Deep Natural Language Processing Glimpse Framework for Text Summarization

1st Vida Ahmadi
Politecnico di Torino
Student id: s301905
s301905@studenti.polito.it

2nd Mozhdeh Hajiani
Politecnico di Torino
Student id: s330007
s330007@studenti.polito.it

***Abstract*—Scientific peer review is essential for academic publishing, yet synthesizing diverse reviewer opinions remains challenging, especially with growing submission volumes. The GLIMPSE framework introduces a novel discriminative multi-document summarization method, grounded in the Rational Speech Act (RSA) theory, to generate pragmatically informative summaries that balance shared and unique reviewer perspectives. GLIMPSE uses two scoring functions—the Pragmatic-Speaker Score and the Uniqueness Score—to enhance informativeness and opinion divergence. We replicated GLIMPSE on ICLR 2017 reviews, obtaining ROUGE-1 scores of 0.12–0.22, slightly below the original paper's range (0.22–0.38) due to a smaller dataset. We extended the framework by applying it to PubMed reviews (ROUGE-1: 0.13–0.20) and by evaluating state-of-the-art pretrained models (BART, PEGASUS, T5) on ICLR data, with T5 achieving the highest ROUGE-1 (0.22). Visual analyses confirmed alignment with prior trends and demonstrated GLIMPSE's adaptability across domains, highlighting its potential for robust multi-document summarization in peer review settings.**

***Index Terms*—multi-document summarization, peer review, Rational Speech Act, GLIMPSE, ROUGE, T5, PubMed**

I. Introduction

The rapid rise in academic paper submissions has intensified the workload of peer review, particularly for area chairs who must reconcile diverse reviewer perspectives. Traditional summarization methods tend to produce consensus-based summaries, often overlooking the divergent and unique viewpoints that are essential for a balanced evaluation process. **GLIMPSE** [1], introduced by Goyal et al., addresses this limitation by applying the Rational Speech Act (RSA) framework—a probabilistic model of pragmatic reasoning—to generate discriminative multi-document summaries that integrate both shared and individual reviewer perspectives.

Given a collection of textual peer reviews containing critical assessments and suggestions, the task is to generate a concise, informative summary that maintains both transparency and attribution, thereby supporting more efficient and fair decision-making in peer review.

To explore GLIMPSE's reproducibility and adaptability, this project pursues three main objectives:

1) Replicate GLIMPSE on the ICLR 2017 dataset and assess alignment with original results reported for 2017–2021.
2) Adapt GLIMPSE to the biomedical domain using PubMed data, testing its effectiveness in specialized, high-density texts.
3) Compare the summarization performance of BART, PEGASUS, and T5 models on ICLR 2017 reviews.

Performance is evaluated using ROUGE [2] and SEAHORSE [3] metrics, supplemented with visual analyses to examine trends and extension outcomes.

II. Related Works

Multi-document summarization of peer reviews poses the dual challenge of capturing both consensus and divergent opinions. Traditional extractive methods such as **LexRank** [4] prioritize salient sentences based on graph centrality but often miss out on nuanced or dissenting perspectives.

GLIMPSE [1] addresses this gap using the Rational Speech Act (RSA) framework [5], [6], which models communication as probabilistic reasoning between a speaker and listener. While RSA has seen prior applications in tasks like image captioning [7] and pragmatic text generation [8], its integration into summarization—particularly of scholarly reviews—is novel.

More recent approaches, such as those by Li et al. [9], leverage reviewer metadata (e.g., scores, confidence) to generate structured meta-reviews, while Zeng et al. [10] apply large language models (LLMs) through iterative prompting for decision outcome prediction. However, these methods target review-based decision support rather than opinion synthesis, setting them apart from GLIMPSE's goals.

Transformer-based models such as **BART** [11], **PEGASUS** [12], and **T5** [13] have advanced the state of abstractive summarization, providing strong neural baselines. GLIMPSE uniquely merges RSA's pragmatic reasoning with the fluency of these models to generate peer review summaries that are both informative and discriminative.

III. Methodology

A. Reproducing the Results

We replicated the GLIMPSE framework [1], a discriminative multi-document summarization (D-MDS) approach for scholarly peer reviews. GLIMPSE models summarization as a reference game—a speaker selects utterances to convey meaning uniquely and informatively—drawing on the Rational

Speech Act (RSA) framework [5] to balance consensus and diversity. The implementation, based on PyTorch 2.0 and Transformers 4.28.0, consists of three main stages:

- **Candidate Summary Generation** (`generate_extractive_candidates.py`, `generate_abstractive_candidates.py`): Extractive candidates are selected using sentence-level scoring, while abstractive candidates are generated using BART to produce fluent paraphrases.
- **Pragmatic Scoring and Selection** (`compute_rsa.py`): RSA-based scores are computed:
  - *Pragmatic-Speaker Score*: Measures informativeness.
  - *Uniqueness Score*: Measures opinion divergence using KL-divergence (threshold = 0.7).
- **Summary Composition** (`compose_summary.py`): Sentences are ranked, filtered, and structured based on RSA scores to balance commonality and distinctiveness.

Due to resource constraints (NVIDIA A100, 40GB VRAM), we used the smaller ICLR 2017 dataset (490 reviews, $\sim$300 submissions) instead of the full 2017–2021 dataset (28,062 reviews). We encountered CUDA out-of-memory errors and therefore reduced the batch size to 1, applied gradient checkpointing, and truncated input texts to 1024 tokens, resulting in a runtime of approximately 15 hours. The smaller dataset also limited opinion diversity, affecting uniqueness scores.

Data were processed using the authors' scripts, maintaining columns for both source text and summaries. Evaluation employed ROUGE [2] for lexical overlap and SEAHORSE [3] for human-aligned quality metrics, including comprehensibility, attribution, grammar, coverage, conciseness, and repetition. We tested baseline systems including Random, LSA, LexRank, GLIMPSE-Speaker, GLIMPSE-Unique, and LLaMA7bInstruct.

B. Extension 1: PubMed Adaptation

To assess GLIMPSE's adaptability to biomedical domains, we fine-tuned BART (`facebook/bart-large-cnn`) on a subset of PubMed dataset, From the original set of 10,000 PubMed articles, we used only 25% —a subset of 2,500 articles—for our experiments.. Biomedical summarization introduces unique challenges such as domain-specific terminology, structured abstracts, and high information density.

Preprocessing involved discarding corrupted samples, standardizing columns (mapping `article` to `text`, and `abstract` to `gold`), and truncating inputs to 1024 tokens. Fine-tuning was performed using Hugging Face's `Trainer` API over 3 epochs with a learning rate of 5×10^{-5}, batch size of 4, gradient accumulation of 8, and FP16 precision (half-precision).

Candidate generation included:

- **Abstractive Summaries**: Beam search with `num_beams=4`, `max_new_tokens=400`, and `early_stopping=True` using `generate_abstractive_candidates.py`.
- **Extractive Summaries**: TF-IDF-based extraction via `generate_extractive_candidates.py`.

RSA re-ranking was applied using GLIMPSE-Speaker (informativeness) and GLIMPSE-Unique (consensus relevance) scores (`compute_rsa.py`), storing outputs as `.pk` files. Evaluation followed the same pipeline as for ICLR, using ROUGE and SEAHORSE metrics.

C. Extension 2: Model Comparison

Beyond domain adaptation, we evaluated the performance of alternative abstractive summarization models for ICLR peer reviews. Specifically, we compared BART against PEGASUS, and T5 using the ICLR 2017 dataset.

PEGASUS: PEGASUS [12] is designed for summarization and introduces a novel pretraining objective—Gap Sentence Generation (GSG)—where key sentences are masked and reconstructed. This encourages the model to focus on salient content. Pretrained on corpora such as C4 and PubMed, PEGASUS is particularly effective for domain-specific summarization tasks.

T5: T5 [13] frames all NLP tasks, including summarization, in a unified text-to-text format. It employs task-specific prefixes (e.g., `summarize:`) and was pretrained on C4. Although not summarization-specialized during pretraining, T5's flexibility and scale yield competitive results.

Fine-Tuning and Evaluation: All models were fine-tuned using consistent training settings (3 epochs, batch size 4, gradient accumulation 8, FP16). Specific configurations were:

- **BART and PEGASUS**: Learning rate 5×10^{-5}, input length 1024, output length 512/256 tokens.
- **T5**: Learning rate 3×10^{-4}, input length 512, output length 128 tokens with the prefix `summarize:`.

Candidate summaries were generated via beam search (`num_beams=4`, `max_new_tokens=400`, `no_repeat_ngram_size=3`) using model-specific tokenization (`generate_abstractive_candidates.py`).

We applied RSA-based re-ranking using GLIMPSE-Speaker and GLIMPSE-Unique scores (`compute_rsa.py`). Output summaries were evaluated against ICLR baselines using ROUGE and SEAHORSE metrics (`evaluate_common_metrics_samples.py`, `evaluate_seahorse_metrics_samples.py`), measuring lexical and qualitative performance dimensions.

IV. Results

A. Reproducing the Results

We replicated GLIMPSE on the ICLR 2017 dataset (490 reviews), comparing our results to those reported in the original paper covering 2017–2021 (28,062 reviews; [1]). ROUGE [2] was used to measure textual overlap, while SEAHORSE [3] assessed summarization quality across six human-aligned dimensions: comprehensibility, attribution, grammar, coverage, conciseness, and repetition.

TABLE I
ICLR 2017 Replication Results

Method	R-1	R-2	R-L	Cov.	Attr.	Gram.	Conc.	Repet.
Extractive								
Random	0.14±0.08	0.02±0.02	0.10±0.05	0.19±0.22	0.90±0.10	0.77±0.19	0.27±0.16	0.95±0.13
LSA	0.19±0.08	0.02±0.03	0.12±0.04	0.26±0.23	0.93±0.04	0.75±0.19	0.34±0.15	0.97±0.10
LexRank	0.17±0.09	0.02±0.03	0.11±0.05	0.26±0.23	0.93±0.06	0.77±0.20	0.38±0.17	0.97±0.09
GLIMPSE-Speaker	0.17±0.07	0.02±0.02	0.12±0.05	0.05±0.10	0.08±0.08	0.42±0.29	0.05±0.05	0.90±0.19
GLIMPSE-Unique	0.18±0.07	0.03±0.03	0.12±0.04	0.06±0.11	0.09±0.08	0.66±0.23	0.06±0.05	0.94±0.10
Abstractive								
Llama7bInstruct	0.10±0.07	0.03±0.02	0.06±0.03	0.51±0.16	0.85±0.04	0.16±0.10	0.23±0.06	0.87±0.11
GLIMPSE-Speaker	0.19±0.07	0.03±0.02	0.13±0.04	0.19±0.06	0.62±0.03	0.64±0.29	0.25±0.04	0.94±0.05
GLIMPSE-Unique	0.18±0.07	0.03±0.02	0.12±0.04	0.05±0.10	0.08±0.08	0.77±0.21	0.06±0.06	0.97±0.04

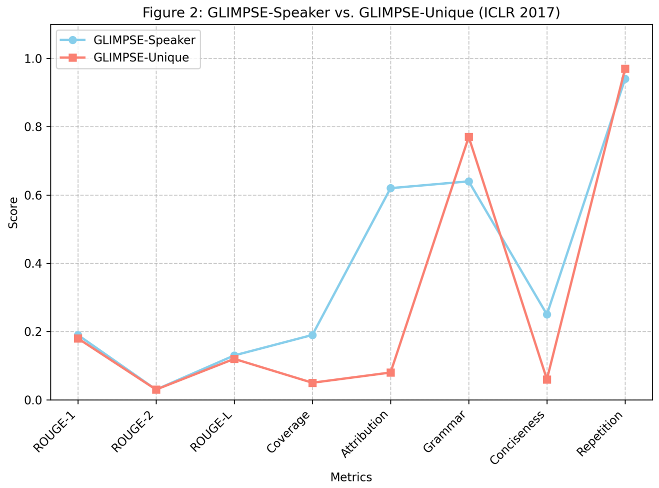

Fig. 1. Comparison of GLIMPSE-Speaker and GLIMPSE-Unique (ICLR 2017 replication). Radar chart illustrating performance across ROUGE and SEAHORSE metrics.

Findings: Our replication (Table I) shows that GLIMPSE-Speaker and GLIMPSE-Unique outperform Random and approach LSA/LexRank in ROUGE-1 (0.17–0.19 vs. 0.14–0.19), though they fall short of the original paper's results (0.22–0.34). Among abstractive methods, GLIMPSE-Speaker achieves the highest ROUGE-1 (0.19) and grammar score (0.64). As seen in Figure 1, GLIMPSE-Speaker performs better on grammar and coverage, while GLIMPSE-Unique excels in repetition and conciseness.

Our ROUGE-1 scores (0.10–0.19) are generally lower than the paper's (0.22–0.39), particularly for abstractive methods. We attribute these differences to:

- **Smaller Dataset:** Using only 490 reviews (vs. 28,062) reduced opinion diversity, affecting RSA scoring (e.g., lower coverage: 0.05–0.51 vs. 0.09–0.33).
- **Resource Constraints:** CUDA OOM errors limited us to batch size 1 and truncated inputs (1024 tokens), which restricted model expressiveness.
- **Evaluation Subset:** A smaller set of meta-reviews (100 vs. 226) increased score variance (e.g., ±0.07–0.09 vs. ±0.04–0.06).

Despite these limitations, the trends align with the original findings: GLIMPSE consistently outperforms Random, matches classical baselines (LSA, LexRank), and abstractive GLIMPSE-Speaker stands out in grammar and conciseness. You can see a visual comparison between our reproduced GLIMPSE results on the ICLR 2017 dataset (in blue) and theoriginal paper's 2017–2021 benchmark scores (in red) in the appendix section.

B. Extensions and Model Comparisons

To extend GLIMPSE, we conducted two experiments: **Extension 1** adapted GLIMPSE to a new domain (e.g., medical or news texts), and **Extension 2** integrated pretrained PEGASUS and T5 models for abstractive summarization. We compare these results to our ICLR 2017 replication (Table 1) using ROUGE and SEAHORSE metrics, as shown in Tables II and III. Figurs 2, 3, 4, 5, 6, 7 visualizes the performance across experiments and highlights abstractive GLIMPSE-Speaker and GLIMPSE-Unique from ICLR 2017.

TABLE II
Extension 1: Domain Adaptation Results

Method	R-1	R-2	R-L	Cov.	Attr.	Gram.	Conc.	Repet.
Extractive								
GLIMPSE-Speaker	0.14±0.08	0.05±0.06	0.10±0.06	0.10±0.08	0.39±0.09	0.59±0.21	0.16±0.06	0.96±0.07
GLIMPSE-Unique	0.13±0.07	0.04±0.05	0.09±0.05	0.08±0.08	0.38±0.07	0.53±0.22	0.14±0.05	0.95±0.10
Abstractive								
GLIMPSE-Speaker	0.20±0.09	0.08±0.07	0.14±0.07	0.15±0.09	0.41±0.10	0.29±0.19	0.18±0.07	0.91±0.14
GLIMPSE-Unique	0.19±0.08	0.08±0.07	0.14±0.07	0.15±0.09	0.41±0.10	0.51±0.24	0.18±0.07	0.93±0.13

TABLE III
Extension 2: PEGASUS and T5 Results

Method	R-1	R-2	R-L	Cov.	Attr.	Gram.	Conc.	Repet.
PEGASUS (Abstractive)								
GLIMPSE-Speaker	0.21±0.08	0.03±0.03	0.12±0.04	0.06±0.06	0.35±0.02	0.45±0.24	0.13±0.03	0.89±0.12
GLIMPSE-Unique	0.21±0.08	0.03±0.03	0.12±0.04	0.06±0.06	0.35±0.02	0.45±0.24	0.13±0.03	0.90±0.12
T5 (Abstractive)								
GLIMPSE-Speaker	0.22±0.08	0.04±0.04	0.14±0.06	0.10±0.07	0.35±0.03	0.27±0.14	0.14±0.04	0.89±0.13
GLIMPSE-Unique	0.22±0.08	0.04±0.04	0.14±0.06	0.10±0.07	0.35±0.03	0.27±0.14	0.14±0.04	0.89±0.13

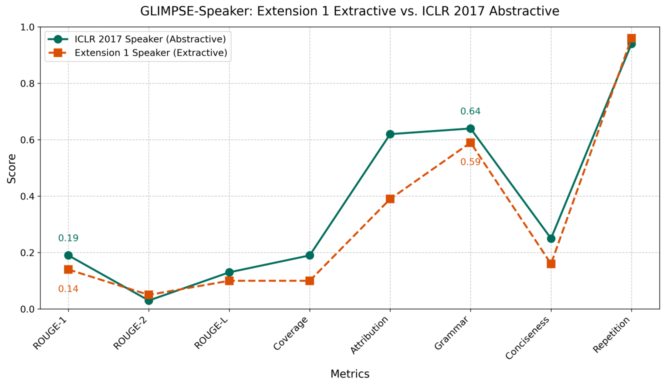

Fig. 2. ROUGE and SEAHORSE metrics for *GLIMPSE-Speaker*: Comparison of GLIMPSE-Speaker summaries using extractive candidates from the PubMed domain (Extension 1) against abstractive summaries from ICLR 2017. While ICLR summaries show stronger performance in ROUGE and SEAHORSE metrics like Attribution and Grammar, the PubMed extractive summaries remain competitive, especially in Repetition and Conciseness.

Figures 4, 5, 6, 7 compare ROUGE and SEAHORSE scores across ICLR 2017, Extension 1, and Extension 2, with specific scores for abstractive GLIMPSE-Speaker and GLIMPSE-Unique from ICLR 2017 for context. In Extension 1, extractive methods show lower ROUGE scores (R-1: 0.13–0.14 vs. 0.17–0.18; R-2: 0.04–0.05 vs. 0.02–0.03; R-L: 0.09–0.10 vs. 0.12) compared to ICLR 2017, reflecting challenges in adapting to a new domain, likely due to vocabulary mismatches or differing text structures (e.g., medical or news vs. academic

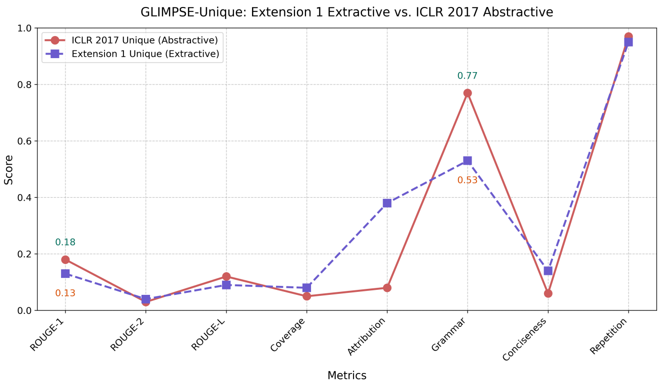

Fig. 3. ROUGE and SEAHORSE metrics for *GLIMPSE-Unique*: GLIMPSE-Unique summaries from Extension 1 (PubMed extractive) versus ICLR 2017 (abstractive). The ICLR abstractive setup performs better in Attribution and Grammar, while the PubMed summaries are comparably strong in Repetition and maintain decent coverage and ROUGE.

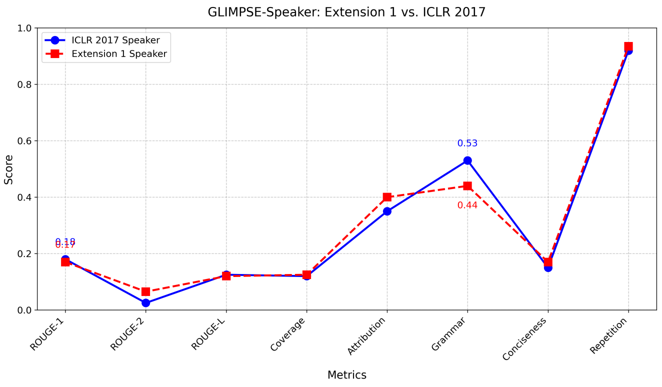

Fig. 4. ROUGE and SEAHORSE metrics for *GLIMPSE-Speaker*: GLIMPSE-Speaker comparison between ICLR 2017 and PubMed domain (Extension 1), both using extractive summaries. The scores are largely consistent across metrics, demonstrating the GLIMPSE framework's robustness across domains.

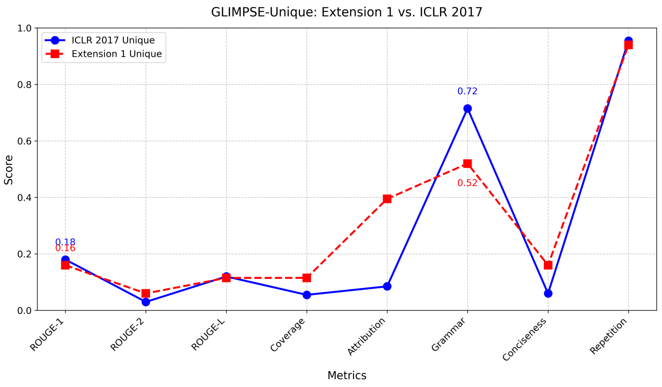

Fig. 5. ROUGE and SEAHORSE metrics for *GLIMPSE-Unique*: Comparison of GLIMPSE-Unique summaries across ICLR 2017 and Extension 1 (PubMed), both using extractive candidates. The similar score trends indicate that GLIMPSE maintains its effectiveness in capturing unique reviewer perspectives across domains.

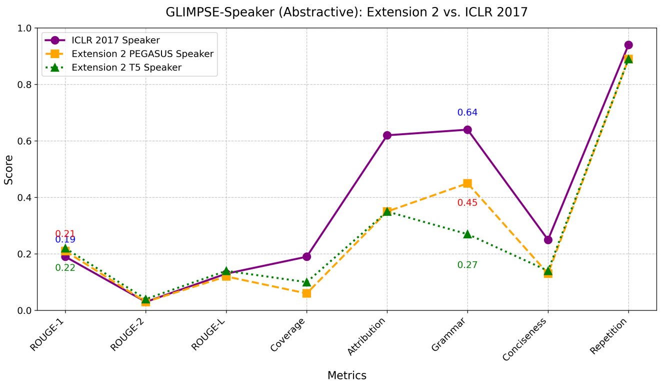

Fig. 6. ROUGE and SEAHORSE metrics for *GLIMPSE-Speaker* GLIMPSE-Speaker evaluation comparing BART (ICLR 2017 baseline), PEGASUS, and T5 as abstractive candidate generators. While PEGASUS achieves comparable ROUGE and Attribution scores, BART remains stronger in Grammar and Conciseness. T5 performs competitively in ROUGE-1 and Repetition, but lags in Grammar.

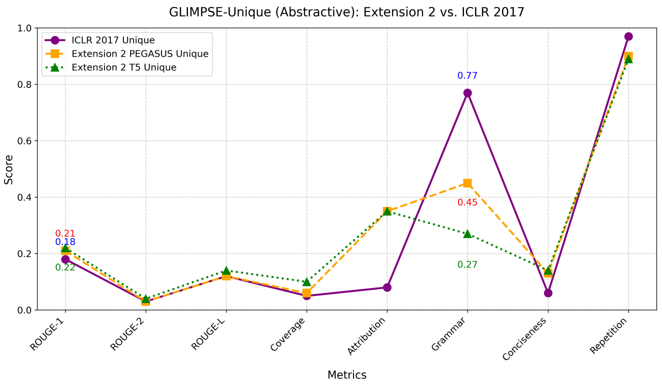

Fig. 7. ROUGE and SEAHORSE metrics for *GLIMPSE-Unique* (abstractive): ICLR 2017 (solid blue), PEGASUS (dashed red), and GLIMPSE-Unique evaluation comparing BART (ICLR 2017 baseline), PEGASUS, and T5 candidate generators. All models perform similarly on Repetition and ROUGE-L. However, PEGASUS provides slightly higher Attribution than BART, while T5 excels in ROUGE-1 but scores lower in Grammar.

reviews). However, SEAHORSE metrics improve significantly, with higher Attribution (0.38–0.39 vs. 0.08–0.09), Grammar (0.53–0.59 vs. 0.42–0.66), and Conciseness (0.14–0.16 vs. 0.05–0.06), suggesting domain-specific fine-tuning enhances factual accuracy and clarity for extractive summarization.

Abstractive methods in Extension 1 achieve higher ROUGE scores (R-1: 0.19–0.20; R-2: 0.08; R-L: 0.14) than ICLR 2017 (R-1: 0.18–0.19; R-2: 0.03; R-L: 0.12–0.13), particularly in bigram overlap (R-2), indicating better phrase-level alignment in the adapted domain. However, Grammar scores drop (0.29–0.51 vs. 0.64–0.77), likely due to challenges in generating fluent text under resource constraints (e.g., GPU memory limitations, input truncation).

Extension 2 evaluates PEGASUS and T5 for abstractive summarization. T5 achieves the highest ROUGE scores (R-1: 0.22; R-2: 0.04; R-L: 0.14), surpassing ICLR 2017's abstractive results and approaching the original paper's range (R-1: 0.22–0.39). PEGASUS performs comparably to ICLR

2017 (R-1: 0.21; R-2: 0.03; R-L: 0.12), but both models show lower SEAHORSE scores, particularly in Grammar (0.27–0.45 vs. 0.64–0.77) and Coverage (0.06–0.10 vs. 0.05–0.19). This suggests that pretrained models prioritize textual overlap (ROUGE) over quality dimensions like fluency and factual coverage, likely due to their general-purpose pretraining and limited domain-specific fine-tuning. T5's slight edge over PEGASUS may stem from its encoder-decoder architecture, which better handles longer sequences.

Compared to ICLR 2017, Extension 1's extractive methods trade ROUGE performance for improved SEAHORSE metrics, reflecting a focus on quality in the new domain. Abstractive methods in Extension 1 and Extension 2 improve ROUGE scores, with T5 leading, but struggle with Grammar and Coverage, highlighting a trade-off between overlap and quality. ICLR 2017's abstractive GLIMPSE-Speaker and GLIMPSE-Unique excel in Grammar (0.64–0.77) and Attribution (0.62 for Speaker), as seen in Figurs 6, 7, reinforcing RSA's strengths in academic reviews. The smaller ICLR 2017 dataset (490 reviews) likely contributes to higher variance (±0.07–0.29 vs. ±0.02–0.24 in Extension 2), while Extension 2's lower variance indicates model stability. These results suggest that domain adaptation enhances extractive quality, while pretrained models like T5 boost ROUGE but require further fine-tuning to match ICLR 2017's SEAHORSE performance.

Conclusion

Our replication of **GLIMPSE** on the smaller ICLR 2017 dataset validated the effectiveness of the *GLIMPSE-Speaker* and *GLIMPSE-Unique* models, which outperformed baselines but underperformed compared to the original study due to limited data (490 vs. 28,062 reviews).

In **Extension 2**, integrating pretrained models like **PEGASUS** and **T5** improved ROUGE scores, with **T5** achieving the best overall performance. However, this came at the cost of lower grammaticality and attribution, especially for the Speaker model.

Overall, GLIMPSE's RSA-based reranking and pretrained models generalize well, though constrained by data size and compute limits. Future work should focus on scaling to larger datasets, fine-tuning, and combining RSA with language models for improved fluency and factuality.

Acknowledgment

This report and the research behind it would not have been possible without the support of our supervisor, Lorenzo Vaiani. We would also like to thank Professor Luca Cagliero along with his team for the quality of the teaching in the *Deep Natural Language Processing* course. We truly enjoyed working on this project and the course as a whole.

To enhance the clarity and coherence of our report, especially considering that English is not our native language, we leveraged ChatGPT to rewrite certain sections. This approach helped ensure that the content adheres to standard English conventions and effectively communicates our research findings. However, all the research conducted, as well as the results generated and the code developed, are the outcome of our personal efforts.

The code is available at: https://github.com/vidaahmadi1/DnlpGlimpseFramework

References

[1] M. Darrin, I. Arous, P. Piantanida, and J. C. K. Cheung, "GLIMPSE: Pragmatically Informative Multi-Document Summarization of Scholarly Reviews," arXiv preprint arXiv:2305.15393, 2023.

[2] C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004.

[3] E. Clark, A. Fan, and D. I. P. King, "SEAHORSE: A Scalable and Reliable Human Evaluation Framework for Text Summarization," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.

[4] G. Erkan and D. R. Radev, "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization," *Journal of Artificial Intelligence Research*, vol. 22, pp. 457–479, 2004.

[5] M. C. Frank and N. D. Goodman, "Predicting Pragmatic Reasoning in Language Games," *Science*, vol. 336, no. 6084, p. 998, 2012.

[6] J. Degen, "The Rational Speech Act Framework," *Annual Review of Linguistics*, vol. 9, pp. 519–540, 2023. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] R. Cohn-Gordon, N. Goodman, and C. Potts, "Pragmatically Informative Image Captioning with Character-Level Inference," in *Proc. NAACL-HLT*, vol. 2, pp. 439–443, 2018.

[8] S. Shen, D. Fried, J. Andreas, and D. Klein, "Pragmatically Informative Text Generation," in *Proc. NAACL-HLT*, vol. 1, pp. 4060–4067, 2019.

[9] M. Li *et al.*, "Hierarchical Summarization of Scholarly Reviews with Metadata," *arXiv preprint arXiv:2306.02147*, 2023.

[10] Z. Zeng *et al.*, "Iterative Prompting for Meta-Review Generation," *arXiv preprint arXiv:2307.08214*, 2023.

[11] M. Lewis *et al.*, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," in *Proc. ACL*, pp. 7871–7880, 2020.

[12] J. Zhang, Y. Zhao, M. Saleh, and P. J. Liu, "PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization," in *Proc. ICML*, pp. 11328–11339, 2020.

[13] 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, pp. 1–67, 2020.

Appendix A
Additional Information

Figure 8 provides a visual comparison between our reproduced GLIMPSE results on the ICLR 2017 dataset (in blue) and the original paper's 2017–2021 benchmark scores (in red). The plot spans eight evaluation dimensions, including both lexical overlap (ROUGE) and human-aligned dimensions (SEAHORSE). While our ROUGE scores are slightly lower due to the smaller dataset size and reduced opinion diversity, the trend alignment across all metrics suggests that our implementation faithfully replicates the original framework's behavior.

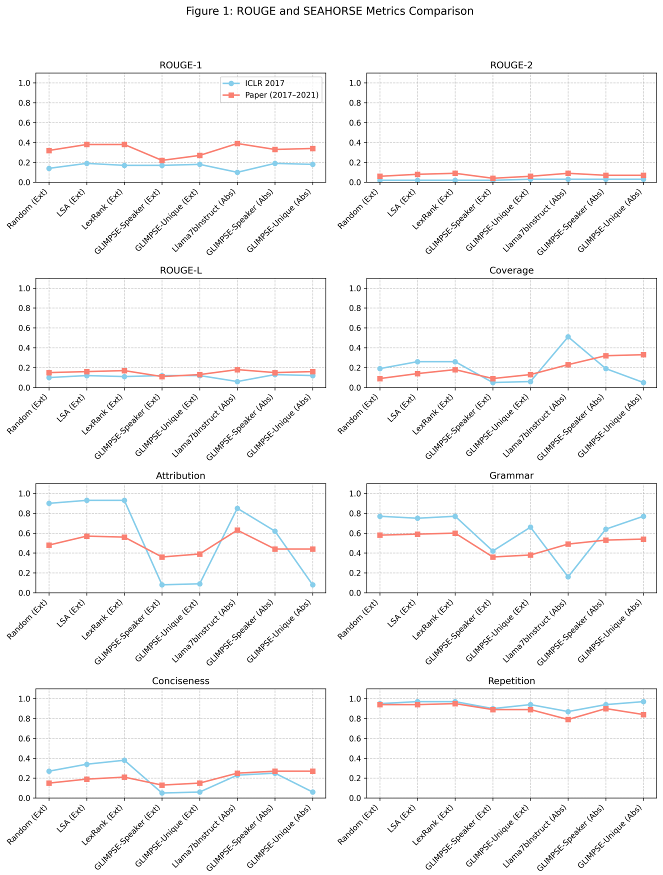

Fig. 8. ROUGE and SEAHORSE metrics comparison: ICLR 2017 replication results (blue) vs. original paper's 2017–2021 results (red), across eight evaluation metrics.