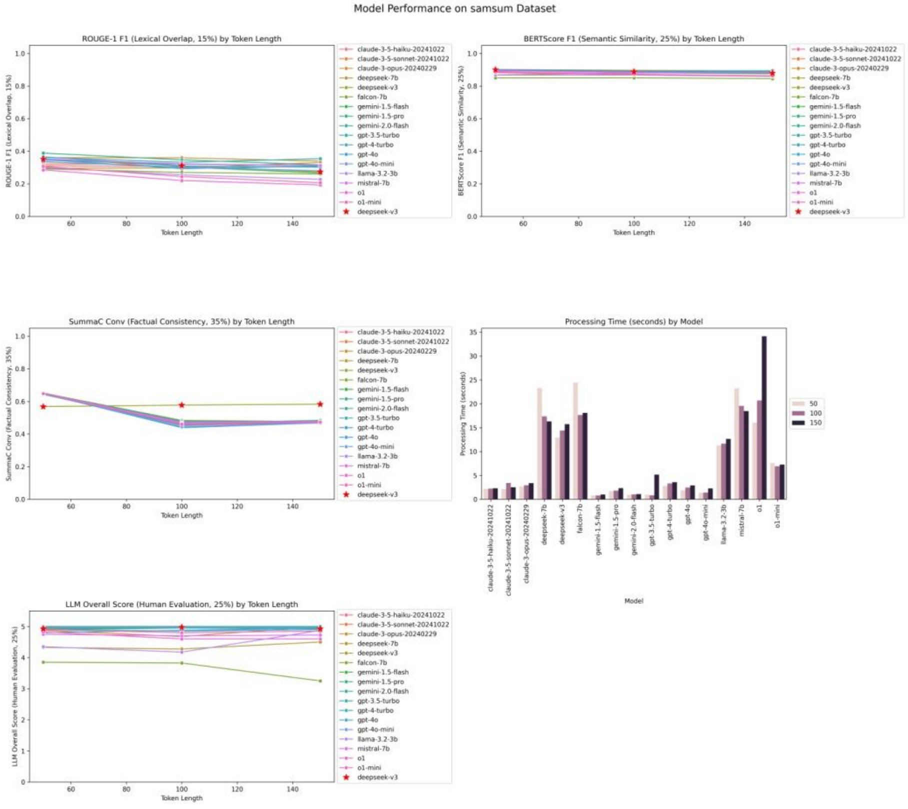


Figure 12: SAMSum dataset performance across metrics and models

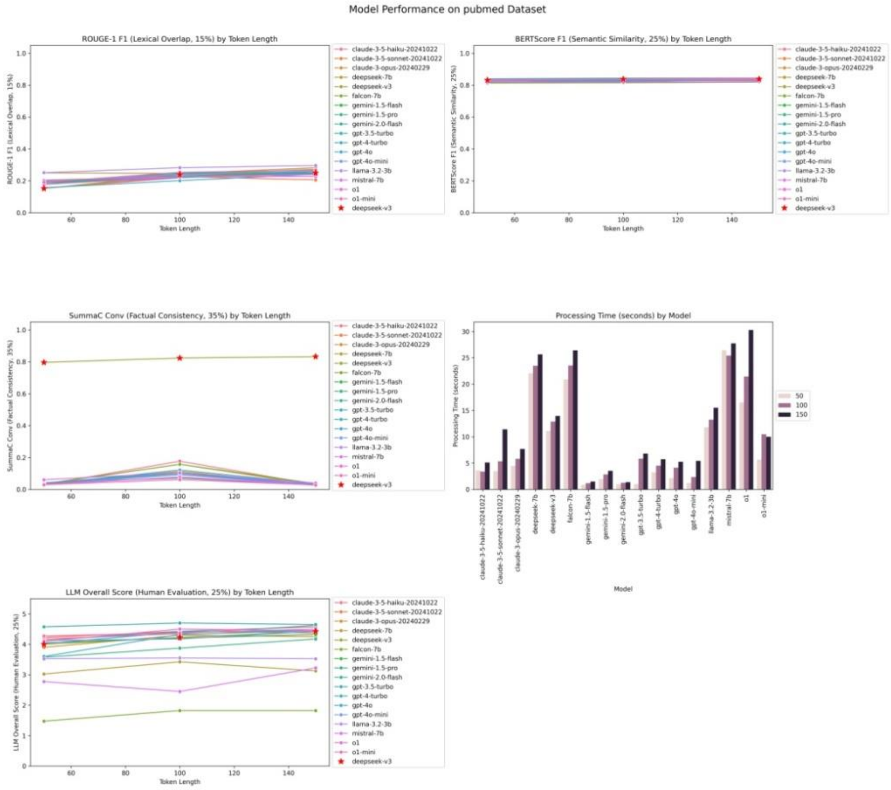


Figure 13: PubMed dataset performance across metrics and models

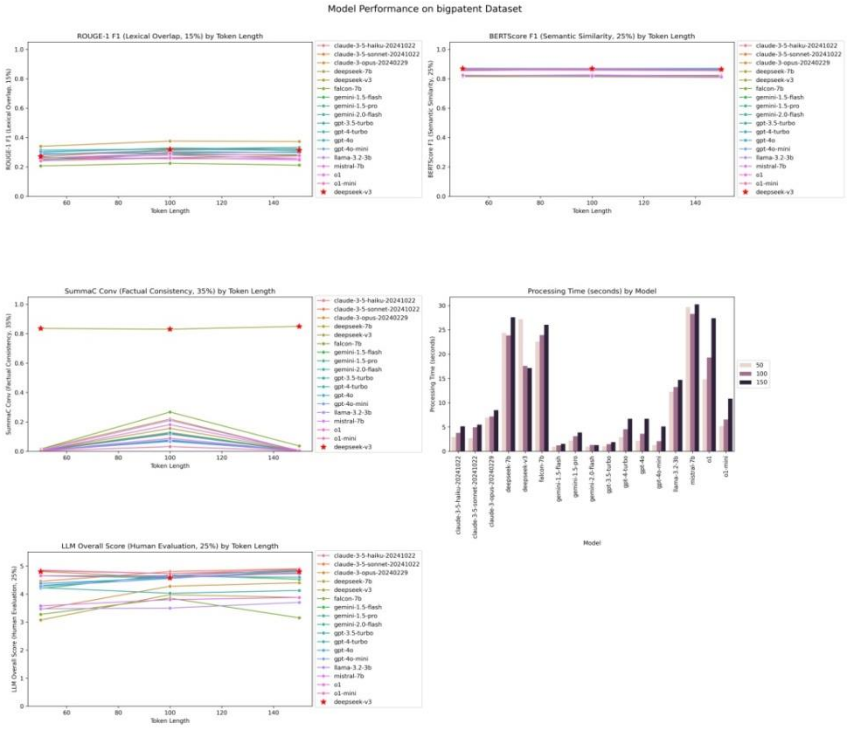


Figure 14: BigPatent dataset performance across metrics and models

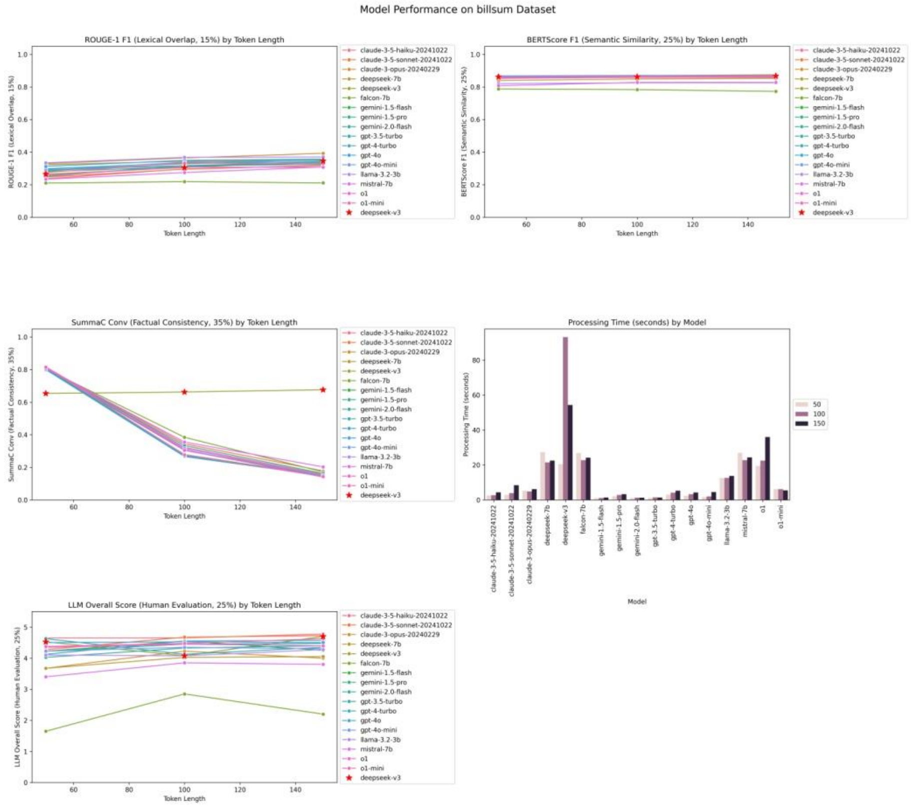


Figure 15: BillSum dataset performance across metrics and models

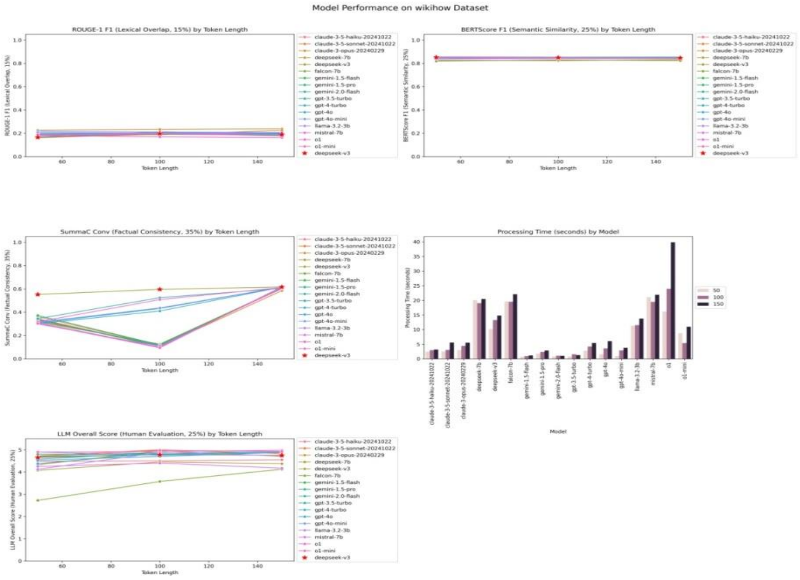


Figure 16: WikiHow dataset performance across metrics and models