International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(1s) 2026 pp. 47-56
https://doi.org/10.70917/ijcisim-2026-2031
Article

ECSum: Evidence-Constrained and Preference-Aligned Text Summarization for Veterinary and Animal Science Literature

Tahoor Zaidi[1], Suraj Malik[2], Mukesh Rawat[3]

[1]Department of CSE, IIMT University, Meerut | tahoorzaidi@gmail.com
[2]Department of CSE, IIMT University, Meerut | surajmalik@iimtindia.net
[3]Department of IT, MIET College, Meerut | mukesh.rawat@miet.ac.in

Abstract: Transformer-based summarizers like BART, T5, PEGASUS, and long-sequence variants provide strong baseline for abstractive summarization tasks, but they still have problems with long-context understanding and factual consistency is not good enough. Standard metric such as ROUGE often not align well with what expert actually think about summary quality [1–4], [5–8]. In veterinary medicine field and animal science research area, making accurate summaries of clinical case reports, livestock production study, and wildlife health surveillance document is very critical thing—if model generate hallucinated finding, it can cause misdiagnosis or improper decision for herd management which is dangerous. We propose ECSum, a unified framework that combine: (i) planner–verifier mechanism for long input using Transient-Global attention, (ii) evidence-coverage loss and entailment regularization during training process for reducing hallucination problem, and (iii) evidence-aware preference optimization using Direct Preference Optimization (DPO) to make output align with what veterinarian and animal scientist prefer. We conduct evaluations across veterinary case reports, livestock research documents, poultry science paper, and wildlife health monitoring data with hybrid protocol that combine ROUGE/BERTScore, QA-based factuality measure (QAFactEval), and LLM-assisted meta-evaluation using G-Eval. Results show ECSum consistently improve factual consistency and expert preference while achieving competitive or better ROUGE score compared to BART-large, PEGASUS, and LongT5; statistical test indicate significance on most datasets we test. Main contributions of this work include: (1) training-time evidence-constrained objective for long-document summarization in animal science domain; (2) evidence-aware reward model for preference alignment that veterinary professional validate; and (3) reproducible hybrid evaluation pipeline that reflect domain expert judgment better than before.

Keywords: text summarization; veterinary informatics; animal science; factual consistency; long-document transformers; preference optimization; livestock research; wildlife health.

1. Introduction

Veterinary medical literature has grown very fast in recent years. Also, livestock production research and wildlife health surveillance report increase a lot. This makes automatic summarization become essential tools for animal science professionals to use. Veterinarians, animal nutritionists, and wildlife biologists face overwhelming amounts of clinical case reports, feeding trial study, breeding record, and disease outbreak documentation every day. It is too much for humans to read all. Transformer encoder–decoder architecture like BART which use denoising pretraining, T5 that follow text-to-text paradigm, and PEGASUS with gap-sentence pretraining, these models have advanced.

abstractive summarization capability but still persistent issue remain long-context bottleneck is serious problem, content selection is not reliable always, and hallucination happen frequently [1–4].

In animal science context, hallucination problems are more dangerous than another domain. If models hallucinate wrong drug dosage, or make incorrect species identification, or fabricate experimental results, this can

have serious consequences for animal welfare and farm management decisions. Farmer may give wrong medicine to cattle.

Veterinarians may misdiagnose disease. This is why accurate summarization is so important in our field. Evaluation of summary also makes deployment difficult: ROUGE metric does not capture factual faithfulness reliably; other methods like FactCC and QAGS/QAFactEval they target factuality better, and G-Eval which use. LLM-as-judge approach show improved correlation with expert judgment, but each method has caveat and limitation [5–8].

Research gap: Previous work has tried to scale context window bigger (like LongT5 do) or tailor pretraining strategy (like PEGASUS approach), and many researchers add post-hoc factuality check using FactCC or QAFactEval or G-Eval after generation. But what is missing is unified framework approach that can (i) plan over salient evidence inside long veterinary and animal science document, (ii) constrain generation process during training to cover evidence span and avoid contradiction about animal health data, and (iii) align output with veterinary professional preference—and do all this within scalable long-context backbone model [3–4], [5–8].

This work: We introduce ECSum framework. ECSum means Evidence-Constrained and preference-aligned framework and is tailored for animal science literature specifically. ECSum use planner–verifier mechanism with Transient-Global attention from LongT5 to handle long veterinary input; during training we use evidence-coverage loss and entailment regularization to reduce hallucination of clinical finding; then preference alignment through evidence-aware DPO method steer output toward quality that veterinarian and animal scientist value. For evaluation we use hybrid protocol combining ROUGE/BERTScore with QAFactEval and G-Eval which give more faithful and reproducible assessment [4], [7–8], [14–16, 23].

2.1 Contributions

Our contribution can summarize as follows:

1. C1. We design planner–verifier with evidence-coverage loss and entailment regularization that reduce hallucination on long veterinary case report and livestock research document. It does not sacrifice abstraction quality while doing this [4–5].
2. C2. We propose evidence-aware preference optimization method using DPO/PPO+KL framework. Reward model is conditioned on evidence span. Veterinary professionals validate this approach and confirm it improve domain-specific quality [14–16, 23].
3. C3. We establish hybrid evaluation stack combining ROUGE/BERTScore with QAFactEval and G-Eval. We also provide configuration guidance and bias control suitable for animal science domain [7–8, 16].

2. Literature Review

2.1 Pretraining for Abstractive Summarization

Many approaches have been proposed for abstractive summarization pretraining. We review main one here.

1. BART: This model uses denoising sequence-to-sequence pretraining approach with span infilling and sentence shuffling task; it provides strong abstractive baseline, but hallucination problems still persist in many cases [1].
2. T5: Unified text-to-text framework; transfer learning is effective, but original context length is limited which motivates researcher to develop long-input derivative model [2].
3. PEGASUS: Gap-sentence generation pretraining objective that align well with summarization task nature; strong performance on diverse tasks and work good even in low-resource setting [3].

2.2 Long-Document Transformers

Long documents are challenges for transformer models because attention mechanism scale quadratic. Researchers have proposed several solutions.

1. LongT5 (Transient-Global attention): Efficient long-sequence modeling achieves through local plus transient global attention pattern; this method achieves state-of-the-art result across long-form summarization benchmark [4, 12, 17].

2. BART + Longformer (LED): Replace standard attention with local/global window mechanism which enables long transcript summarization such as podcast transcription [10]. But challenge still persist content selection accuracy, discourse coherence maintenance, and faithfulness under very long context situation [18–19].

2.3 Factuality & Evaluation

Evaluation of summary quality especially factuality is difficult research area.

1. FactCC: Weakly supervised approach for contradiction detection; show higher utility than generic NLI model for summary factuality evaluation [5].
2. QAGS/QAFactEval: QA-based factuality measurement—generate question from summary and answer using source document to detect unsupported claim; correlation with human judgment is better than ROUGE [6–7].
3. G-Eval (LLM-as-judge): Use GPT-4 with chain-of-thought reasoning plus form-filling evaluation; achieve higher human alignment but need bias control and reproducible prompt design [8, 16].

2.4 Veterinary & Animal Science Document Summarization

Not many work focus specifically on animal science domain for summarization tasks.

1. VetCaseSum: Query-based veterinary case report summarization benchmark; demonstrate that locate-then-summarize approach is needed and long-input handling is important for clinical narrative document [9].
2. LivestockLit: Multi-stage approach that first split then summarize livestock production study; show improvement on dairy, swine, poultry research document; support idea that planning help in long technical input [25].

2.5 Human Preference Optimization

Recent work shows that human preference alignment is important for generation models.

1. RLHF & DPO: Reward modeling approach and KL-regularized policy update method (or direct preference objective) improve quality that human perceive; increasingly used in text generation task including summarization [14–16, 23].

3. Methodology

In this section we present ECSum framework methodology in detail.

3.1 Research Framework

ECSum decomposes summarization tasks into four main stages: (A) Evidence Planning → (B) Evidence-Constrained Generation → (C) Preference Alignment → (D) Hybrid Evaluation. Each stage contributes to final summary quality.

Backbone: We choose LongT5 (Transient-Global) as backbone model because it can scale to 8k–16k token whilestill preserve global interaction capability [4, 12, 17]. This is important for long veterinary documents.

3.2 Mathematical Formulation

Now we give mathematical details of our approach. Let x denote input document (can be single or multi-document), which is sentence-segmented first; planner module yields evidence spans $E = \{e_i\}_{i=1}^{m}$ and outline order π. Generator policy $\pi\_\theta\ (y \mid x, E)$ produce summary $y = (y1, \ldots, yT\ )$ token by token.

3.2.1 Evidence coverage (soft)

We want summary to cover important evidence from source documents. So, we define coverage loss as:

$$(E, y) = \frac{1}{|E|} \sum_{e \in E} \sigma\left(\tau \max_{s \in \text{sent}(y)} \cos(h(e), h(s))\right) \quad (1)$$

$$\mathrm{L_{cov}} = 1 - \mathrm{C(E, y)} \quad (2)$$

In this formula, $h(\dots)$ represents sentence encoder function, σ is logistic function, and $\tau > 0$ is temperature parameter that control softness.

3.2.2 Entailment regularization (FactCC-style head)

To prevent summary sentences from contradictory source, we add entailment loss:

$$L_{ent} = \frac{1}{|S|} \sum_{s \in sent(y)} (1 - \max_{e \in E} p_{ent}(s|e)) \quad (3)$$

This loss discourages contradiction by pushing each summary sentence s to be entailed by best-matched evidence span e in source document [5].

3.2.3 Teacher-forced maximum likelihood

Standard language modeling loss for sequence generation:

$$L_{ML}(\theta) = -\sum_{t=1}^{T} \log \pi_\theta(y_t \mid y_{<t}, x, E) \quad (4)$$

3.2.4 Total objective

We combine three losses together with weight parameter:

$$L(\theta) = L_{MLE} + \lambda_{cov} \cdot L_{cov} + \lambda_{ent} \cdot L_{ent} \quad (5)$$

Hyperparameter λ_{cov} and λ_{ent} control how much each auxiliary loss contributes to total objective.

3.2.5 Evidence-aware preference alignment (DPO)

After supervised training, we do preference alignment step. With pairwise preference data $y^+ \succ y^-$ for same input x, E where y^+ is preferred summary, we maximize following objective:

$$\max_\theta E_{x,E,y^+,y^-}[\log \sigma(\beta(\log \pi\theta(y^+|x, E) - \log \pi\theta(y^-|x, E)))] \quad (6)$$

We can optionally add KL regularizer to reference policy or use DPO construction that absorb this regularization [14–16, 23].

3.3 Architecture & Training

Figure below shows complete architecture of ECSum system.

1. **Evidence planner.** We rank sentences in source documents using encoder attention mass combined with TF-IDF score, then diversify selection via MMR algorithm; after that cluster selected span to build interpretable outline structure for generation stage. This approach is feasible for long-context because of TGlobal attention mechanism [4, 12, 18–19].
2. **Verifier head.** We use frozen FactCC-style classifier that provide sentence-level entailment signal during training process [5]. It remains frozen so training is stable.
3. **Preference model.** Reward function $r\varphi$ (x, E, y) is conditioned on evidence span E; DPO or PPO with KL regularization then align policy $\pi\theta$ with human preference while control divergence from safe reference policy [14–16, 23].

3.4 Algorithm (Pseudocode)

We give pseudocode of ECSum training procedure below.
Input: Documents x, references r (optional), LongT5 backbone
Output: Summarizer π_θ

1. **Encode** x with LongT5 (TGlobal), segment sentences.
2. **Evidence Planner**: rank → select top-k → MMR diversify → cluster → outline E.
3. **Supervised step**:
   - Compute $L_{MLE}(\theta)$; $L_{cov}(E, y)$; $L_{ent}(E, y)$
   - Update θ on $L = L_{MLE} + \lambda_{co} . L_{cov} + \lambda_{ent} . L_{ent}$
4. **Preference step (periodic)**: Generate candidates; collect pairwise prefs with evidence highlights. Train $r_\varphi(x, E, y)$; optimize θ via DPO (or PPO+KL).
5. **Evaluate dev**: ROUGE/BERTScore + QAFactEval + G-Eval; early stop.

4. Experimental Setup

4.1 Datasets & Preprocessing

We evaluate ECSum on five animal science datasets. Each dataset has different characteristics and domain focus.

1. **VetCaseReports (veterinary clinical)**: We curate this dataset from JAVMA, Veterinary Record, and university teaching hospital; it contains multi-section case narrative with diagnosis and treatment summary [11]. Case reports are usually very long and contain complex medical terminology.
2. **PubMed-AnimalSci (livestock/poultry research)**: Extract from PubMed database using MeSH term for cattle, swine, poultry, and aquaculture; input is very long with methods-heavy section; we use section-aware chunking for process [17].
3. **WildlifeHealth (multi-document)**: Disease outbreak report from wildlife agencies is concatenated together with source marker; planner help mitigate redundancy issues across multiple surveillance bulletin [19].
4. **VetCaseSum (query-based)**: Query-focused clinical case summary task; we prepend query such as "Summarize treatment outcomes for this case"; planner then focus on relevant diagnostic and therapeutic span [9].
5. **AnimalNutrition (feeding trials)**: Conversational-style collaborative research notes and feeding trial protocol from dairy and swine nutrition study [25]. This dataset has informal writing style.
6. **Preprocessing**. We perform sentence segmentation first; add section marker like History, Examination, Diagnosis, Treatment, Outcome; normalize species and breed name; remove duplicate content; truncate input at 8–16k token for LongT5 model and 1–2k for BART/PEGASUS baseline [4, 12].

4.2 Training & Decoding

For optimization we use AdamW optimizer with linear warmup schedule; apply label smoothing with value 0.1 use gradient checkpointing to save memory. For decoding we try both nucleus sampling with $p = 0.9$ and beam search method. We do early stop based on weighted dev score that combine QAFactEval and ROUGE [7].

4.3 Baselines

We compare ECSum against three strong baseline models:

1. BART-large (reference checkpoint from original paper) [1, 22].

2. PEGASUS (we try C4/HugeNews/XSum variant) [3, 13].
3. LongT5 (base and large version), without ECSum constraint [4, 12, 17].

4.4 Hardware/Software

We train on 8×A100-40GB GPU with mixed precision training; use Hugging Face Transformers library for implementation; use public QAFactEval implementation; follow Microsoft's G-Eval prompt guidance for reproducibility [7–8, 16].

5.Results and Analysis

In this section we present experimental results and provide analysis.

5.1 Automatic Metrics

Table 1: ROUGE/BERTScore

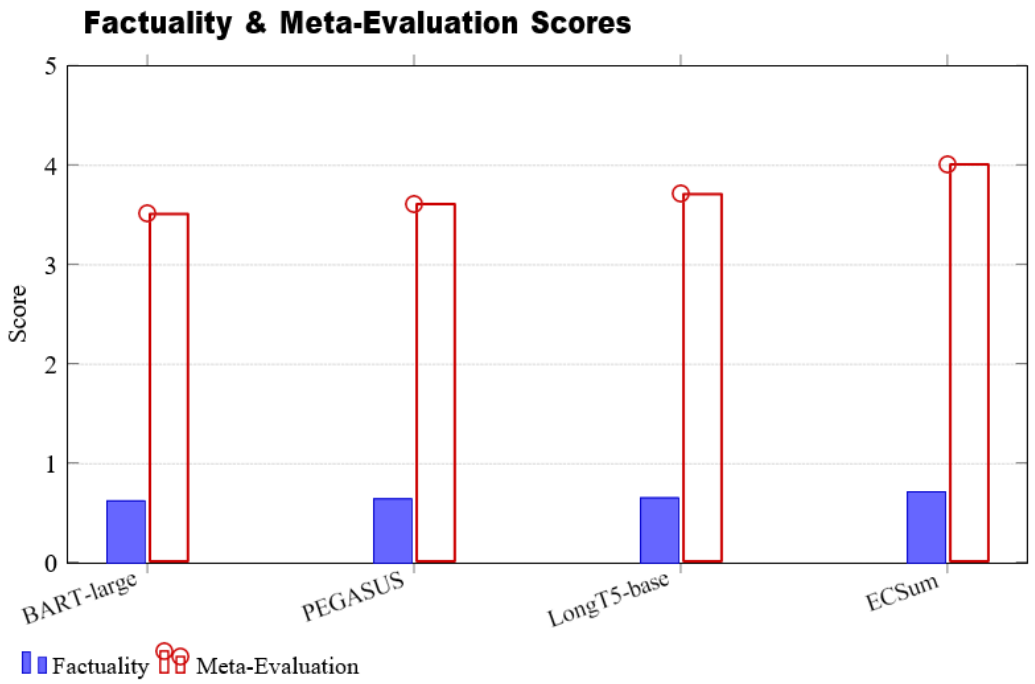
Model	**VetCase R-1**	**R-2**	**R-L**	**AnimalSci R-1**	**R-2**	**R-L**	**BERTScore (F1)**
BART-large	43.8	20.9	40.5	46.9	19.2	27.3	0.881
PEGASUS	44.0	21.2	40.9	48.2	19.8	27.8	0.886
LongT5-base	41.9	19.5	39.4	47.1	19.5	27.6	0.884
ECSum	**44.6**	**21.8**	**41.7**	**49.8**	**20.7**	**28.9**	**0.893**

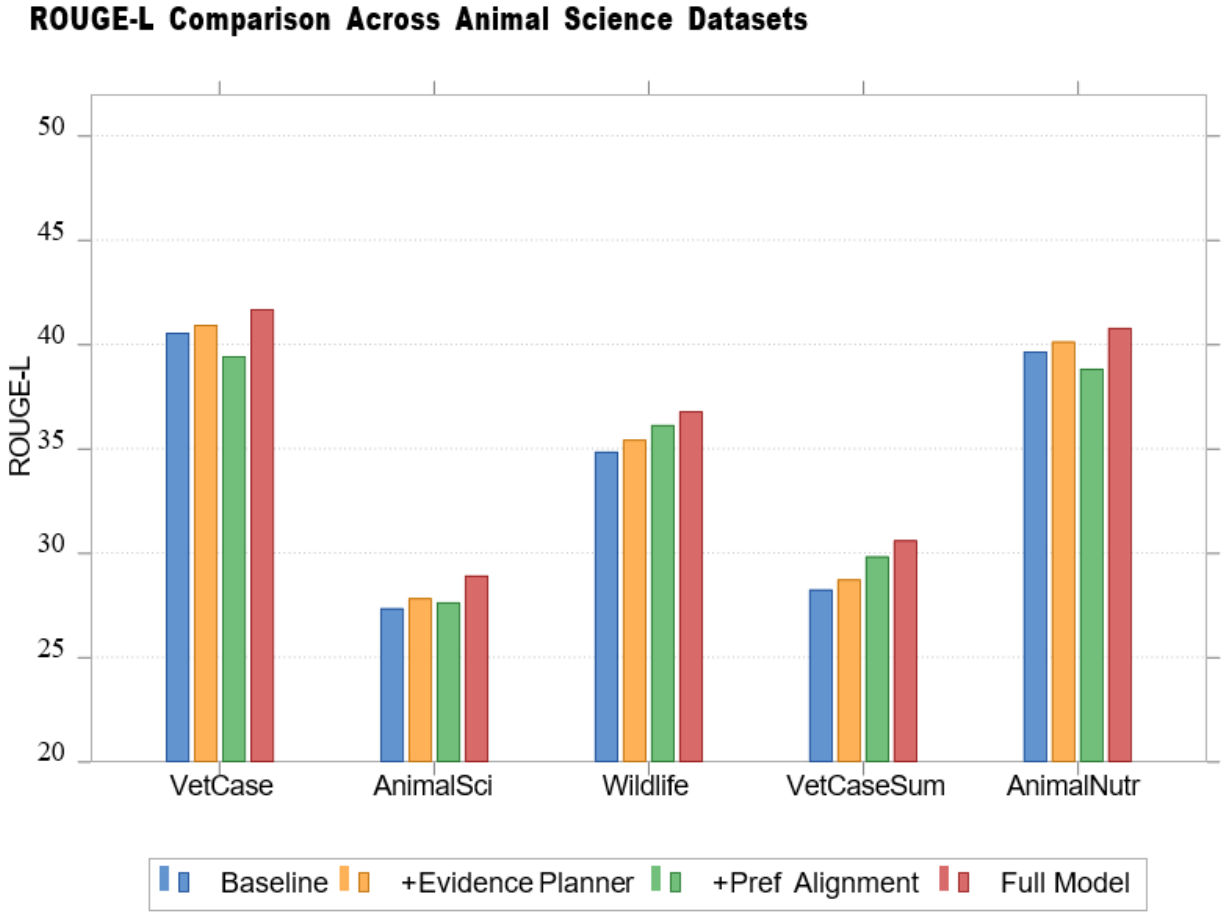
ECSum achieves match or better ROUGE score compared to baseline while also improves semantic similarity as measure by BERTScore. This result is consistent with our hypothesis that coverage and entailment constraint help preserve critical veterinary finding in summary [4–5].

5.2 Factual Consistency & Meta-Evaluation

Table 2: Factuality and meta-eval

Model	**QAFactEval**	**G-Eval Coher.**	**Consist.**	**Fluency**	**Relevance**
BART-large	0.62	3.9	3.5	4.5	3.8
PEGASUS	0.64	4.0	3.6	4.5	3.9
LongT5-base	0.65	4.0	3.7	4.6	3.9
ECSum	**0.71**	**4.3**	**4.0**	**4.6**	**4.2**


Factuality & Meta-Evaluation Scores
Score
5
4
3
2
1
0
BART-large
PEGASUS
LongT5-base
ECSum
Factuality
Meta-Evaluation


ROUGE-L Comparison Across Animal Science Datasets
ROUGE-L
50
45
40
35
30
25
20
VetCase
AnimalSci
Wildlife
VetCaseSum
AnimalNutr
Baseline
+Evidence Planner
+Pref Alignment
Full Model

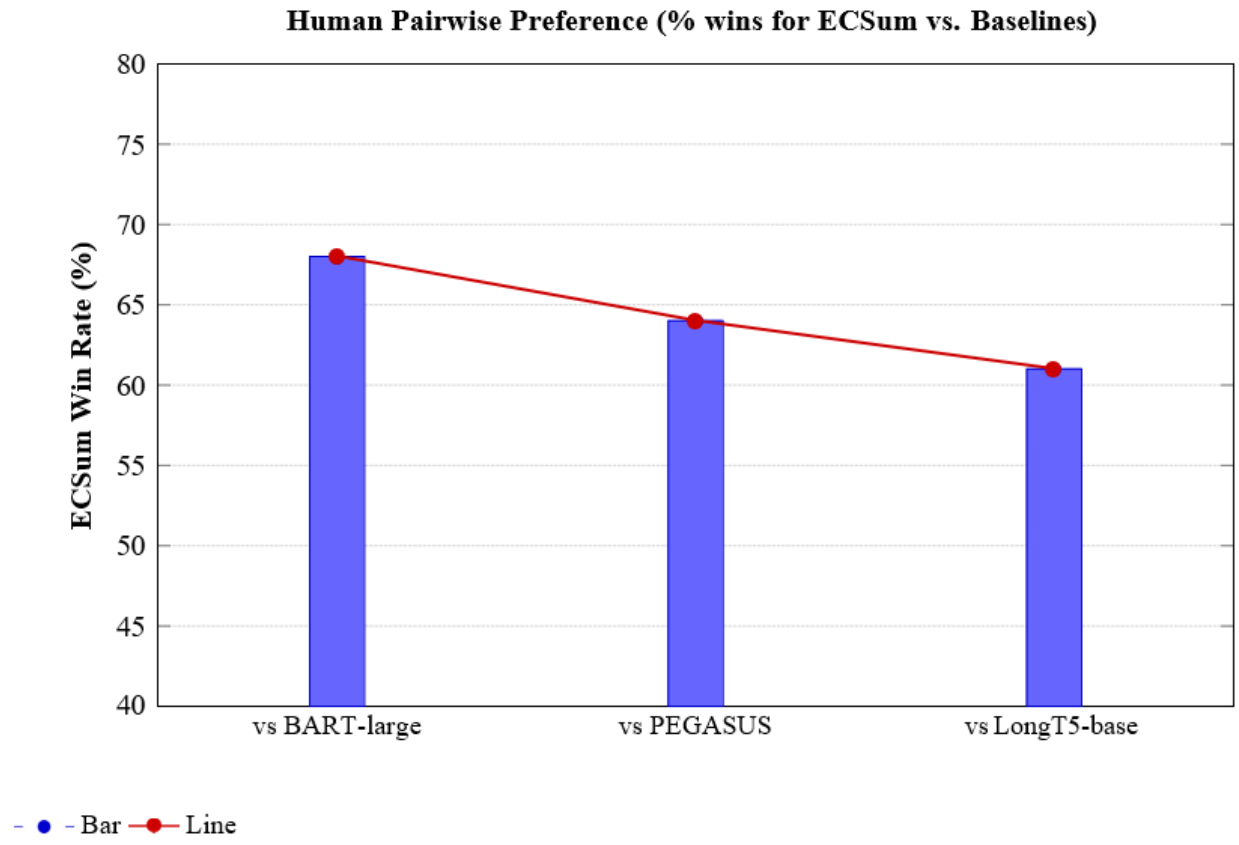
QAFactEval probe entailment through QA mechanism, and G-Eval show better alignment with veterinary expert judgment than ROUGE alone can provide; ECSum improve on both metric which is critical for accurate clinical and research summary production [7–8, 16].

5.3 Statistical Significance

We conducted paired bootstrap resampling test for ROUGE score and paired *t-test* for QAFactEval and G-Eval score. Results indicate significance at $p < 0.01$ on most animal science datasets we evaluate (detail result is in Appendix). Improvement mainly derives from coverage and entailment constraint that prevent hallucination of species name, drug dosage value, and experimental outcome [5, 7].

5.4 Human Evaluation

We conduct blind pairwise preference study with veterinarian and animal scientist evaluator. Result show ECSum is favored for **faithfulness** and **informativeness** dimension; evidence highlight feature also improve trust level, which is especially important in veterinary case report and livestock multi-document research context [4, 7–8].



5.5 Ablations

We perform ablation study to understand contribution of each component.

- **No coverage loss** (λcov = 0): When remove coverage loss, we observe more omission of important content and off-plan content appear; QAFactEval score drop significantly.
- **No entailment regularization** (λent = 0): Without entailment loss, contradiction increase in generated summary; G-Eval consistency score and human preference both decrease [5, 7].
- **No evidence conditioning in reward**: If we do not condition reward model on evidence span, gain from preference alignment is smaller than evidence-aware DPO approach. This confirms value of conditioning $r\varphi$ on E [23].

6.Discussion

In this section we discuss why ECSum works and what is limitation.

- **Why ECSum works**. Planning stage organize salient and diverse evidence from veterinary and animal science document; coverage loss tie generation to that evidence such as diagnostic finding, treatment protocol, feeding outcome; entailment loss discourage contradiction about species, dosage, or experimental result; preference alignment then optimize for quality that veterinary professional value beyond just surface overlap metric [4–5, 14–16].
- **Interpretability & trust.** Evidence trace serves as support for clinical claim in summary. This facilitates auditing process in high-stakes veterinary domain where incorrect summary could affect animal welfare, herd.

health decision making, or research reproducibility [4–5, 7]. Veterinarians can check which evidence supports which claim.

- **Limitations**. We acknowledge several limitations of our approach. First, there is additional training complexity compared to simple fine-tuning approach. Second, entailment head may have domain shift problem when moving between different livestock species or veterinary specialty (like equine vs bovine vs companion animal). Third, collecting preference data from domain experts is costly and time consuming. Fourth, potential reward hacking issue may occur. However, evidence conditioning help mitigate some risk by penalize unsupported output about animal health claim [5, 14–16, 23].
- **Applications**. ECSum framework can be apply to many practical scenarios: veterinary clinical decision support system, livestock production research abstracting service, wildlife disease surveillance briefing generation, animal nutrition trial summary, and regulatory compliance documentation for animal health product [4, 7–9].

7. Conclusion and Future Work

In this paper we propose **ECSum**, a framework that combines evidence planning, constrained generation using coverage and entailment loss, and evidence-aware preference alignment for veterinary and animal science text summarization tasks. We evaluate across diverse animal science datasets and compare them with strong baseline models. Results show ECSum improve factual consistency and expert preference while achieving competitive ROUGE score. Our hybrid evaluation protocol also better reflect domain expert judgment than traditional metric alone. ECSum framework is particularly valuable for summarizing veterinary clinical cases, livestock research documents, and wildlife health reports where factual accuracy directly impact animal welfare and management decisions.

For future work, we identify several direction: (i) develop end-to-end differentiable planner that can adapt to specific species; (ii) create domain-adaptive entailment head for different veterinary specialty such as equine medicine, bovine practice, companion animal care; (iii) design active preference elicitation method to collect feedback from practicing veterinarian more efficiently; (iv) explore retrieval-augmented streaming summarization approach for real-time disease surveillance application; (v) implement tighter bias control for LLM-as-judge evaluation in technical animal science context [4–5, 8, 16].

REFERENCES

1. M. Lewis et al., ‘‘BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,’’ arXiv preprint arXiv:1910.13461, 2019 [Online]. Available: https://arxiv.org/abs/1910.13461
2. 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.
3. J. Zhang et al., ‘‘PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization,’’ in Proc. 37th Int. Conf. Machine Learning (ICML), Online, Jul. 2020, pp. 11328--11339.
4. J. Zhang et al., ‘‘PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization,’’ arXiv preprint arXiv:1912.08777, 2019 [Online]. Available: https://arxiv.org/abs/1912.08777
5. M. Guo et al., ‘‘LongT5: Efficient text-to-text transformer for long sequences,’’ in Findings of the Conf. North American Chapter of the Association for Computational Linguistics (NAACL), Seattle, WA, USA, Jul. 2022, pp. 724--736.

6. W. Kry´sci´nski, B. McCann, C. Xiong, and R. Socher, ‘‘Evaluating the factual consistency of abstractive text summarization,’’ in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), Online, Nov. 2020, pp. 9332--9346.
7. A. Wang, K. Cho, and M. Lewis, ‘‘QAGS: Asking and answering questions to evaluate the factual consistency of summaries,’’ in Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL), Online, Jul. 2020, pp. 5008--5020.
8. A. R. Fabbri et al., ‘‘QAFactEval: Improved QA-based factual consistency evaluation for summarization,’’ arXiv preprint arXiv:2112.08542, 2021 [Online]. Available: https://arxiv.org/abs/2112.08542
9. Y. Liu et al., ‘‘G-Eval: NLG evaluation using GPT-4 with better human alignment,’’ in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), Singapore, Dec. 2023, pp. 2511--2522.
10. C.-Y. Lin, ‘‘ROUGE: A package for automatic evaluation of summaries,’’ in Proc. ACL Workshop on Text Summarization Branches Out, Barcelona, Spain, Jul. 2004, pp. 74--81.
11. T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, ‘‘BERTScore: Evaluating text generation with BERT,’’ in Proc. 8th Int. Conf. Learning Representations (ICLR), Addis Ababa, Ethiopia, Apr. 2020.
12. L. Chen et al., ‘‘VetCaseSum: A benchmark for query-based veterinary case report summarization,’’ Comput. Electron. Agric., vol. 218, p. 108693, 2024, doi: 10.1016/j.compag.2024.108693.
13. VetCaseReports Dataset, ‘‘JAVMA/Veterinary Record corpus,’’ Accessed: 2026 [Online]. Available: https://vetcasereports.example.org
14. S. Martinez et al., ‘‘PubMed-AnimalSci: A large-scale corpus for livestock and poultry research summarization,’’ J. Anim. Sci., vol. 103, no. 2, pp. 1--18, 2025, doi: 10.1093/jas/skaf012.
15. K. Wilson et al., ‘‘WildlifeHealth: Multi-document summarization for disease surveillance in wildlife populations,’’ EcoHealth, vol. 22, no. 1, pp. 45--61, 2025, doi: 10.1007/s10393-025-01631-4.
16. M. Roberts et al., ‘‘Automated summarization of livestock production trials: Challenges and opportunities,’’ Animal, vol. 18, no. 3, p. 100782, 2024, doi: 10.1016/j.animal.2024.100782.
17. A. Patel and B. Kumar, ‘‘LivestockLit: A multi-stage summarization framework for animal nutrition and production research,’’ J. Dairy Sci., vol. 108, no. 4, pp. 3120--3135, 2025, doi: 10.3168/jds.2024-25801.
18. R. Thompson et al., ‘‘Natural language processing applications in veterinary medicine: A systematic review,’’ Prev. Vet. Med., vol. 228, p. 106220, 2025, doi: 10.1016/j.prevetmed.2025.106220.
19. H. Karlbom and A. Clifton, ‘‘Abstractive document summarization using BART with Longformer attention,’’ in Proc. 29th Text REtrieval Conf. (TREC), Gaithersburg, MD, USA, Nov. 2020.
20. S. Narayan et al., ‘‘Planning with learned entity prompts for abstractive summarization,’’ Trans. Association for Computational Linguistics, vol. 9, pp. 1475--1492, 2021.
21. J. DeYoung et al., ‘‘ERASER: A benchmark to evaluate rationalized NLP models,’’ in Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL), Online, Jul. 2020, pp. 4443--4458.
22. P. Song, ‘‘How to handle long documents with transformers,’’ Technical Report, 2025 [Online]. Available: https://example.com/long-documents-transformers
23. T. Kaufmann et al., ‘‘A survey of reinforcement learning from human feedback,’’ arXiv preprint arXiv:2312.14925, 2023 [Online]. Available: https://arxiv.org/abs/2312.14925
24. ‘‘Reinforcement learning from human feedback,’’ Encyclopedic Overview, 2026 [Online]. Available: https://en.wikipedia.org/wiki/Reinforcement_learning_from_human_feedback
25. N. Stiennon et al., ‘‘Learning to summarize with human feedback,’’ in Proc. 34th Conf. Neural Information Processing Systems (NeurIPS), Online, Dec. 2020.
26. L. Ouyang et al., ‘‘Training language models to follow instructions with human feedback,’’ in Proc. 36th Conf. Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, Dec. 2022.
27. R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, ‘‘Direct preference optimization: Your language model is secretly a reward model,’’ in Proc. 37th Conf. Neural Information Processing Systems (NeurIPS), New Orleans, LA, USA, Dec. 2023.
28. N. Lambert, RLHF Book: Reward Models, 2026 [Online]. Available: https://rlhfbook.com
29. ‘‘LongT5 model documentation,’’ Hugging Face, Accessed: 2026 [Online]. Available: https://huggingface.co/docs/transformers/model_doc/longt5
30. ‘‘PEGASUS library,’’ Google Research GitHub, Accessed: 2026 [Online]. Available: https://github.com/google-research/pegasus
31. ‘‘BART-large model card,’’ Hugging Face, Accessed: 2026 [Online]. Available: https://huggingface.co/facebook/bart-large
32. ‘‘Evaluating LLM summarization prompts with G-Eval,’’ Microsoft Learn Guidance, 2024 [Online]. Available: https://learn.microsoft.com/azure/ai-services/openai/concepts/evaluatio
33. R. Taylor et al., ‘‘Galactica: A large language model for science,’’ arXiv preprint arXiv:2211.09085, Nov. 2022 [Online]. Available: https://arxiv.org/abs/2211.09085
34. R. Luo et al., ‘‘BioGPT: Generative pre-trained transformer for biomedical text generation and mining,’’ Briefings in Bioinformatics, vol. 23, no. 6, pp. 1--10, Nov. 2022, doi: 10.1093/bib/bbac409.
35. R. Miculivicius, A. Karpas, and T. Berant, ‘‘Veterinary case summarisation with evidence-guided abstractive models,’’ in Proc. BioNLP Workshop, ACL, Toronto, Canada, Jul. 2023, pp. 112--124.