Supplementary Material for 'MetaSearch: Search-Augmented LLM with Reasoning for Consensus Resolution in Peer Review' Paper

Workflow of MetaSearch

This section provides a detailed explanation of the workflow implemented in **MetaSearch: Search-Augmented LLM with Reasoning for Consensus Resolution in Peer Review**. The accompanying diagram (**Figure 1**) visually represents the structured process followed by our system, from data acquisition to consensus resolution.

1. Dataset and Preprocessing

MetaSearch utilizes the **oaimli/Peersum dataset** from **HuggingFace**, which contains structured peer review data, including reviewer critiques, meta-reviews, and acceptance decisions. We perform **random sampling** to select a diverse subset of papers for analysis.

2. Critique Points Extraction

Critique points are extracted using two complementary approaches:

- **LLM-Based Approach**: We employ **Gemini 2.0 Flash-Lite** to identify and extract critique points from reviews.
- **NLP-Based Approach**: We use **semantic similarity matching** to retrieve critique points and compute **disagreement scores** based on cosine similarity.

3. Disagreement Detection

To quantify disagreements among reviewers, we implement:

- **LLM-Based Scoring**: Gemini 2.0 Flash-Lite provides structured disagreement details.
- **Similarity-Based Scoring**: A disagreement score is derived as **1 - (cosine similarity)** between extracted critique embeddings.

4. Search-Augmented Peer Review Agent

To enhance critique verification and consensus resolution, we integrate a **Peer Review Agent** that retrieves relevant information from:

- **ArXiv API**
- **Google Scholar API**
- **Semantic Scholar API**
- **Tavily Search**

The retrieved literature provides **state-of-the-art (SoTA) references** for validation.

5. Reasoning and Consensus Resolution

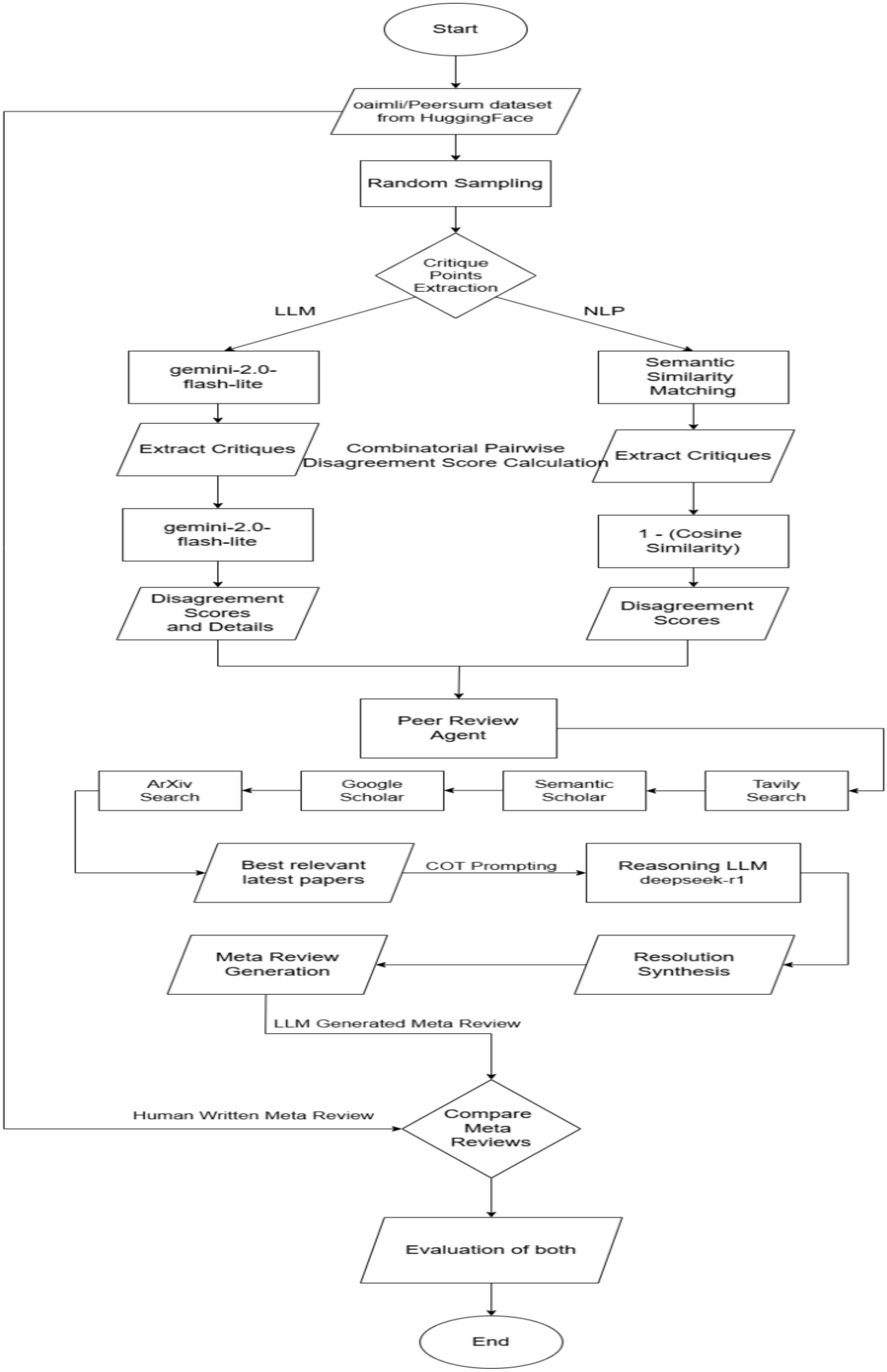
MetaSearch incorporates **DeepSeek-R1**, a specialized **reasoning LLM**, to synthesize and resolve reviewer disagreements. The system generates:

- **Meta Reviews** summarizing reviewer critiques.
- **Resolution Synthesis**, identifying key consensus points.

6. Evaluation and Comparison

We compare **LLM-generated meta-reviews** against the actual **meta-reviews in Peersum**, evaluating:

- Agreement with ground truth meta-reviews.
- Quality of consensus resolution.

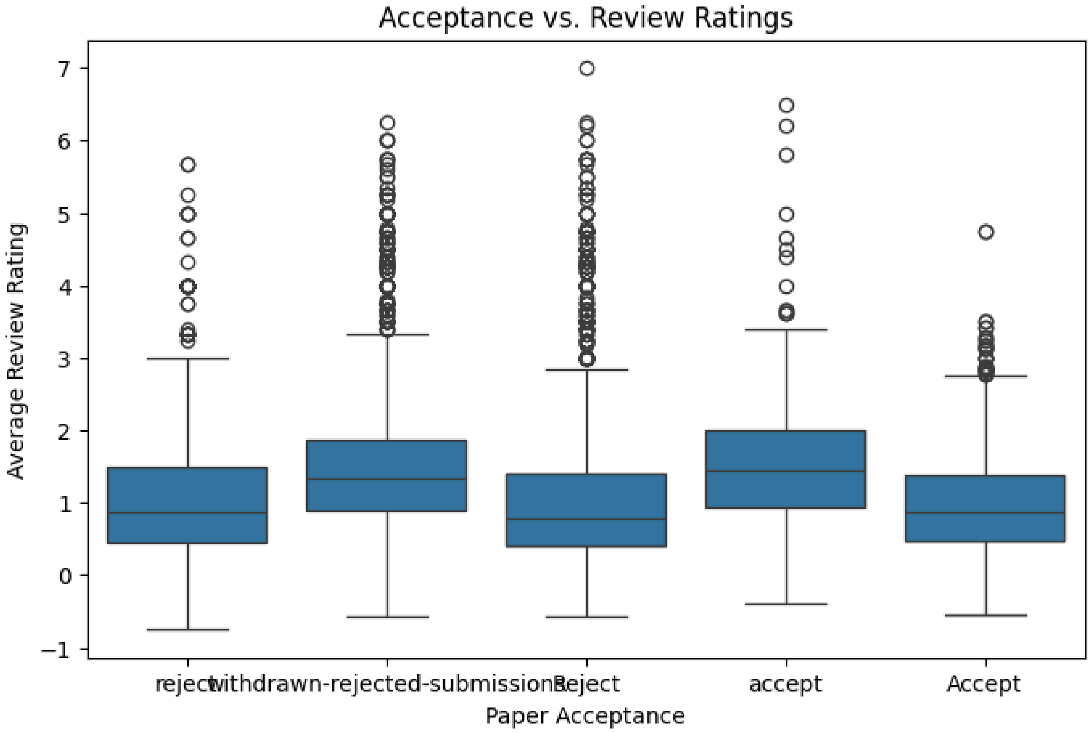

Start
oaimli/Peersum dataset from HuggingFace
Random Sampling
Critique Points Extraction
LLM
NLP
gemini-2.0-flash-lite
Semantic Similarity Matching
Extract Critiques
Combinatorial Pairwise Disagreement Score Calculation
Extract Critiques
gemini-2.0-flash-lite
1 - (Cosine Similarity)
Disagreement Scores and Details
Disagreement Scores
Peer Review Agent
ArXiv Search
Google Scholar
Semantic Scholar
Tavily Search
Best relevant latest papers
COT Prompting
Reasoning LLM deepseek-r1
Meta Review Generation
Resolution Synthesis
LLM Generated Meta Review
Human Written Meta Review
Compare Meta Reviews
Evaluation of both
End

Figures

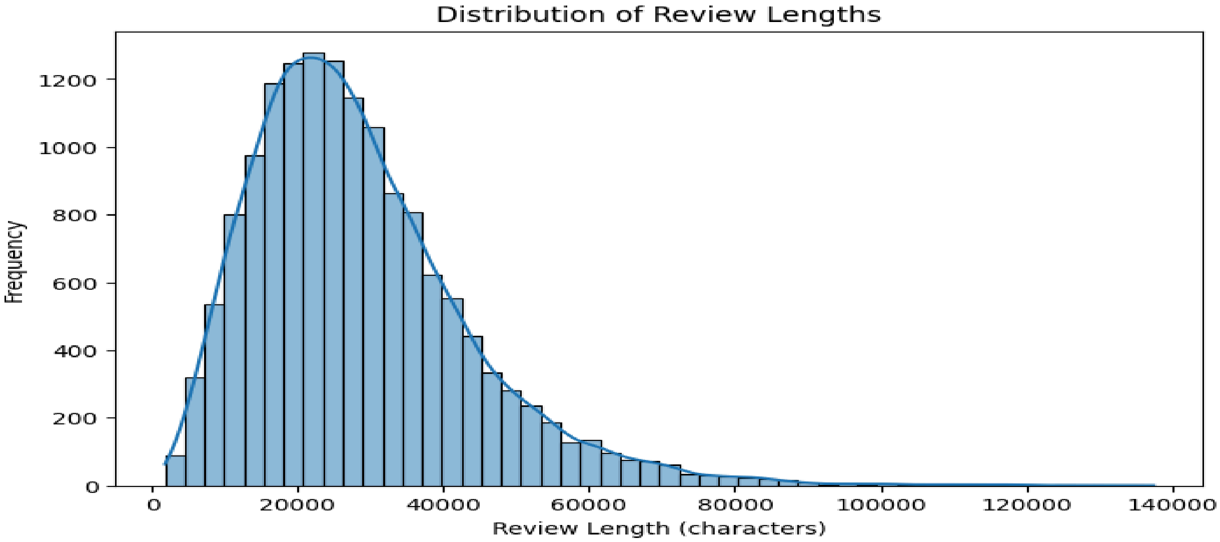
This section provides an overview of the figures included in the repository. Each figure is designed to illustrate key aspects of our analysis and methodology, offering insights into review dynamics, disagreement resolution, and the performance of our meta review generation. Each of these figures are mentioned in the 'Results and Analysis' section of the paper.

Dataset Characteristics and Reviewer Behavior

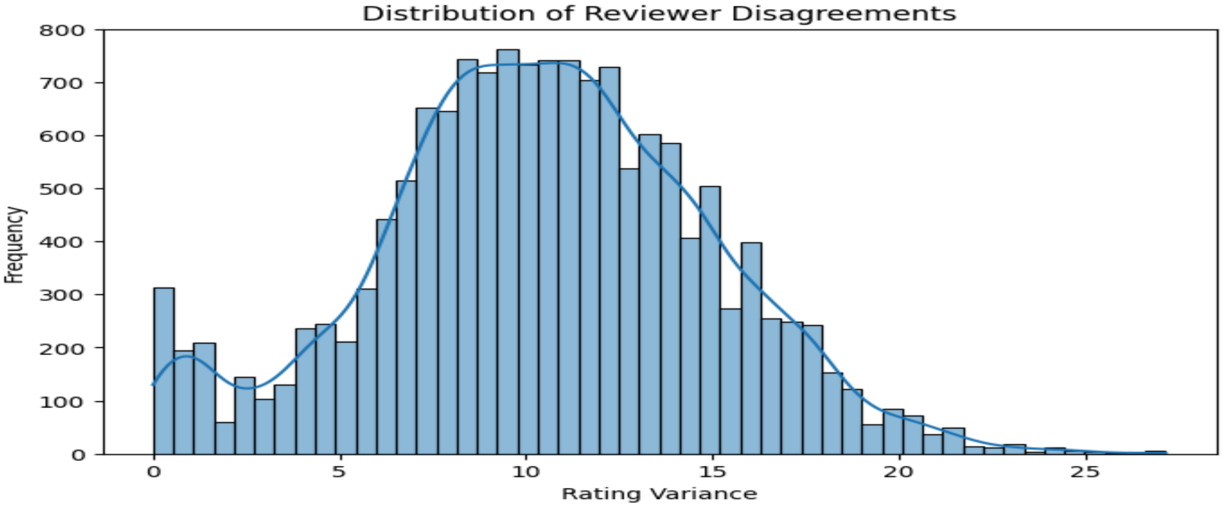
1. **Acceptance vs. Review Ratings**
   File: *Acceptance vs. Review Ratings.png*
   Subsection: Dataset Characteristics and Reviewer Behavior
   Explanation:
   This figure visualizes the relationship between review ratings and paper acceptance outcomes. It shows that higher ratings tend to be associated with accepted papers and highlights patterns such as a peak at zero variance (complete consensus) and a right-skewed distribution of scores.



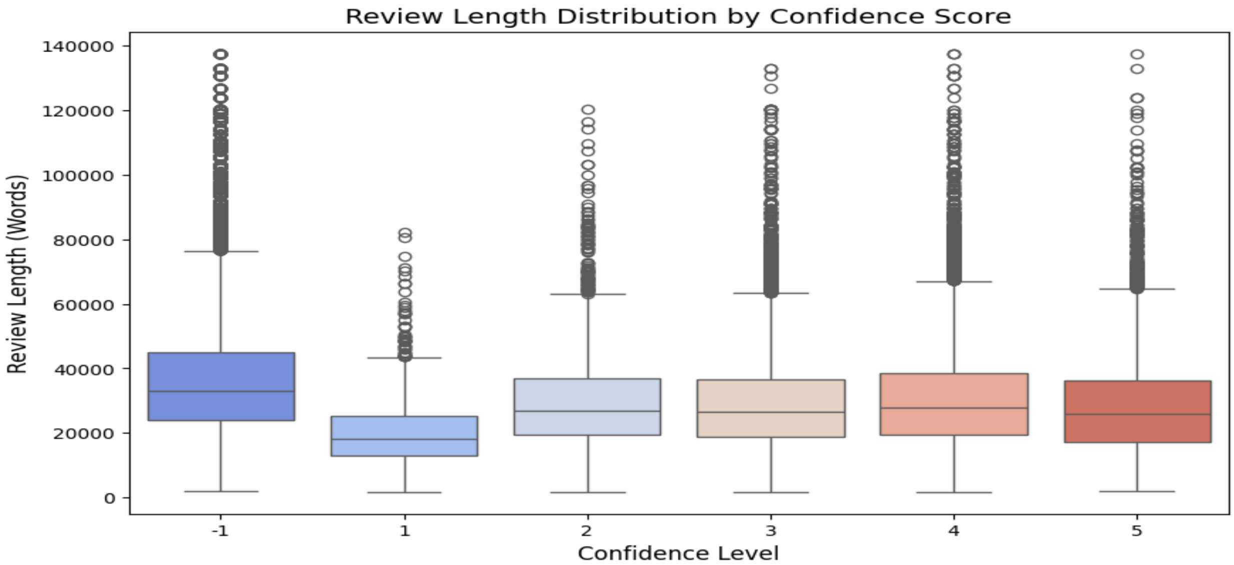
2. **Distribution of Review Lengths**
 File: *Distribution of Review Lengths.png*
 Subsection: Dataset Characteristics and Reviewer Behavior
 Explanation:
 This image presents the overall distribution of review lengths, providing a baseline measure of how detailed reviewers are in their evaluations.



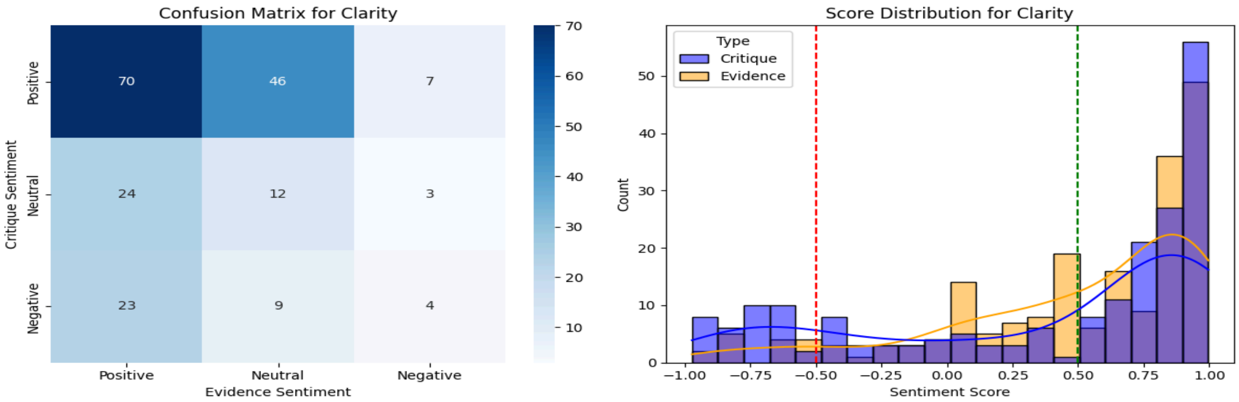
3. **Distribution of Reviewer Disagreements**
 File: *Distribution of Reviewer Disagreements.png*
 Subsection: Dataset Characteristics and Reviewer Behavior
 Explanation:
 This figure quantifies the extent of disagreement among reviewers. It reinforces the narrative that reviewer opinions vary widely, emphasizing the inherent subjectivity of academic evaluations.



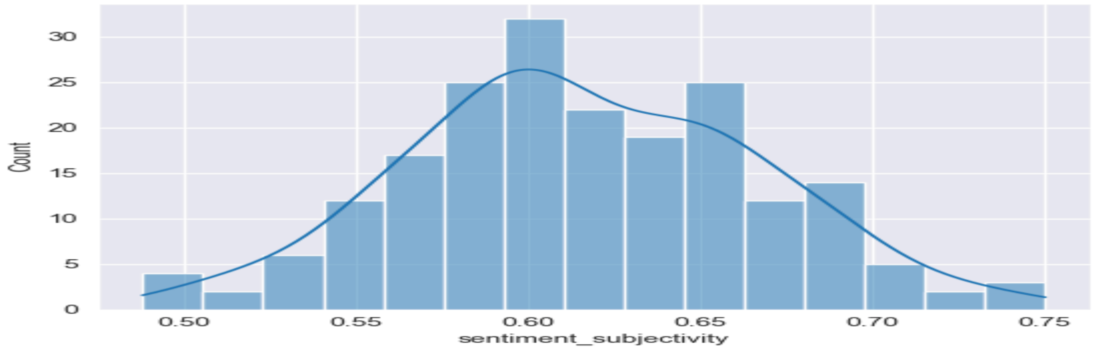
4. **Review Length Distribution by Confidence Score**
   File: *Review Length Distribution by Confidence Score.png*
   Subsection: Dataset Characteristics and Reviewer Behavior
   Explanation:
   This visualization breaks down review lengths by reviewer confidence. It indicates that reviewers with lower confidence show higher variability in length—either being overly verbose or notably brief.



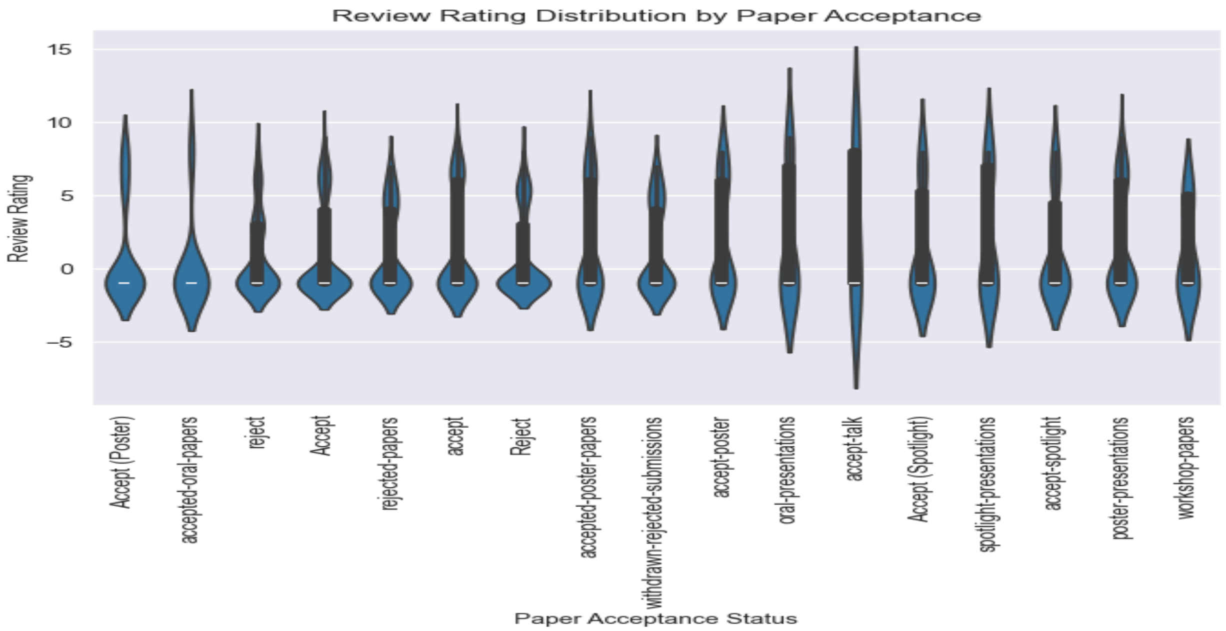
5. **Sentiment Score Distribution for Clarity**
   File: *Sentiment Score Distribution for Clarity.png*
   Subsection: Dataset Characteristics and Reviewer Behavior
   Explanation:
   This figure focuses on sentiment scores related to the clarity of reviews. It provides insight into how reviewers assess manuscript clarity and contributes to understanding the overall critical tone of the feedback.



6. **Sentiment Subjectivity Histogram**
   File: *Sentiment Subjectivity Histogram.png*
   Subsection: Dataset Characteristics and Reviewer Behavior
   Explanation:
   This histogram shows that reviews are predominantly opinion-based, with high subjectivity scores. This supports the claim that peer reviews are largely driven by personal perspectives rather than purely objective assessments.

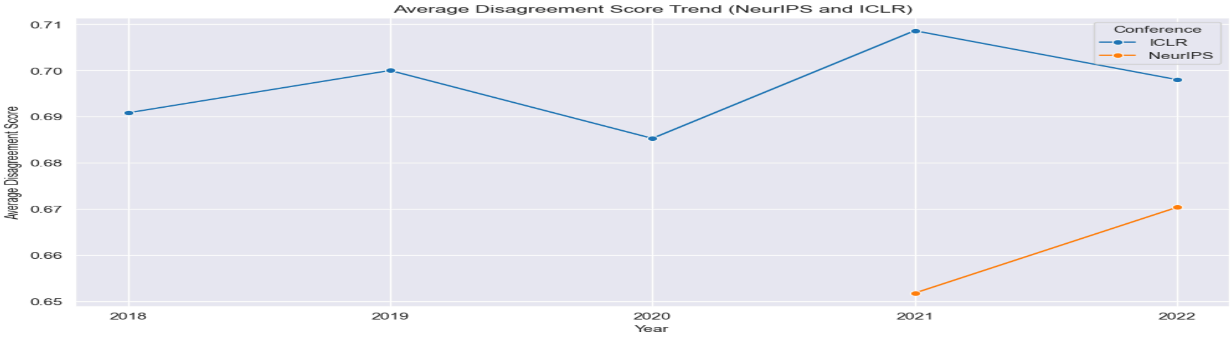


7. **Review Rating Distribution by Paper Acceptance**
   File: *Review Rating Distribution by Paper Acceptance.png*
   Subsection: Dataset Characteristics and Reviewer Behavior
   Explanation:
   This detailed breakdown further illustrates the relationship between review ratings and paper acceptance. It confirms that higher ratings typically correlate with acceptance, underscoring the internal consistency of the review process.

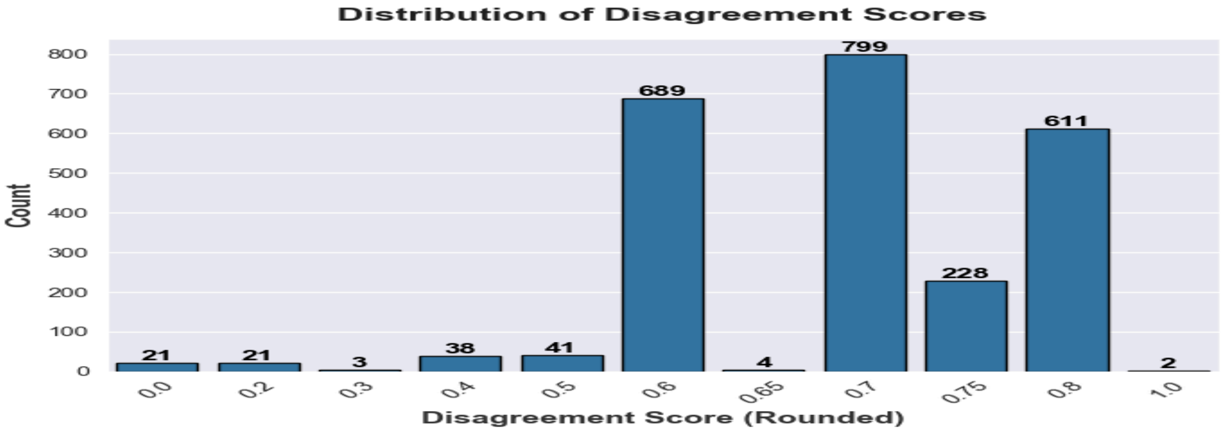


Critique Points Extraction and Disagreement Detection

1. **Average Disagreement Score Trend (NeurIPS and ICLR)**
   File: *Average Disagreement Score Trend (NeurIPS and ICLR).png*
   Subsection: Critique Points and Disagreement Detection
   Explanation:
   This figure compares the average disagreement scores between NeurIPS and ICLR submissions. It highlights that while both conferences exhibit significant reviewer disagreement, NeurIPS tends to have marginally lower disagreement scores, illustrating differences in reviewer consensus across venues.



2. **Distribution of Disagreement Scores**
   File: *Distribution of Disagreement Scores.png*
   Subsection: Critique Points and Disagreement Detection
   Explanation:
   This histogram shows the spread of disagreement scores across review pairs. The right-skewed distribution emphasizes that higher disagreement scores are common, reinforcing the challenges of synthesizing diverse reviewer opinions.



Search Verification

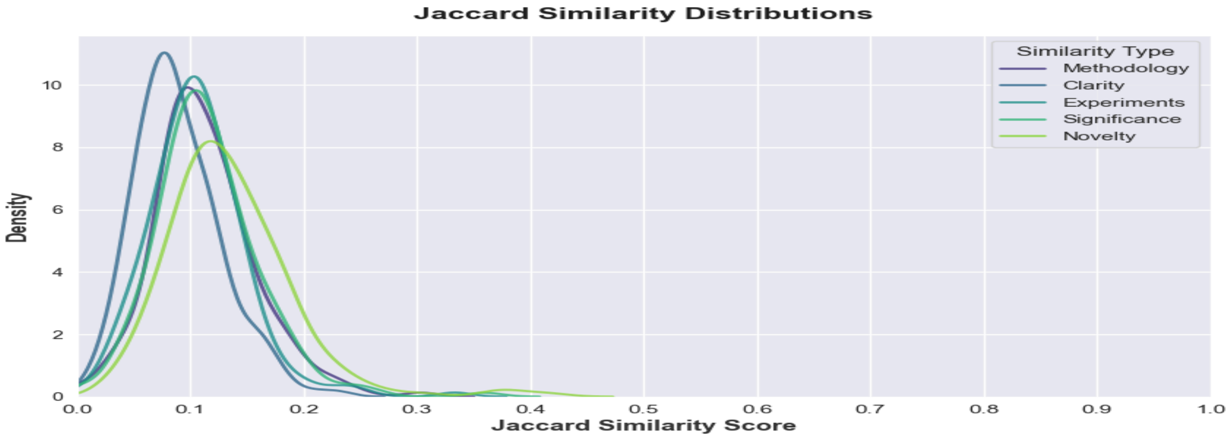
1. **Jaccard Similarity Distributions**

   File: *Jaccard Similarity Distributions.png*

   Subsection: Search-based Evidence and Literature Retrieval

   Explanation:

   This figure presents Jaccard similarity scores between the extracted critique points and the retrieved evidence. Despite high semantic similarity, the low lexical overlap confirms that the retrieval process captures novel and relevant content rather than mere textual matches.



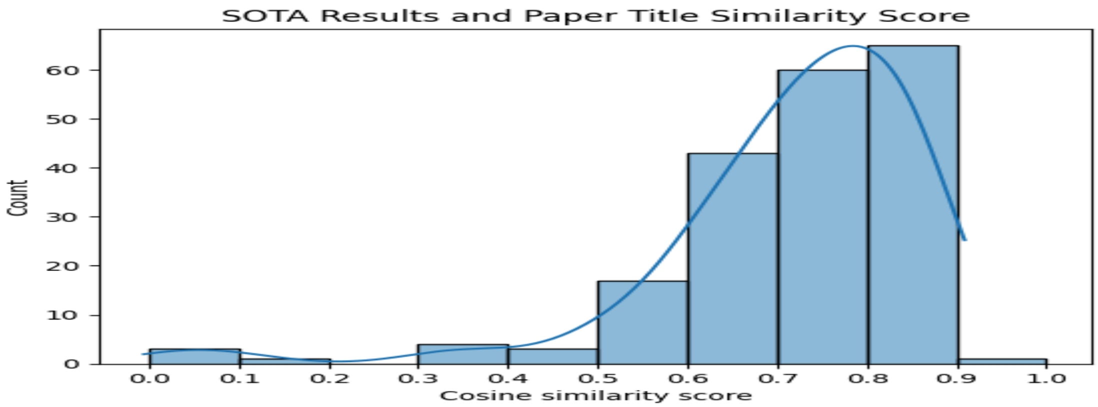
2. **SOTA Results and Paper Title Similarity Score**

   File: *SOTA Results and Paper Title Similarity Score.png*

   Subsection: Search-based Evidence and Literature Retrieval

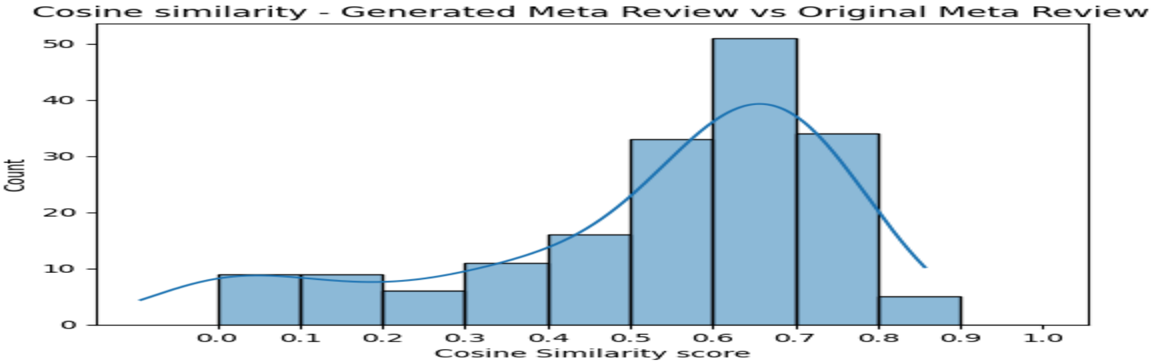
   Explanation:

   This visualization correlates state-of-the-art (SOTA) paper retrieval outcomes with similarity scores based on paper titles. It validates the multi-source search pipeline by demonstrating that the retrieved literature is semantically aligned with the target work.

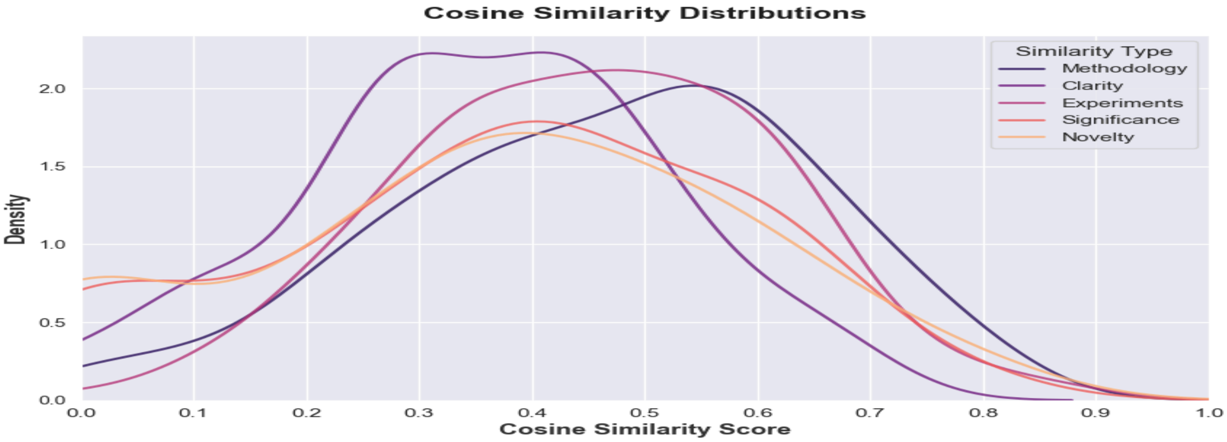


Consensus Resolution and Meta Review Generation

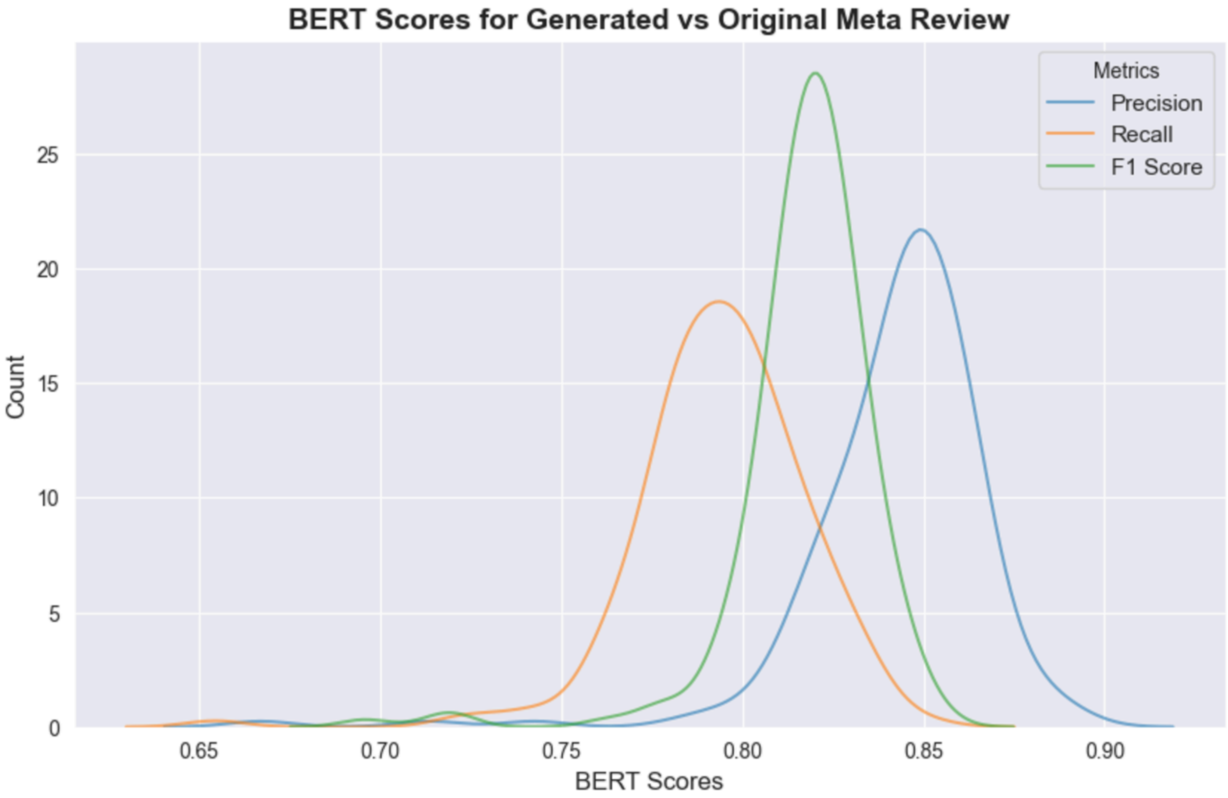
1. **Cosine Similarity – Generated Meta Review vs Original Meta Review**
   File: *Cosine similarity - Generated Meta Review vs Original Meta Review.png*
   Subsection: Consensus Resolution and Meta-Review Generation
   Explanation:
   This figure depicts the distribution of cosine similarity scores comparing AI-generated meta-reviews with human-written ones. The high semantic similarity indicates that the generated meta-reviews effectively capture the essential evaluative content of the original reviews.



2. **Cosine Similarity Distributions of Accepted Critiques vs Final Resolution Summary**
   File: *cosine.png*
   Subsection: Consensus Resolution and Meta-Review Generation
   Explanation:
   This image shows the similarity between accepted critique points and the final resolution summaries. The right-skewed distribution demonstrates that while most critiques are closely aligned with the generated summaries, there is a range of similarity that reflects the nuanced synthesis process.



3. **BERT Precision, Recall and F1 Scores for AI Generated vs Original Meta-Reviews**
   File: *meta_review_bert.png*
   Subsection: Consensus Resolution and Meta-Review Generation
   Explanation:
   This figure evaluates the quality of the AI-generated meta-reviews using BERT-based metrics. With F1 scores predominantly above 0.75, it confirms that the generated reviews successfully capture the semantic content of the original meta-reviews despite differences in phrasing, underscoring the efficacy of the consensus resolution mechanism.


BERT Scores for Generated vs Original Meta Review
Metrics
Precision
Recall
F1 Score
Count
BERT Scores