ALIGNING REVIEWER GUIDELINES AND REVIEWER FEEDBACK: A DATA-DRIVEN STUDY

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

ÖZER TANRISEVER

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF DATA INFORMATICS

JANUARY 2026

Aligning Reviewer Guidelines and Reviewer Feedback: A Data-Driven Study

Submitted by **ÖZER TANRISEVER** in partial fulfillment of the requirements for the degree of **Master of Science in Data Informatics Department, Middle East Technical University** by,

Date: 14.01.2026

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Özer Tanrısever

Signature :

ABSTRACT

ALIGNING REVIEWER GUIDELINES AND REVIEWER FEEDBACK: A DATA-DRIVEN STUDY

Tanrısever, Özer
M.S., Department of Data Informatics
Supervisor: Prof. Dr. Tuğba Taşkaya Temizel

January 2026, 126 pages

Academic venues have established reviewer guidelines to enable standardized evaluation. This study proposes a framework designed to categorize reviewer inquiries and systematically measure their alignment with institutional guidelines. The analysis utilizes peer review data from the 2024 International Conference on Learning Representations (ICLR) on the OpenReview platform. The framework employs a question extraction pipeline utilizing Large Language Models (LLMs) and embedding similarity to decompose unstructured review inquiry text into semantically coherent chunks. A stratified sample of 427 reviews selected from 22,358 yielded 760 question chunks following multi-agent consensus validation. Generative topic modeling identified 13 topics within this dataset, including *Methodology* and *Computational Efficiency*. Finally, these empirically derived topics are mapped against evaluation criteria extracted from eight unique reviewer guidelines representing ten top-tier AI venues. Topic generation and mapping underwent manual and LLM-based verification, with an additional cross-check by a senior academic to reduce subjectivity. The results reveal a divergence between prescribed and practiced criteria. *Ethical Considerations* appears in six out of eight guidelines but ranks as the least frequent topic in the dataset. In contrast, *Presentation and Figures* appears only in two guidelines, but accounts for 8.29% of the reviewer inquiries. The findings offer actionable insights for conference organizers to refine guidelines and enable authors to understand the prioritized reviewer criteria in practice, ultimately enhancing the reliability of the peer review ecosystem.

Keywords: Large Language Model, Generative Topic Modeling, LLM-as-a-Judge, Data-Driven Guideline Alignment, Academic Peer Review

ÖZ

HAKEM YÖNERGELERİ İLE GERİ BİLDİRİMLERİNİN UYUMU: VERİ TEMELLİ BİR ÇALIŞMA

Tanrısever, Özer
Yüksek Lisans, Veri Bilişimi Bölümü
Tez Yöneticisi: Prof. Dr. Tuğba Taşkaya Temizel

Ocak 2026, 126 sayfa

Akademik yayın mecraları, değerlendirme sürecinde bir standart oluşturması adına hakem yönergeleri oluştururlar. Bu çalışma, hakemlerin değerlendirmelerinde sormuş oldukları soruları kategorize etmek ve ilgili yönergelerle uyumunu sistematik olarak ölçmek üzere tasarlanmış bir yapı sunmaktadır. Analiz için, OpenReview platformunda bulunan International Conference on Learning Representations (ICLR) mecrasının 2024 iterasyonunda bulunan değerlendirme verileri kullanılmıştır. Sunulan yapı, hakem sorgu metinlerini anlamsal olarak tutarlı öbeklere ayrıştırmak amacıyla Büyük Dil Modelleri (BDM'ler) ve gömü benzerliğinden yararlanan bir yapı içermektedir. 22.358 değerlendirmeden tabakalı örneklem ile seçilen 427 değerlendirme, çoklu ajan mutabakat doğrulamasından geçirilmiştir ve içerisinden 760 soru öbeği elde edilmiştir. Üretken konu modellemesi ile bu veri seti içerisinde, *Metodoloji* ve *Hesaplama Verimliliği* de dahil olmak üzere 13 farklı konu tespit edilmiştir. Karşılaştırma ölçütlerini belirlemek amacıyla, on adet önde gelen Yapay Zekâ (YZ) yayın mecrasını temsil eden sekiz farklı hakem yönergesindeki değerlendirme kriterleri çıkarılmıştır. Çalışmanın son aşamasında, ampirik olarak elde edilen bu konular, söz konusu kriterlerle eşleştirilmiştir. Konu üretimi ve eşleştirme süreçleri hem manuel hem de BDM tabanlı doğrulamadan geçirilmiş; öznelliği en aza indirmek amacıyla deneyimli bir akademisyen tarafından ek bir çapraz kontrole tabi tutulmuştur. Sonuçlar, yönergeler ile uygulanan kriterler arasındaki farkı ortaya koymaktadır. *Etik Hususlar* sekiz yönergenin altısında yer almasına rağmen, veri setinde en düşük sayıda dokümana sahip olan konu olarak sıralanmıştır. Buna karşılık, *Sunum ve Şekiller* yalnızca iki yönergede yer almasına rağmen, hakem sorgularının %8,29'unu oluşturmaktadır. Elde edilen bulgular, konferans organizatörlerinin yönergeleri iyileştirmesi ve yazarların hakemler

tarafından önceliklendirilen kriterlerini anlaması adına uygulanabilir içgörüler sunmakta ve nihayetinde hakem değerlendirme ekosisteminin güvenilirliğini artırmayı hedeflemektedir.

Anahtar Kelimeler: Büyük Dil Modelleri, Üretken Konu Modelleme, Değerlendirici Olarak Büyük Dil Modelleri, Veri Odaklı Yönerge Uyumu, Akademik Hakem Değerlendirmesi



To those who take action

ACKNOWLEDGMENTS

First and foremost, I wish to extend my deepest appreciation to my supervisor, Prof. Dr. Tuğba Taşkaya Temizel. Her mentorship, insightful feedback, and academic rigor have been instrumental in defining the direction and quality of this research. I count myself fortunate to have benefited from her expertise and invaluable guidance throughout this process.

I also wish to thank Middle East Technical University, and specifically the Informatics Institute, for providing a strong academic environment that fosters intellectual and professional growth.

I gratefully acknowledge the support of EPAM Systems for providing the essential computational resources required to conduct the experimental phase of this study.

My heartfelt thanks go to my family for their constant support, understanding, and encouragement throughout this journey.

Finally, I thank everyone who contributed, directly or indirectly, to the realization of this thesis.

TABLE OF CONTENTS

ABSTRACT iv

ÖZ vi

DEDICATION viii

ACKNOWLEDGMENTS ix

TABLE OF CONTENTS x

LIST OF TABLES xvii

LIST OF FIGURES xx

LIST OF ABBREVIATIONS xxii

CHAPTERS

1 INTRODUCTION 1

1.1 Research Questions 2

1.2 Contributions of the Study 3

1.3 Organization of the Thesis 4

2 RELATED WORK 5

2.1 Computational Analysis of Peer Reviews 5

2.2 Evolution of Topic Modeling Approaches 6

2.3 Limitations in Current Peer Review Analysis 7

3 DATASET AND EXPLORATORY DATA ANALYSIS 9
3.1 Data Acquisition 9
3.1.1 The OpenReview Platform as a Data Source 9
3.1.2 Selection of the ICLR 2024 Venue 10
3.2 Dataset Construction 10
3.2.1 Reproducibility and Pipeline Management 11
3.2.2 The OpenReview Data Model 12
3.2.2.1 The "Note" Object and Relational Aggregation 12
3.2.3 Data Ingestion and Schema Validation 12
3.2.3.1 Feature Parsing and Normalization 12
3.2.3.2 Database Structure 13
3.2.4 Data Filtering and Integrity Verification 13
3.3 Exploratory Data Analysis 14
3.3.1 Submission Outcomes 14
3.3.2 Data Licenses 14
3.3.3 Reviewer and Community Participation 15
3.3.4 Analysis of Quantitative and Textual Metrics 17
3.3.4.1 Review Distribution per Submission 17
3.3.4.2 Quantitative Metrics 17
3.3.4.3 Textual Metrics 22
3.3.5 Correlation Analysis of Features 23
3.3.5.1 Selection of Correlation Metric 23

3.3.5.2 Results and Interpretation 24

4 METHODOLOGY 27

4.1 Experimental Design and Data Sampling 28

4.1.1 Stratified Sampling Strategy 28

4.1.2 Distribution Verification 29

4.2 Source-Anchored Information Extraction 30

4.2.1 Non-Informative Text Filtering 32

4.2.2 Context-Aware Extraction 32

4.2.3 Hallucination Mitigation 33

4.3 Multi-Stage Data Validation Framework 34

4.3.1 Stage 1: Deterministic Fidelity Filtering 34

4.3.2 Stage 2: Semantic Consensus via LLM-as-a-Judge 34

4.3.2.1 Judge Selection and Diversity 35

4.3.2.2 Inter-Rater Reliability 35

4.3.2.3 Final Consensus and Dataset Pruning 36

4.4 Generative Topic Modeling 38

4.4.1 Final Corpus Preparation 38

4.4.2 Domain-Specific Prompt Adaptation 38

4.4.3 Multi-Run Topic Discovery and Consolidation 39

4.5 Logit-Based Soft Topic Classification 39

4.5.1 Soft Classification 40

4.5.2 Domain Coverage Verification 40

4.5.3 Quantification of Semantic Uncertainty 41
4.5.3.1 Normalized Shannon Entropy 41
4.5.3.2 Probability Margin 41
4.5.3.3 Operational Interpretation and the RLHF Effect 42
4.5.4 Topic Separability via Jensen-Shannon Divergence 42
4.6 Guideline Alignment Methodology 43
4.6.1 Venue Selection and Guideline Corpus Construction 43
4.6.2 Thematic Mapping and Gap Analysis 44
4.6.2.1 Directive Extraction and Semantic Chunking 45
4.6.2.2 Deductive Topic Assignment 45
4.6.2.3 Human-in-the-Loop Validation 45
5 ANALYSIS AND RESULTS 47
5.1 The Schema of Reviewer Inquiries 47
5.1.1 Semantic Definitions and Scope 47
5.1.2 Frequency Distribution and Operational Priorities 49
5.2 Analysis of Semantic Uncertainty 50
5.2.1 Quantification of Reviewer Ambiguity 50
5.2.2 Topic Separability and Structural Overlap 53
5.3 The Guideline-Practice Gap 53
5.3.1 Consensus and Divergence Across Venues 53
5.3.2 Alignment Analysis: Prescribed vs. Practiced 55
5.3.2.1 Unwritten Norms 55

5.3.2.2 Neglected Criteria 56
5.3.3 Guideline-Exclusive Categories 56
6 DISCUSSION 59
6.1 The Hierarchy of Reviewer Priorities 59
6.2 Structural Coherence and Model Behavior 60
6.2.1 Semantic Interdependence of Topics 60
6.2.2 The Influence of RLHF on Uncertainty 61
6.3 Contextualization within the Peer Review Literature 61
6.4 The Guideline–Practice Gap 62
6.4.1 Implicit Norms Beyond Formal Guidelines 62
6.4.2 Institutional Lag and Emergent Criteria 62
7 VALIDITY THREAT MITIGATION 65
7.1 Construct Validity 65
7.2 Internal Validity 66
7.3 External Validity 67
7.4 Reliability and Reproducibility 67
8 CONCLUSION 69
8.1 Research Summary 69
8.2 Synthesis of Findings 69
8.2.1 The Schema of Criticism (RQ1) 70
8.2.2 The Guideline-Practice Gap (RQ2) 70
8.3 Implications 71

8.4 Limitations and Future Works 71

8.5 Concluding Remarks 72

REFERENCES 73

APPENDICES

A DATA SOURCE AND CONTEXT 79

A.1 OpenReview Interface 79

A.2 Scoring Rubrics and Definitions 81

A.3 Administrative Comments 82

B EXTENDED STATISTICAL ANALYSIS 83

B.1 Supplementary EDA Plots for Quantitative Metrics 83

B.1.1 Review-Level Score Distributions 83

B.1.2 Submission-Level Aggregated Metrics 86

B.2 Additional EDA Plots for Textual Metrics 88

B.3 Statistical Significance Table 90

B.4 Cosine Similarity Statistics 90

B.5 LLM-as-a-Judge Agreement 93

B.5.1 Distribution of Agreement Levels 93

B.5.2 Individual Judge Verdicts 93

B.6 Validation Dataset Processing Pipeline 94

C TOPIC MODELING RESULTS 97

C.1 Topic Generation Outputs 97

C.2 Qualitative Examples of Reviewer Inquiries 97

D LLM PROMPT TEMPLATES FOR DATA PROCESSING AND VALIDATION . . . 115

D.1 Non-Informative Text Filtering 115

D.2 Semantically Coherent Question Chunking 117

D.3 LLM as a Judge for Question Separation 119

D.4 TopicGPT Prompt for Topic Generation 122

D.5 Probabilistic Topic Assignment 124

D.6 Reviewer Guideline Topic Assignment 124

LIST OF TABLES

Table 1 Comparison of submission volumes for ICLR 2024 and ICLR 2025. 2

Table 2 Volume of data artifacts collected in MongoDB (ICLR 2024). 13

Table 3 Disposition of all ICLR 2024 submissions. 15

Table 4 Verification of data licensing for key user-generated artifacts. 15

Table 5 Descriptive statistics of reviewer workload (Reviews per Reviewer). 16

Table 6 Descriptive statistics for review count per submission. 17

Table 7 Frequency distribution of individual rating scores. 18

Table 8 Descriptive statistics of the averaged review ratings per submission. 19

Table 9 Descriptive statistics for all quantitative metrics in individual reviews. 21

Table 10 Descriptive statistics for average scores per submission. 21

Table 11 Descriptive statistics for word counts of review text fields. 22

Table 12 Review counts per rating category across the sequential sampling stages. 29

Table 13 Selected top-tier venues for guideline analysis. 44

Table 14 The consolidated schema of reviewer inquiries. 48

Table 15 Frequency distribution of reviewer inquiries. 50

Table 16 Summary statistics for probabilistic uncertainty metrics. 52

Table 17 Validation statistics for automated guideline topic assignment. 55

Table 18 Alignment of data-driven topics with official reviewer guidelines across top-tier venues. 57

Table 19 Verbatim descriptions for ICLR 2024 confidence scores. 81

Table 20 Verbatim descriptions for ICLR 2024 rating scores. 81

Table 21 Shared scoring rubric for soundness, presentation, and contribution for ICLR 2024. 81

Table 22 Kendall's Tau correlation p-values (N=22,358). Non-significant values are bolded. 91

Table 23 Descriptive statistics for cosine similarity scores of extracted question chunks. 92

Table 24 Distribution of agreement levels across 802 judged question chunks. 93

Table 25 Total *valid* and *invalid* verdicts cast by each LLM judge. 93

Table 26 Topic generation output - Run 1 (Model: GPT-4o, Data: Subset 1, $n = 403$). . 98

Table 27 Topic generation output - Run 2 (Model: Gemini 2.5 Flash, Data: Subset 2, $n = 403$). 99

Table 28 Topic generation output - Run 3 (Model: GPT-4o, Data: Full Dataset, $N = 760$). 100

Table 29 Representative examples for the *Inquiry About Methodology* topic, stratified by confidence level. 101

Table 30 Representative examples for the *Inquiry About Comparative Analysis* topic, stratified by confidence level. 102

Table 31 Representative examples for the *Inquiry About Applicability and Limitations* topic, stratified by confidence level. 103

Table 32 Representative examples for the *Inquiry About Theoretical Foundations* topic, stratified by confidence level. 104

Table 33 Representative examples for the *Inquiry About Terminology and Clarity* topic, stratified by confidence level. 105

Table 34 Representative examples for the *Inquiry About Presentation and Figures* topic, stratified by confidence level. 106

Table 35 Representative examples for the *Inquiry About Interpretability and Explanation* topic, stratified by confidence level. 107

Table 36 Representative examples for the *Inquiry About Performance and Metrics* topic, stratified by confidence level. 108

Table 37 Representative examples for the *Inquiry About Validity and Reproducibility* topic, stratified by confidence level. 109

Table 38 Representative examples for the *Inquiry About Data* topic, stratified by confidence level. 110

Table 39 Representative examples for the *Inquiry About Motivation and Contribution* topic, stratified by confidence level. 111

Table 40 Representative examples for the *Inquiry About Computational Efficiency* topic, stratified by confidence level. 112

Table 41 Representative examples for the *Inquiry About Ethical Considerations* topic, stratified by confidence level. 113

LIST OF FIGURES

Figure 1 Distribution of official comments' authors by role. 16

Figure 2 Distribution of the number of reviews per submission. 18

Figure 3 Distribution of individual rating scores across all reviews. 19

Figure 4 Distribution of average rating scores per submission. 20

Figure 5 Kendall's Tau (τ_b) correlation matrix for review features. 24

Figure 6 Sankey diagram illustrating the data flow from the full ICLR 2024 corpus to the stratified Main Dataset and the Validation Dataset. 30

Figure 7 Comparative distribution of review ratings. 31

Figure 8 Cohen's Kappa inter-rater agreement matrix for the three LLM judges. . . 36

Figure 9 The complete data processing pipeline for the Main Dataset. 37

Figure 10 Frequency distribution of reviewer inquiries across the 13 generated topics. 49

Figure 11 Boxplot of Probability Margin per topic on Main Dataset. 51

Figure 12 Jensen-Shannon Divergence (JSD) heatmap of topic centroids. Values closer to 1 (darker) indicate higher semantic separability, while values closer to 0 (lighter) indicate greater overlap. 54

Figure 13 Screenshot of a submission page on OpenReview. 79

Figure 14 Screenshot of a review on OpenReview. 80

Figure 15 Distribution of confidence scores across individual reviews. 83

Figure 16 Distribution of soundness scores across individual reviews. 84

Figure 17 Distribution of presentation scores across individual reviews. 84

Figure 18 Distribution of contribution scores across individual reviews. 85

Figure 19 Distribution of average confidence scores per submission. 86

Figure 20 Distribution of average soundness scores per submission. 86

Figure 21 Distribution of average presentation scores per submission. 87

Figure 22 Distribution of average contribution scores per submission. 87

Figure 23 Distribution of word counts for the *Strengths* field in reviews. 88

Figure 24 Distribution of word counts for the *Weaknesses* field in reviews. 89

Figure 25 Distribution of word counts for the *Questions* field in reviews. 89

Figure 26 Frequency distribution of reviewer inquiries across the 13 generated topics on Validation Dataset. 94

Figure 27 Data processing flow for the Validation Dataset. 95

Figure 28 Boxplot of Probability Margin per topic on Validation Dataset. 96

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
API	Application Programming Interface
ARR	ACL Rolling Review
CV	Computer Vision
ICL	In-Context Learning
ICLR	International Conference on Learning Representations
IRR	Inter-Rater Reliability
JSD	Jensen-Shannon Divergence
LDA	Latent Dirichlet Allocation
LLM	Large Language Model
ML	Machine Learning
NLP	Natural Language Processing
NPMI	Normalized Pointwise Mutual Information
RLHF	Reinforcement Learning from Human Feedback
SOTA	State of the Art

CHAPTER 1

INTRODUCTION

Academic publishing is the fundamental mechanism for the dissemination, validation, and archival of scientific knowledge. Peer review is the main process driving this mechanism. It serves as an important quality control system where independent experts scrutinize a manuscript's qualities, including its validity, significance, and originality. The primary objective is to filter out flawed methodologies and ensure that published work is sound and contributes meaningfully to the existing literature.

To maintain the integrity of this evaluation, several reviewing models have been established, and they are primarily distinguished by their *identity transparency* [1] levels. The most common models are single-blind and double-blind review models. In a single-blind review, reviewers know the authors' identities, which carries a risk of bias based on reputation or affiliation. In contrast, the double-blind review conceals the identities of both authors and reviewers from one another. This method is widely regarded as the gold standard for reducing implicit bias and ensuring that papers are judged solely on their scientific merit. Consequently, it has become the dominant format in top-tier computer science and Artificial Intelligence (AI) venues to promote fairness in a highly competitive landscape.

The exponential growth in research output has led to a shortage of qualified reviewers, a phenomenon well-documented in recent studies [2, 3]. This surge is particularly severe in top-tier venues. As illustrated in Table 1, the volume of submissions to the International Conference on Learning Representations (ICLR)[1] has reached historic highs, placing immense pressure on the peer review ecosystem and resulting in a severe shortage of senior reviewers. The selection of ICLR as the focal point of this study is motivated by its high impact in the field, evidenced by its Google Scholar ranking as the third most influential publication venue in the *Engineering & Computer Science* category, with an h5-index of 362 and an h5-median of 652 [4].

To manage this scale, conference organizers are compelled to recruit a large cohort of junior reviewers to supplement the limited pool of senior experts. While this expansion is necessary to handle the volume of submissions, it introduces significant variability in the quality and depth of feedback. To mitigate this inconsistency and ensure high standards, venues have developed comprehensive reviewer guidelines. These documents are frequently updated to reflect the community's current needs, evolving research standards, and specific focus areas for each iteration of the conference. These guidelines serve as an important standard for the

[1] `https://iclr.cc/Conferences/2024`

Table 1: Comparison of submission volumes for ICLR 2024 and ICLR 2025.

Metric	**ICLR 2024**	**ICLR 2025**
Accepted Papers (all acceptance types)	2,261	3,708
Rejected Papers (excluding desk-rejected)	3,519	5,019
Total Decided Submissions	**5,780**	**8,727**

increasingly heterogeneous reviewer pool, explicitly instructing reviewers on what constitutes a high-quality evaluation.

However, the alignment between prescribed criteria and actual reviewer practice remains a largely unexamined empirical question. This study specifically investigates this gap by analyzing the *Questions* field of reviewer feedback to identify latent topics and evaluate their alignment with prescribed criteria such as reproducibility.

The analysis of peer review data through Natural Language Processing (NLP) has become an active area of research, with efforts primarily focused on predictive tasks or quality assessment. For instance, the introduction of the PeerRead dataset facilitated tasks such as predicting paper acceptance from review text [5], while other work has focused on automatically evaluating the quality of the reviews themselves [6]. Methodologically, these analyses have evolved alongside the field of topic modeling. Early approaches often relied on traditional probabilistic models, which struggled to capture deep semantic meaning. The advent of contextual embeddings led to more coherent models like BERTopic [7], which uses transformers to generate richer topic representations. Most recently, the field has shifted towards directly using Large Language Models (LLMs) for topic generation, with frameworks like TopicGPT [8] demonstrating the ability to produce higher-quality, more interpretable topics that align better with human judgment.

The focus of prior works had been mostly on the predictive power or the intrinsic quality of the review text. To our knowledge, no prior work has systematically investigated the alignment between the emergent topics of reviewer feedback questions and the explicit criteria of institutional reviewer guidelines. This thesis addresses this gap through a data-driven analysis of peer reviews from the ICLR 2024 conference. To facilitate this investigation, the dataset was sourced from `OpenReview`[2] platform.

1.1 Research Questions

The main objective of this study is to empirically investigate the latent topics within academic peer review, focusing specifically on the *Questions* field of the reviews associated with ICLR 2024 submissions on the OpenReview platform. By employing a data-driven approach, this research first identifies the underlying topics and subsequently examines the alignment between these reviewer inquiries and the formal institutional guidelines.

[2] `https://openreview.net/`

The research questions that guide this study are as follows:

Research Question 1: What are the main topics that receive the highest degree of criticism in peer reviews?

- **RQ1.1:** What are the defining semantic properties, distributional patterns, and representative examples of each identified theme?
- **RQ1.2:** How can the semantic separability of the generated topics be quantitatively assessed to validate the coherence and structure of the topic model?

Research Question 2: To what extent do the data-driven topics of reviewer criticism align with the official reviewing guidelines of top-tier venues?

- **RQ2.1:** What are the primary topics of criticism that emerge from reviewer feedback, but are not explicitly requested or detailed in the formal guidelines?
- **RQ2.2:** Which explicitly requested evaluation criteria in the formal guidelines are systematically de-prioritized or ignored in the actual feedback provided by reviewers?

1.2 Contributions of the Study

The contributions of this thesis are as follows:

1. *A Source-Grounded Framework for Question Extraction and Soft-Classification:* We propose a multi-stage framework designed to extract semantically coherent inquiries, generate latent topics, and perform soft classification on unstructured peer-review text. Unlike traditional topic modeling or hard-classification approaches, this framework leverages the probabilistic nature of Large Language Models (LLMs) to quantify semantic uncertainty. This approach enforces strict source-grounding to mitigate hallucinations and ensure the extraction of coherent question units. Furthermore, it introduces granular metrics, specifically Probability Margin and Normalized Entropy, to measure the ambiguity of reviewer intent. This allows for a nuanced distinction between clearly defined inquiries and multi-faceted, ambiguous feedback, providing a more accurate representation of the review discourse.
2. *Empirical Quantification of the Guideline-Practice Gap:* To our knowledge, this study provides the first systematic comparison between the *prescribed* criteria of top-tier venue guidelines and the *practiced* criteria of actual reviewers using the *Questions* field within the ICLR 2024 dataset. We identify specific thematic areas where practice diverges from policy. This analysis reveals both the explicitly requested criteria that reviewers neglect or de-prioritize, and the emergent community norms that are currently absent from formal guidelines.

1.3 Organization of the Thesis

The remainder of this thesis is organized as follows:

Chapter 2 reviews the evolution of computational peer review analysis and topic modeling methodologies, tracing the shift from probabilistic models to generative LLM frameworks. It concludes by identifying the research gap concerning the alignment between reviewer inquiries and institutional guidelines.

Chapter 3 details the construction of the ICLR 2024 corpus, from the programmatic acquisition of data via the OpenReview API to the implementation of a reproducible storage architecture using MongoDB and Hydra. It then presents a comprehensive exploratory analysis of submission dynamics, reviewer metrics, and statistical correlations between evaluative scores and textual feedback.

Chapter 4 details the experimental framework, encompassing stratified sampling, source-anchored information extraction, and multi-agent consensus validation. It subsequently defines the generative topic discovery process and the *Logit-Based Soft Classification* technique used to quantify semantic uncertainty and assess guideline alignment.

Chapter 5 reports the empirical findings, presenting a 13-topic schema of reviewer inquiries and its frequency distribution. It then assesses the structural coherence of the schema using probabilistic metrics and concludes by quantifying the *Guideline-Practice Gap*, revealing the divergence between practiced and prescribed review criteria.

Chapter 6 interprets the empirical results, examining the hierarchy of reviewer priorities and the influence of model behavior on semantic structure. It subsequently analyzes the *Guideline-Practice Gap* attributing the observed divergence to the reliance on implicit community norms and the distinct functional roles of institutional policy versus active critique.

Chapter 7 addresses the methodological rigor of the study by systematically evaluating validity threats. It details the mitigation strategies employed to safeguard construct and internal validity, including source-anchored extraction and stratified sampling, while defining the boundaries of external generalizability and reproducibility.

Chapter 8 synthesizes the research findings, confirming the divergence between institutional guidelines and reviewer practice. It discusses the broader implications of this gap for conference stakeholders, acknowledges the study's limitations, and outlines future research directions for automated peer review analysis.

CHAPTER 2

RELATED WORK

This chapter presents the existing literature relevant to the contributions of this thesis. We first examine prior efforts in constructing peer review datasets and the integration of AI into the review ecosystem. Second, we trace the methodological evolution of topic modeling, highlighting the shift from probabilistic models to Large Language Model (LLM) based approaches. Finally, we discuss existing research regarding reviewer guidelines and the evaluation of review quality, identifying the gap between institutional policy and reviewer practice.

2.1 Computational Analysis of Peer Reviews

The analysis of academic peer review has been significantly advanced by the availability of structured datasets and computational methods. Staudinger et al. [3] present common NLP tasks applied to the peer review domain. These tasks include acceptance prediction and review generation. Beyond early corpora such as PeerRead [5], the field has seen the development of specialized resources. Recent work includes large-scale multidisciplinary datasets [9] and specialized taxonomies for comparison discussions [10].

Initial research in this area often treated paper acceptance as a predictive task. These studies used the review text to predict the acceptance decision of submissions. However, the field has evolved to incorporate more nuanced techniques. Methods such as aspect-based sentiment analysis of scientific reviews are now being used to achieve more accurate predictions [11], and deep neural architectures such as DeepSentiPeer [12] are employed to predict the final decision. They also predict recommendation scores by using sentiment features extracted from reviews [12]. Furthermore, researchers are using cross-attention mechanisms between the review and the manuscript content. This allows models to mimic the complex interactions that inform the decisions of editors and area chairs [13].

Despite these advancements, researchers have identified limitations in traditional sentiment analysis. Studies demonstrate that the distribution of aspect-based sentiment varies significantly between accepted and rejected papers, suggesting that aggregate sentiment scores fail to capture the full picture of scientific critique [11]. To facilitate deeper analysis, new datasets with multi-layered annotations have been developed. These resources categorize review text by the specific paper section and the aspect being critiqued, such as novelty or soundness [6]. By scrutinizing the thematic content of reviews through these granular annota-

tions, researchers can uncover latent biases missed by surface-level sentiment, paving the way for a more equitable and transparent peer review process.

Most recently, the focus has expanded to the application of Large Language Models (LLMs) within the peer review ecosystem. Surveys have mapped how AI can support research assessment and publishing [14, 15]. These works have progressed from theoretical inquiries about automation to empirical evaluations of LLM performance. For instance, Yuan et al. [16] investigated the feasibility of using NLP models to generate scientific paper reviews. Studies also assessed biases and factual reliability of LLMs when they are tasked with generating reviews [17]. More recently, researchers have begun using LLM-based agents to simulate the review ecosystem. This allows for controlled experiments on review dynamics like discussion and consensus-building [18].

2.2 Evolution of Topic Modeling Approaches

Researchers have long relied on topic modeling to uncover latent topics within unstructured review text. However, the methodology for this task has evolved significantly, transitioning from traditional approaches based on word frequency and co-occurrence statistics to modern methods utilizing semantic embeddings and generative frameworks. This shift addresses the need for more coherent and interpretable topics, particularly when analyzing the short, specialized, and often noisy text characteristic of academic peer reviews.

For over a decade, Latent Dirichlet Allocation (LDA) [19] served as the standard for unsupervised topic discovery [20]. In the context of scientific literature, LDA has been widely used to map research trends and categorize submissions. However, as a probabilistic model, LDA operates on a *bag-of-words* assumption, treating documents as collections of tokens without considering their order or context. This simplification is a major drawback, as it ignores crucial semantic information. Consequently, LDA's performance degrades significantly on short texts, such as individual reviewer comments, where the sparsity of word co-occurrence patterns makes it difficult to infer meaningful topics [21].

The advent of powerful contextual embeddings from models like BERT [22] marked a significant advancement. These advancements lead to the development of clustering-based topic models such as BERTopic [7]. This approach first generates dense vector representations for each document, capturing the semantic meaning in a high-dimensional space. These vectors are then clustered using algorithms such as HDBSCAN [23] to group semantically similar documents. The key advantage is the ability to capture contextual nuances that probabilistic models miss, leading to more coherent topic clusters [7]. However, a persistent challenge with these models is topic representation, especially on short texts such as reviewer comments. The output is typically a list of keywords for each cluster, which requires subjective human interpretation to assign a clear and meaningful definition [7].

The most recent and major advancement in this field involves leveraging the generative and reasoning capabilities of Large Language Models (LLMs). Frameworks such as TopicGPT have pioneered this approach, moving beyond clustering methods to active generation [8]. Unlike previous methods, these models utilize the extensive world knowledge and linguistic under-

standing of LLMs. That way, these frameworks not only identify topics but are also capable of generating high-quality, human-readable labels and their detailed definitions. This is often achieved through various prompting strategies, including zero-shot and few-shot approaches, which guide the LLM to produce topics that are coherent, distinct, and comprehensive.

Despite these generative advancements, evaluating the quality of topic models remains a fundamental challenge [24]. Traditional automated metrics like coherence scores (e.g., Normalized Pointwise Mutual Information (NPMI)) have been shown to be inconsistent and often poorly correlated with human judgment, a problem highlighted by studies questioning if automated evaluation is fundamentally broken [25, 26]. On the other hand, human evaluation, while considered the gold standard for assessing topic quality, is expensive, time-consuming, and difficult to scale, which limits its use in large-scale experiments [25]. To address this, recent research and this study has adopted the *LLM-as-a-Judge* framework, where LLMs serve as scalable proxies for human judgment to assess topic quality along multiple criteria [27].

2.3 Limitations in Current Peer Review Analysis

While the computational analysis of peer review content is well-established, the systematic comparison of this content against institutional guidelines remains a largely unexamined domain. Existing studies have mainly focused on intrinsic review quality, predictive outcomes, or system-level effects of review policies. This section presents prior work on review quality assessment and outlines methodological precedents.

The assessment of review quality has been approached through various criteria, including helpfulness, constructiveness, and informativeness [28, 29]. Helpfulness is defined by whether a review provides useful insights for the author, while constructiveness relates to the provision of actionable feedback [29]. To automate this assessment, researchers have employed techniques such as argument mining to identify and extract the core argumentative content from review text, which is crucial for decision-making [30, 31]. However, many of the tools developed to measure review quality have not undergone rigorous validation, and the concepts they evaluate vary widely, limiting their reliability and applicability [29].

The discrepancy between formal guidelines and reviewer practices can be attributed to systemic challenges within the academic publishing ecosystem. The peer review process involves multiple stakeholders, namely authors, conference organizers, and reviewers. Each stakeholder has differing objectives, which can lead to misaligned incentives and policies [32]. The sheer volume of submissions places a significant burden on reviewers, which may result in lower-quality feedback and a deviation from comprehensive evaluation criteria [32].

To the best of our knowledge, despite extensive work on review quality and reviewer behavior, no prior work uses reviewer inquiries as a basis for examining a guideline–practice gap in peer reviews. This thesis addresses this gap by generating latent topics within reviewer inquiries from the ICLR 2024 conference. Our investigation contains two stages. First, we evaluate the alignment of these generated topics with the institutional reviewer guidelines of ICLR 2024. Second, to understand broader community norms, we conduct a comparative analysis against the evaluation criteria of ten highly reputable AI venues.

CHAPTER 3

DATASET AND EXPLORATORY DATA ANALYSIS

This study is based on a corpus constructed from the peer review data of the 2024 International Conference on Learning Representations (ICLR), hosted on the OpenReview platform. Developed at the University of Massachusetts Amherst, OpenReview is an open-source scientific platform with features such as transparent peer review and open access publishing [33]. The dataset was programmatically collected via authorized requests using the official OpenReview Application Programming Interface (API), which provided a comprehensive capture of all publicly available artifacts for each submission. This raw data, mirroring the content of the submission pages in OpenReview, was then systematically processed and filtered to create the final curated corpus for this study.

3.1 Data Acquisition

3.1.1 The OpenReview Platform as a Data Source

OpenReview is a web-based system designed to manage the entire lifecycle of the academic peer review process, from paper submission and reviewer assignment to rebuttals and the publication of final decisions. The main objective of the platform is to increase transparency and openness in scientific evaluation. The suitability of this platform is as follows:

Transparency: The platform provides direct access to the complete peer review lifecycle, including initial reviews, author rebuttals, and final decisions.

Data Structure: OpenReview employs a highly structured data model where submissions, reviews, rebuttals, and decisions are organized as discrete, machine-readable entities with associated metadata. Visual examples of a submission and a review, which illustrate these structured fields, are provided in Figures 13 and 14 in Appendix A.1. This structured format is essential for the programmatic analysis of review components and conversational threads.

Accessibility: The provision of a public Application Programming Interface (API) allows for systematic and reproducible data acquisition, a fundamental requirement for computational research.

Platform Policies: OpenReview's Terms of Use [34] and Privacy Policy [35], at the time of data acquisition, establish a clear and supportive environment for the analysis of peer

review data. These policies specify that all submitted "comments", which are a category that includes reviews, rebuttals, and meta-reviews, are released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license.[1] This license grants the public the right to share and adapt the material for any purpose, including the computational analysis conducted in this thesis, provided appropriate attribution is maintained. Furthermore, the "Metadata Dedication" clause places all metadata associated with a submission to the public domain under a Creative Commons Public Domain Dedication (CC0 1.0).[2] This applies to structured data fields such as presentation and confidence scores, removing any licensing restrictions on their use in quantitative analysis.

3.1.2 Selection of the ICLR 2024 Venue

Within the OpenReview ecosystem, ICLR was chosen as the specific focus of this study. As a top-tier conference in the field of artificial intelligence, ICLR attracts a high volume of quality submissions, ensuring a large and representative dataset. The 2024 iteration was selected because it was the most recent to complete its full review cycle at the time of data acquisition.

However, the principal motivation for this choice was the organizers' newly introduced modification to the review process: a highly structured, multi-part review form. This updated format mandates that reviewers provide feedback across distinct, predefined sections, including *Strengths*, *Weaknesses*, and, most critically for this work, a dedicated *Questions* section.

The explicit *Questions* section provides a cleaner corpus for investigating the actionability of reviewer comments and the subsequent author-reviewer dialogue. Furthermore, the conference's double-blind review policy and its open rebuttal phase create a suitable corpus for studying the dynamics of scholarly discourse.

3.2 Dataset Construction

The dataset for this analysis was constructed from scratch, as no pre-existing corpus met the specific requirements of this research. To this end, a custom data acquisition and storage pipeline was developed. Raw data artifacts were programmatically collected from the OpenReview API using a custom abstraction layer built upon the official `openreview-py`[3] client. These artifacts were subsequently parsed, validated against a predefined schema, and systematically ingested into a `MongoDB`[4] database. This structured approach ensures data integrity from the outset and facilitates the complex querying and filtering procedures detailed in the following sections.

[1] https://creativecommons.org/licenses/by/4.0/
[2] https://creativecommons.org/publicdomain/zero/1.0/
[3] https://github.com/openreview/openreview-py
[4] https://www.mongodb.com/

3.2.1 Reproducibility and Pipeline Management

To ensure the scientific rigor and reproducibility of the data construction process, a modern software development toolchain was employed. The entire codebase is version-controlled using `Git`[5].

Execution Environment: To ensure a consistent and isolated execution context, all code was developed and run within a dedicated `Conda`[6] environment with Python 3.10 interpreter. By isolating the project's dependencies from system-level packages, this approach eliminates potential version conflicts and ensures that the research can be replicated within an identical software context.

Dependency Management: For fully reproducible software environment creation, the project utilized `Poetry`[7] for dependency management. All project dependencies were declared in the `pyproject.toml` file by specifying required packages and their version constraints. During installation, Poetry resolves the complete dependency graph and generates a `poetry.lock` file. Since this lock file records the exact version of every package, it ensures complete environment reproducibility.

Credential Management: To ensure the secure handling of sensitive information, such as OpenReview credentials and access tokens for cloud-based LLM services, a dedicated credential management strategy was implemented. All secrets were stored as key-value pairs in a local `.env` file. The `python-dotenv`[8] package was then used to load these variables into the application's environment at runtime. Crucially, the `.env` file is explicitly listed in the project's `.gitignore` file, ensuring that no sensitive credentials are ever committed to the version control history. This practice effectively decouples secrets from the codebase, enhancing security and preventing their accidental exposure.

Configuration Management: Distinct from credential management, all non-sensitive experimental parameters were managed by `Hydra`[9]. This includes settings such as database connection details, embedding model names, and LLM model names from various vendors. Hydra enables these parameters to be defined in structured YAML files, completely decoupling them from the application source code. This approach facilitates organized, auditable, and easily composable experimental setups, significantly improving the traceability and reproducibility of each data collection and analysis run.

Data and Artifact Storage: A `MongoDB` database was employed as the central repository for all collected data. To ensure data integrity and maintain a clean separation of concerns, all database interactions were channeled through a custom-built abstraction layer. This layer integrates `Pydantic`[10] models to perform strict schema validation on all data artifacts before they are committed to the database, guaranteeing that only well-formed documents are stored.

---

[5] https://git-scm.com/
[6] https://anaconda.com/docs/getting-started/miniconda/main
[7] https://python-poetry.org/
[8] https://github.com/theskumar/python-dotenv
[9] https://github.com/facebookresearch/hydra
[10] https://github.com/pydantic/pydantic

For environmental consistency and portability, the database itself was deployed within a `Docker`[11] container.

3.2.2 The OpenReview Data Model

The OpenReview platform organizes scientific discourse through a flexible, object-oriented architecture. Understanding this schema is important for the accurate parsing of review artifacts and the reconstruction of the peer review dialogue. It is important to note that OpenReview operates two distinct API versions. As the ICLR 2024 venue utilizes API v2, all data acquisition logic in this study was engineered to comply with this specification.

3.2.2.1 The "Note" Object and Relational Aggregation

The fundamental unit of data in OpenReview is the Note. This object serves as the universal container for all artifacts in the peer review process. Consequently, paper submissions, reviews, rebuttals, and decisions are all stored as Note objects. Although these objects contain numerous metadata fields, our extraction pipeline relies on two main identifiers to reconstruct the components:

- **`id`:** A globally unique string identifying the specific Note.
- **`forum`:** The universal identifier for a submission's thread. While every data object possesses a unique `id`, all artifacts, such as decisions and reviews, associated with a specific submission share a common `forum` value. Within our data architecture, this attribute serves as the main aggregation key, allowing the custom database wrapper to programmatically unify these disparate components into a single, cohesive submission record.

3.2.3 Data Ingestion and Schema Validation

A systematic ingestion pipeline was implemented to convert the raw API responses into a queryable corpus. Data was retrieved via authenticated requests to ensure access to all public metadata. Before storage, all incoming data passed through a validation layer constructed using *Pydantic*. This step enforced strict typing and structural integrity, ensuring that all retrieved records were correctly parsed and adhered to the defined schema.

3.2.3.1 Feature Parsing and Normalization

The ICLR 2024 uses composite strings instead of numeric types for quantitative fields in reviews. This approach simplifies interface display, but necessitates an additional preprocessing step for computational analysis. Consequently, metrics such as *rating* are not represented as

[11] `https://www.docker.com/`

integers but as composite strings that concatenate the numeric score with a descriptive label, as illustrated in Figure 14 (e.g., *"10: Strong accept..."*).

To enable statistical analysis, composite string fields were parsed during the ingestion phase. A normalization routine extracted the numerical prefix and cast it to a numerical format. This transformation ensures that ratings and other composite fields are treated strictly as a quantitative value in the database.

3.2.3.2 Database Structure

The validated data was stored in a MongoDB database, organized into distinct collections to mirror the logical components of the review process. Table 2 details the comprehensive inventory of collected artifacts of the ICLR 2024 venue.

To provide context on the specific data types, the `official_comments` collection captures the rebuttal phase, comprising the discourse between authors and venue officials regarding the submissions and reviews. Similarly, the `meta_reviews` collection contains the assessments by Area Chairs, who present a final recommendation to the Program Chairs based on the initial reviews and subsequent rebuttals.

While the acquisition pipeline captured the full breadth of the conference data, including meta-reviews and comments, the analysis presented in this thesis is derived exclusively from the `submissions`, `decisions`, and `reviews` collections.

Table 2: Volume of data artifacts collected in MongoDB (ICLR 2024).

Collection Name	**Count**
`submissions`	7,404
`accepted_submissions`	2,260
`withdrawn_submissions`	1,656
`desk_rejected_submissions`	53
`decisions`	5,780
`reviews`	28,028
`meta_reviews`	5,781
`official_comments`	70,624

3.2.4 Data Filtering and Integrity Verification

While the initial acquisition phase captured all 7,404 submissions to ICLR 2024, a significant portion of these did not complete the peer review cycle due to desk rejections or withdrawals. To ensure the analysis focused on completed peer review interactions, the dataset was filtered to include only submissions that possessed a final `decision` artifact. Within the OpenReview data model, the presence of this artifact serves as the definitive indicator that a submission has successfully traversed the full peer review lifecycle. Applying this filter resulted in a core set of 5,780 submission data for the analysis in Section 3.3.

The data integrity of the ingestion pipeline was verified through a manual audit of 20 randomly selected submissions. We cross-referenced the stored data against the official OpenReview submission pages (see Figures 13 and 14). The audit confirmed exact correspondence and validated the accuracy of the data extraction.

3.3 Exploratory Data Analysis

To characterize the ICLR 2024 peer review ecosystem, a comprehensive exploratory data analysis (EDA) was conducted on the curated corpus of 5,780 submissions that completed the full review cycle. This section investigates the distribution of final decisions, the nature of reviewer participation, and the statistical properties of quantitative assessment scores.

3.3.1 Submission Outcomes

The initial pool of submissions for ICLR 2024 comprised 7,404 manuscripts as mentioned in Table 2. However, a substantial fraction of these did not proceed through the entire peer-review lifecycle. As detailed in Table 3, 1,624 papers (21.9% of the total) were removed from consideration prior to receiving a final decision, with 1,571 being withdrawn by the authors and 53 being desk-rejected by the program committee. This filtering process resulted in a final corpus of 5,780 submissions that underwent full peer review.

Table 3 details the final disposition of all manuscripts submitted to ICLR 2024. Focusing on the cohort that completed the full peer-review cycle, the process remained highly selective with an overall acceptance rate of 39.12% and a rejection rate of 60.88%. In contrast to venues that require authors to select a specific presentation track (e.g., poster or oral) at the time of submission, ICLR employs a unified submission process. All manuscripts enter a single pool, and the distinction between oral, spotlight, and poster presentations is determined by the Program Chairs at the final decision stage.

It is noteworthy that the 'withdrawn' status is not exclusively limited to the pre-decision phase. A deeper analysis of the 1,656 submissions in the `withdrawn_submissions` collection reveals that 85 submissions were retracted by their authors after a final decision had been rendered. Among these, only one had received an *Accept (poster)* decision, while the remaining 84 had received a *Reject* decision. This indicates that authors occasionally retract their work even after a final verdict has been delivered.

3.3.2 Data Licenses

The legal and ethical foundation of this research is secured by OpenReview's permissive content policies. Table 4 presents the results of our empirical audit on the artifacts of 5,780 submissions. that underwent full peer review. The results unambiguously confirm that 100% of these user-generated artifacts are licensed under CC BY 4.0.

Table 3: Disposition of all ICLR 2024 submissions.

Manuscript Status	**Count**	**Percentage (%)**
Accept (oral)	86	1.16
Accept (spotlight)	367	4.96
Accept (poster)	1,808	24.42
Desk-Rejected	53	0.72
Reject	3,519	47.53
Withdrawn by Author (pre-decision)	1,571	21.22
Total	**7,404**	**100.00**

Table 4: Verification of data licensing for key user-generated artifacts.

Artifact Type	**Total Count**	**Verified (CC BY 4.0)**	**Percentage (%)**
Official Comments	66,066	66,066	100.00
Reviews	22,358	22,358	100.00
Decisions	5,780	5,780	100.00

A key distinction in OpenReview's licensing is between user-generated content and submission metadata. While the content is licensed under CC BY 4.0, the submission objects themselves lack an explicit license field. This is intentional, as the platform's policy dedicates all metadata to the public domain via the Creative Commons Public Domain Dedication (CC0 1.0). This dual-licensing framework provides an unambiguous legal basis for the computational analysis conducted in this thesis.

3.3.3 Reviewer and Community Participation

To understand the dynamics of the interactive discussion phase, we analyzed the authorship of the 66,066 `official comments` entries. These artifacts represent the conversational layer of peer review, encompassing author rebuttals, subsequent discussions between reviewers and authors, and general comments. As illustrated in Figure 1, this discourse is primarily driven by authors and reviewers. *Authors* contributed the largest volume of comments, confirming the discussion phase as a direct dialogue between a paper's proponents and its critics. On the other hand, *Area Chairs* were significantly less active, while *Program and Senior Area Chairs* contributed only a handful of comments, which were typically broad institutional announcements, as shown in Appendix A.3.

Separate from the analysis of the discussion phase, we profiled the workload distribution of the formal evaluation process by analyzing the authorship of the 22,358 reviews. The findings, summarized in Table 5, reveal an extreme imbalance. While the maximum number of reviews completed by any single individual was two, most of the reviewers gave only one review in ICLR 2024.

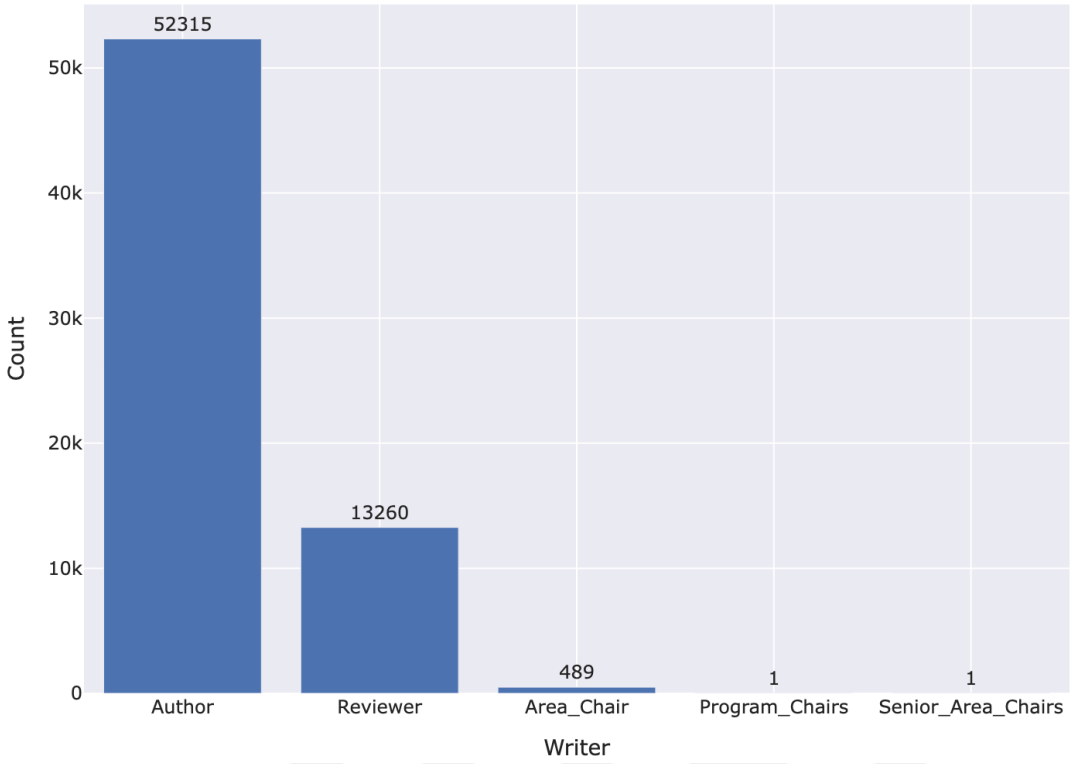


Figure 1: Distribution of official comments' authors by role.

A detailed calculation based on the total number of unique reviewers (22,337) and total reviews (22,358) precisely quantifies this imbalance. Specifically, an overwhelming majority of 22,316 reviewers (99.91%) contributed a single review, while a mere 21 individuals contributed two. This finding demonstrates that the ICLR 2024 reviewer pool was not a stable cohort of dedicated experts but rather a highly transient and heterogeneous group. Given that most reviewers only evaluate a single paper, it underscores the importance of clear, explicit guidelines to ensure evaluation consistency.

Table 5: Descriptive statistics of reviewer workload (Reviews per Reviewer).

Statistic	**Value**
Count (Unique Reviewers)	22,337
Mean	1.0009
Std. Dev.	0.0306
Min	1.00
25% (Q1)	1.00
50% (Median)	1.00
75% (Q3)	1.00
Max	2.00

3.3.4 Analysis of Quantitative and Textual Metrics

To characterize the review process in detail, we conducted a comprehensive analysis of the quantitative scores, the textual properties of the free-text fields, and the relationships between them.

3.3.4.1 Review Distribution per Submission

The number of evaluations per submission is a fundamental aspect of the review process. As summarized in Table 6, each submission of ICLR 2024 received a mean of approximately 3.87 reviews. While the minimum was two, some papers received as many as eight. The histogram in Figure 2 shows that each submission received at least two reviews. This multi-reviewer structure ensures that every submission is evaluated by a diverse panel, thereby mitigating individual reviewer bias and enhancing the robustness of the final decision.

Table 6: Descriptive statistics for review count per submission.

Statistic	**Value**
Count (Submissions)	5,780
Mean	3.87
Std. Dev.	0.62
Min	2
25% (Q1)	4
50% (Median)	4
75% (Q3)	4
Max	8

3.3.4.2 Quantitative Metrics

While the `rating` serves as the definitive metric representing the reviewer's overall recommendation, the evaluation framework also includes distinct metrics to assess specific attributes of the submission. This section details the analysis of the `rating` field, followed by the reviewer `confidence` and the three component quality metrics: `soundness`, `presentation`, and `contribution`.

Individual Review Ratings: The distribution of ratings across individual reviews, detailed in Table 7, reveals a distinct usage pattern. Although the scale nominally spans from one to ten, the data indicates that reviewers exclusively utilized the scores $\{1, 3, 5, 6, 8, 10\}$. The verbatim descriptions corresponding to these observed scores are provided in Table 20 (Appendix A.2). The intermediate integers $\{2, 4, 7, 9\}$ are absent from the dataset. This confirms that while the scale is presented as a 10-point scale, it functions in practice as a 6-point categorical scale.

As illustrated in Figure 3, the distribution exhibits a unimodal shape heavily concentrated around the acceptance threshold. The scores of five ("Marginally below") and six ("Marginally

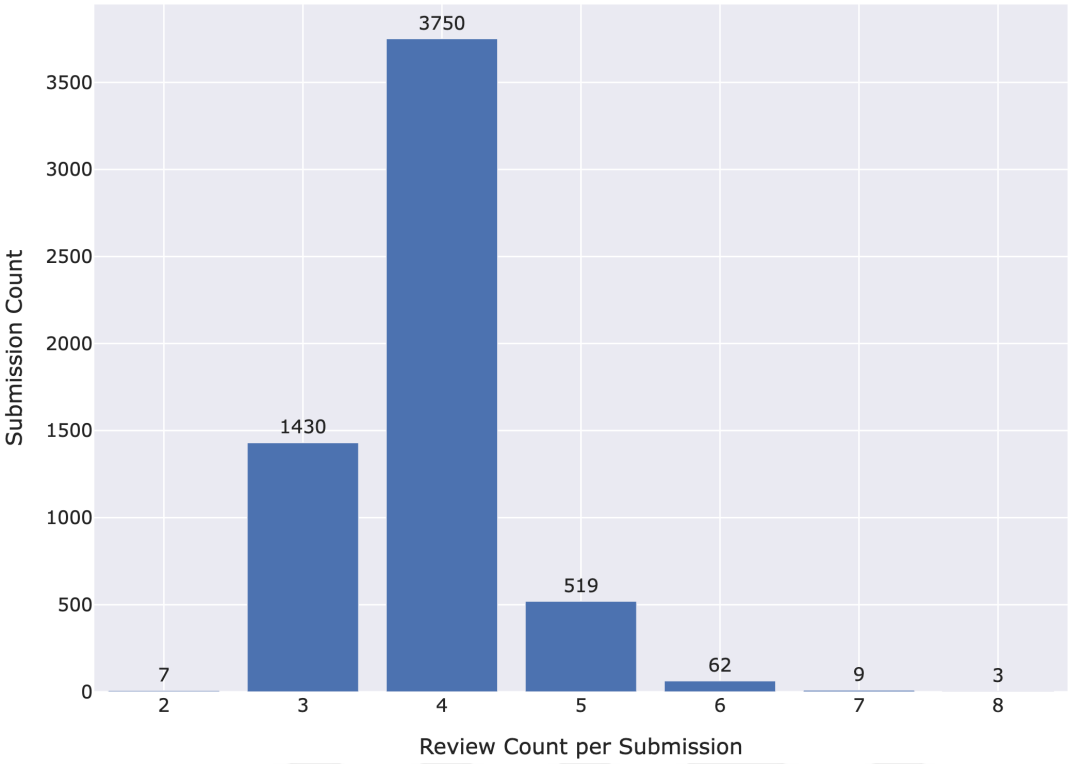


Figure 2: Distribution of the number of reviews per submission.

above") together account for 61.89% of all reviews, indicating that the majority of submissions are viewed as borderline cases. While the mode is six, the distribution displays a notable asymmetry in its tails. The cumulative volume of rejection-leaning scores (1 and 3) outweighs that of the clear acceptance scores (8 and 10). This disparity is most pronounced at the extremes, where strong endorsements (score 10) are significantly rarer than strong rejections (score 1). This reflects a reviewing culture that is more inclined to penalize flaws than to recognize perfection.

Table 7: Frequency distribution of individual rating scores.

Rating Score	Count	Percentage (%)
1	364	1.63
3	4,292	19.20
5	6,186	27.67
6	7,651	34.22
8	3,763	16.83
10	102	0.46
Total	**22,358**	**100.00**

Submission-Level Aggregated Review Ratings: In contrast to the discrete individual scores, the average rating per submission forms a continuous distribution, as illustrated in Figure 4.

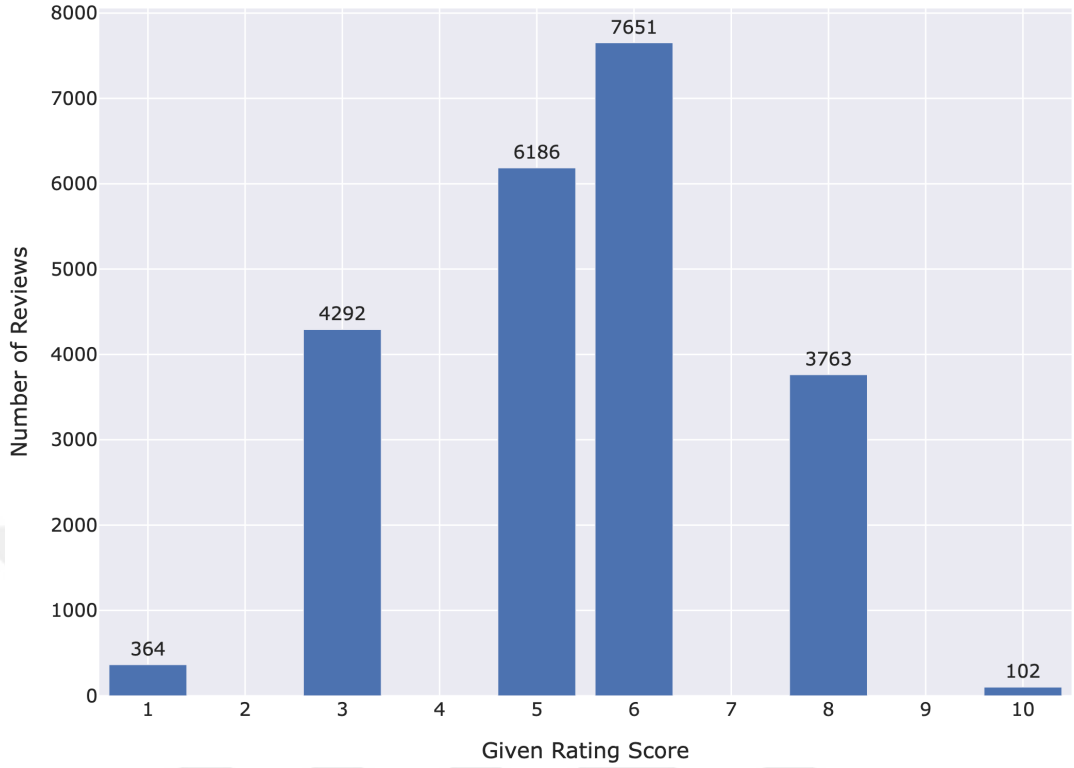


Figure 3: Distribution of individual rating scores across all reviews.

As detailed in Table 8, the aggregated metrics reveal a mean rating of 5.42 and a median of 5.50 across the 5,780 submissions. Notably, despite the presence of individual ratings of 10, the maximum average rating for any single submission is 9.00, indicating that no paper in the dataset achieved a unanimous "Strong Accept" consensus.

Table 8: Descriptive statistics of the averaged review ratings per submission.

Statistic	**Value**
Count (Submissions)	5,780
Mean	5.42
Std. Dev.	1.20
Min	1.00
25% (Q1)	4.67
50% (Median)	5.50
75% (Q3)	6.25
Max	9.00

Reviewer Confidence: The `confidence` score quantifies the reviewer's self-assessed expertise on a five-point scale, ranging from an admitted inability to assess the work (score 1) to absolute certainty in the evaluation (score 5). The complete, verbatim definitions for each level of this scale are provided in Table 19 (Appendix A.2).

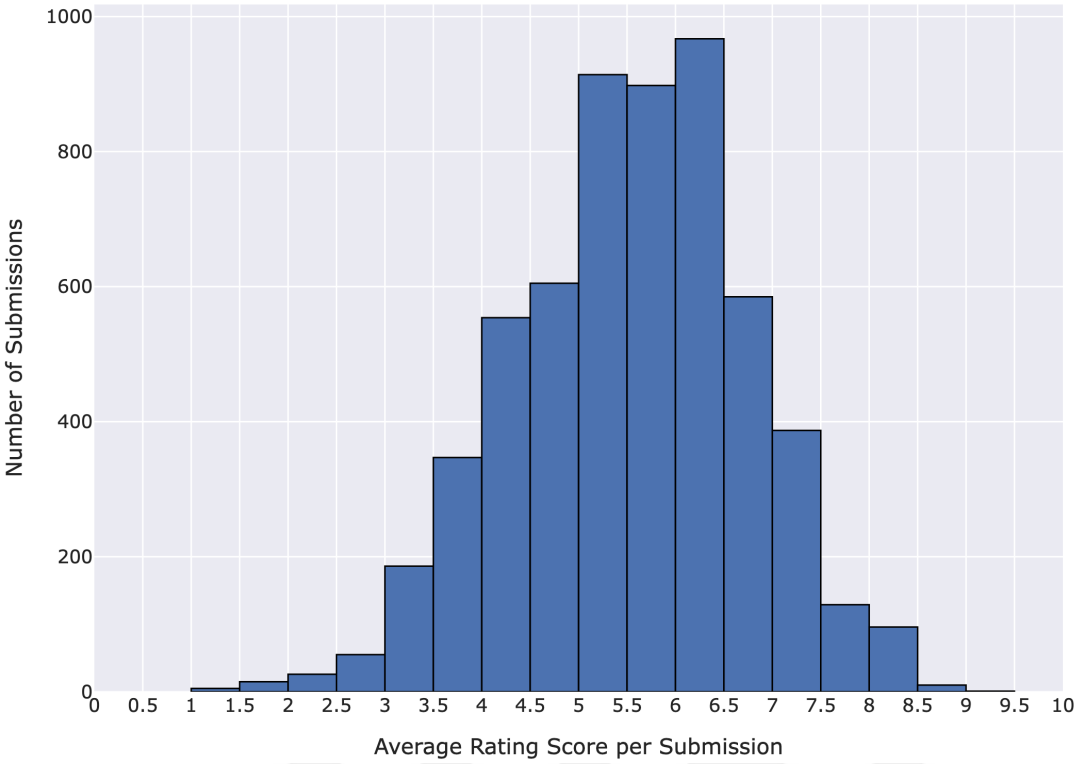


Figure 4: Distribution of average rating scores per submission.

The distribution of individual confidence scores, presented in Figure 15 (Appendix B.1.1), is slightly left-skewed. The mode of the distribution is 4, indicating that the typical reviewer expresses strong confidence in their assessment while acknowledging a slight margin for error, rather than claiming the absolute certainty associated with a maximum score. This concentration at the upper end of the scale suggests that the reviewer pool generally perceives itself as well-qualified for the assigned submissions, a positive indicator for the technical validity of the review process. This pattern of high self-reported expertise persists when the data is aggregated at the submission level. As illustrated in Figure 19 (Appendix B.1.2), the distribution of mean confidence scores per submission demonstrates a comparable concentration in the upper quartile, confirming that the vast majority of papers are evaluated by a panel of reviewers who collectively claim a high level of expertise.

Paper Quality Metrics: The evaluation framework includes three specific quality dimensions: `soundness`, `presentation`, and `contribution`. These metrics share a uniform four-point scale, ranging from '*poor*' (score 1) to '*excellent*' (score 4). The verbatim rubric definitions are provided in Table 21 (Appendix A.2).

A comparative analysis of these metrics reveals a divergence in reviewer sentiment. As illustrated by the distributions in Appendix B.1.1, `soundness` and `presentation` exhibit left-skewed distributions with a mode of three ("good"). In contrast, the distribution for `contribution` is slightly right-skewed, with a mode of two ("fair").

Table 9 quantifies this discrepancy across the 22,358 individual reviews. Among the quality metrics measured on the shared 1-4 scale, `presentation` achieves the highest mean score of 2.70, while `contribution` lags with a mean of 2.45. This gap is most visible in the median scores: the median for `contribution` is two, whereas both `soundness` and `presentation` maintain a median of three. This statistical disparity indicates that while reviewers generally find the `presentation` and `soundness` of the submissions adequate, they remain significantly more critical of the work's `contribution`.

To capture the consensus evaluation of each manuscript, the metrics were aggregated at the submission level. Table 10 summarizes the descriptive statistics for these average scores, while the corresponding frequency distributions are visualized in Appendix B.1.2.

Table 9: Descriptive statistics for all quantitative metrics in individual reviews.

Statistic	**Confidence** *(Scale 1-5)*	**Soundness** *(Scale 1-4)*	**Presentation** *(Scale 1-4)*	**Contribution** *(Scale 1-4)*
Mean	3.60	2.68	2.70	2.45
Std. Dev.	0.81	0.68	0.74	0.70
Min	1.00	1.00	1.00	1.00
25% (Q1)	3.00	2.00	2.00	2.00
50% (Median)	4.00	3.00	3.00	2.00
75% (Q3)	4.00	3.00	3.00	3.00
Max	5.00	4.00	4.00	4.00
Mode	**4.00**	**3.00**	**3.00**	**2.00**

Table 10: Descriptive statistics for average scores per submission.

Statistic	**Avg. Conf.** *(Scale 1-5)*	**Avg. Sound.** *(Scale 1-4)*	**Avg. Pres.** *(Scale 1-4)*	**Avg. Contrib.** *(Scale 1-4)*
Mean	3.60	2.68	2.70	2.45
Std. Dev.	0.48	0.42	0.46	0.44
Min	1.75	1.00	1.00	1.00
25% (Q1)	3.25	2.50	2.50	2.25
50% (Median)	3.67	2.75	2.75	2.50
75% (Q3)	4.00	3.00	3.00	2.75
Max	5.00	4.00	4.00	4.00

3.3.4.3 Textual Metrics

Beyond the quantitative scores, the ICLR 2024 review format includes three free-text fields which provide the narrative justification for a reviewer's assessment (see Figure 14). These fields serve distinct functional roles in the evaluation process:

- **Summary:** This field contains the reviewer's summary of the paper's core arguments, serving as a comprehension check that grounds the subsequent detailed evaluation.
- **Strengths:** In this section, reviewers articulate the perceived merits of the manuscript, highlighting novel contributions and robust methodologies.
- **Weaknesses:** This section is dedicated for identifying deficiencies, such as methodological flaws, missing baselines, or lack of clarity. It represents the core of the critical feedback.
- **Questions:** This field functions as the main channel for inquiry, enabling reviewers to scrutinize specific dimensions of the work, such as its methodological design, theoretical foundations, or empirical validity. These entries often act as conversation starters for the rebuttal phase. Notably, a subset of reviewers utilize this field referentially, inputting directives such as "See weaknesses" or "See above" rather than formulating distinct questions. This phenomenon underscores the interdependence of the review sections and necessitates context-aware processing for accurate analysis.

To characterize the effort allocated to each section, a comparative analysis of their word counts was performed. This approach draws on prior bibliometric research, which frequently utilizes review length as a quantitative proxy for reviewer engagement and the depth of feedback [36]. Although the `Summary` field was described for completeness, our quantitative analysis focuses on the `Strengths`, `Weaknesses`, and `Questions` sections, as these represent the core evaluative and dialogic components of the review. The `Summary` field is therefore excluded from the following statistical comparison. Table 11 presents the descriptive statistics for the length of these three key fields across all 22,358 reviews.

Table 11: Descriptive statistics for word counts of review text fields.

Statistic	Strengths	Weaknesses	Questions
Mean	63.31	181.81	83.94
Std. Dev.	47.89	162.46	111.05
Min	1	1	1
25% (Q1)	33	78	12
50% (Median)	52	141	50
75% (Q3)	80	235	110
95%	149	475	283
99%	229	793	535
Max	1475	2999	1923

The data reveals a clear asymmetry in reviewer effort. The `Weaknesses` section is, on average, nearly three times longer than the `Strengths` section ($\sim$ 181 vs. $\sim$ 63 words), suggesting that the review process is fundamentally oriented toward critique. Furthermore, the data is highly right-skewed. For all three fields, the maximum values far exceed the 99th percentile, indicating the presence of a "long tail" of reviewers who provide exceptionally extensive feedback. For instance, while the median `Weaknesses` length is 141 words, the maximum observed length is 2,999 words. The corresponding frequency distributions, which visually confirm this right-skewed pattern, are presented in Appendix B.2.

3.3.5 Correlation Analysis of Features

To investigate the interplay between the quantitative scores assigned by reviewers and the textual effort they expended, a comprehensive correlation analysis was conducted. This analysis aims to uncover the underlying relationships that shape a reviewer's final evaluation, revealing which factors are most influential and how different components of the review are related.

3.3.5.1 Selection of Correlation Metric

Given the nature of the dataset, the choice of correlation coefficient is critical for a valid analysis. The quantitative scores (`rating`, `confidence`, `soundness`, `presentation`, `contribution`) are measured on ordinal scales, where the values represent ordered categories rather than true interval measurements. Furthermore, as established in Section 3.3.4.3, the word count distributions for the free-text fields are highly right-skewed and non-normal.

Under these conditions, the commonly used Pearson's product-moment correlation coefficient is inappropriate, as its assumptions of linearity and normality are violated. A non-parametric rank correlation method is therefore required. While both Spearman's Rho [37] and Kendall's Tau [38] are suitable non-parametric alternatives, Kendall's Tau-b (τ_b) was selected for this analysis based on methodological evidence regarding its superior performance with Likert-scale data.

1. **Robustness to Tied Ranks:** The dataset contains a large number of tied ranks, an inherent characteristic of discrete Likert-style scales. Prior research [39] explicitly identifies Kendall's Tau-b as the most robust measure of association for such data, demonstrating that it outperforms other coefficients when facing severe departures from normality and uneven marginal splits.
2. **Direct Probabilistic Interpretation:** Kendall's Tau has a more direct and intuitive interpretation. The value of τ represents the difference between the probability that two randomly selected pairs of observations are in the same order (concordant) and the probability that they are in a different order (discordant).

For these reasons, Kendall's Tau provides a more accurate and interpretable measure of the monotonic relationship between the ordinal and skewed variables in this study.

3.3.5.2 Results and Interpretation

The pairwise correlations for all review features were computed across the 22,358 individual reviews. The resulting Kendall's Tau correlation matrix is visualized as a heatmap in Figure 5. Before interpreting these results, it is important to address the statistical context. With a sample size of over 22,000, the analysis possesses extremely high statistical power, meaning even trivial relationships are likely to be statistically significant ($p < 0.05$). Therefore, a p-value's significance is less informative than the magnitude of the correlation coefficient itself. Our interpretation will consequently focus on Kendall's Tau value as a measure of practical significance or effect size. The full table of p-values is provided in Appendix B.3 primarily to document the analysis and highlight the rare instances of non-significant findings.

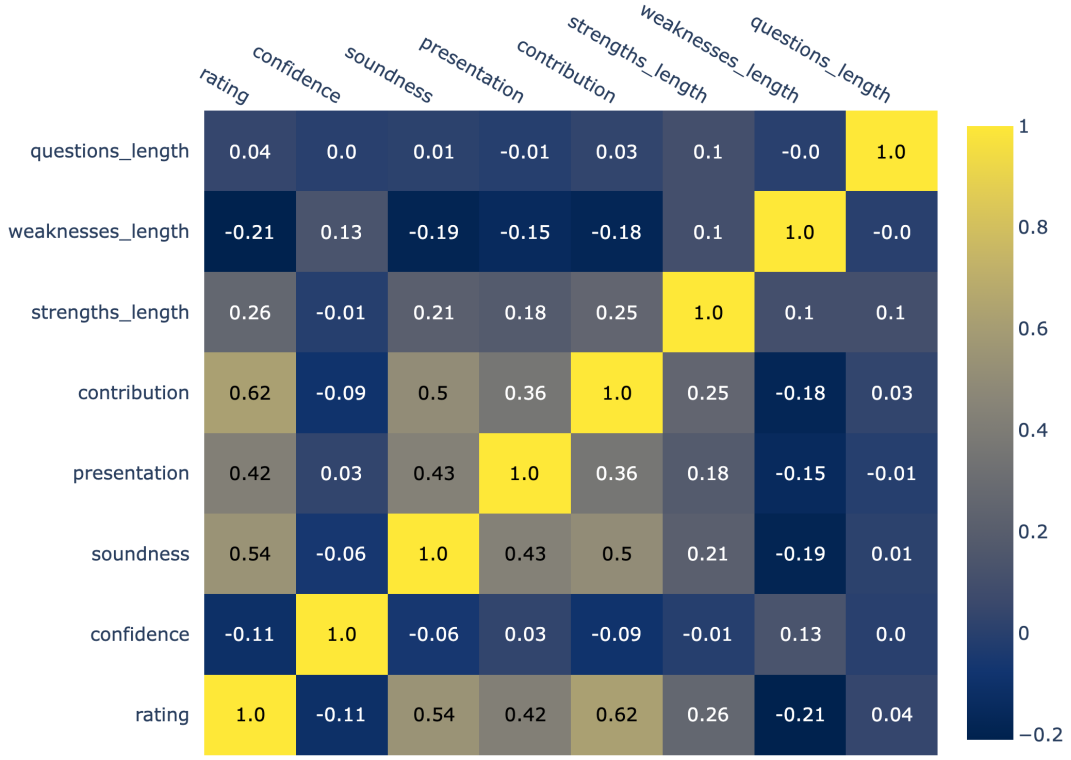


Figure 5: Kendall's Tau (τ_b) correlation matrix for review features.

The analysis yields several insights into the mechanics of peer review evaluation:

Drivers of Overall Rating: The strongest associations of the final `rating` are the paper's perceived quality metrics. `Contribution` exhibits the highest correlation ($\tau = 0.62$), followed by `Soundness` ($\tau = 0.54$) and `Presentation` ($\tau = 0.42$). This hierarchy confirms that a novel and significant contribution is the key driver of high ratings, followed by methodological soundness, with presentation quality ranking as the third most important determinant.

The Role of Textual Effort: The length of the free-text sections shows a clear and divergent relationship with the final rating.

- **Weaknesses Length:** A moderate negative correlation exists between `rating` and `weaknesses_length` ($\tau = -0.21$). This quantitatively confirms that lower-rated papers receive more extensive criticism as reviewers dedicate more effort to documenting flaws.
- **Strengths Length:** Conversely, `strengths_length` has a moderate positive correlation with `rating` ($\tau = 0.26$). Higher-rated papers elicit more detailed praise.
- **Questions Length:** The length of the `questions` section shows a negligible correlation with the final `rating` ($\tau = 0.04$). It shows that the volume of inquiry appears to be independent of the paper's perceived quality. Reviewers pose questions regardless of whether they view the paper favorably or unfavorably.

The Independence of the Questions Section: A particularly noteworthy finding is the statistical independence of the `Questions` section from other key evaluative components. The correlation between `questions_length` and `weaknesses_length` is effectively zero ($\tau \approx -0.00$, p=0.656), as is the correlation with `presentation` ($\tau \approx -0.01$, p=0.234) and `confidence` ($\tau \approx 0.00$, p=0.696). This evidence strongly refutes the hypothesis that the `Questions` section is merely an extension of the critique. Instead, it reinforces the view that this section serves a distinct dialogic function, separate from the evaluative role of the `Strengths` and `Weaknesses` sections, acting as a dedicated channel for clarification and discussion.

Reviewer Confidence: The `confidence` score reveals a complex pattern. It has a weak negative correlation with the overall `rating` ($\tau = -0.11$), suggesting that reviewers who report higher confidence tend to be slightly more critical. One interpretation is that reviewers feel most confident when they identify clear, definitive flaws, which leads to lower ratings. Conversely, borderline or highly novel papers might induce less confidence but receive more nuanced evaluations. The relationship between confidence and textual effort is also weak; for instance, there is no statistically significant relation between a reviewer's confidence and the length of their questions or strengths sections.

CHAPTER 4

METHODOLOGY

This chapter details the computational framework developed to systematically identify the latent topics within the `Questions` field of peer reviews and investigate the alignment between these data-driven topics and the official reviewer guidelines of top-tier venues. The experimental analysis is grounded in the ICLR 2024 peer review dataset, which was programmatically collected via the OpenReview API as detailed in Chapter 3. To ensure the validity of the findings, the data used in this chapter is restricted to the subset of the corpus comprising submissions that received a formal decision, thereby focusing exclusively on manuscripts that completed the full peer review lifecycle.

The methodological process is organized into six distinct stages:

1. *Experimental Design and Data Sampling:* The construction of statistically representative subsets of the ICLR 2024 corpus to facilitate computationally intensive analysis while preserving the distribution of reviewer sentiment.
2. *Source-Anchored Information Extraction:* The development of a custom LLM pipeline designed to extract question chunks while mitigating hallucination via embedding-based verification.
3. *Multi-Stage Data Validation Framework:* The implementation of an LLM-as-a-Judge consensus mechanism to empirically validate the quality of the extracted data using diverse reasoning models.
4. *Generative Topic Modeling:* The adaptation of the TopicGPT [8] framework to derive a schema of reviewer inquiries from the validated corpus.
5. *Logit-Based Soft Topic Classification:* The introduction of a soft-classification technique to quantify semantic uncertainty and evaluate topic separability using log-probabilities of generated tokens.
6. *Guideline Alignment Methodology:* An empirical investigation into the alignment between these data-driven topics and the prescribed reviewer guidelines of top-tier venues.

4.1 Experimental Design and Data Sampling

The ICLR 2024 dataset comprises 22,358 individual reviews from 5780 unique submissions. While this volume provides a comprehensive view of the conference, processing the entire corpus through complex, multi-stage LLM reasoning pipelines presents prohibitive computational constraints. Consequently, it was necessary to construct a smaller, manageable subset of the data for the information extraction phase.

However, a simple random down-sampling would be methodologically flawed due to two inherent characteristics of peer review data: *intra-submission redundancy* and *class imbalance*. Reviews for the same paper often repeat similar points, leading to semantic redundancy. Furthermore, the vast majority of submissions receive borderline ratings (scores of 5 or 6), rendering clear acceptances (score 10) and strong rejections (score 1) statistically rare. A naive random sample would likely exclude these edge cases and over-represent redundant borderline feedback.

To address these challenges, a multi-stage sampling strategy was employed to reduce the corpus to a computationally feasible size while preserving the semantic diversity of reviewer feedback. The reduction process began by randomly selecting a single review from each of the 5,780 submissions that completed the full peer-review cycle. From this normalized pool, a stratified sampling method was applied based on the review rating scores. A critical constraint of this step was the preservation of the minority class, ensuring that all reviews with a perfect rating of 10 were retained in their entirety while the remaining categories were sampled proportionally. This filtering process ultimately yielded a final analytical corpus of $N = 427$ reviews.

4.1.1 Stratified Sampling Strategy

The sampling process was designed to mitigate two specific forms of bias: *intra-submission correlation* and *class imbalance*.

Mitigation of Intra-Submission Correlation: Reviews associated with a single submission frequently exhibit significant thematic overlap, often raising identical concerns regarding the manuscript. Consequently, the inclusion of multiple reviews from a single submission would introduce *semantic redundancy*, artificially inflating the statistical prominence of specific topics. To mitigate this, the dataset was aggregated using the `forum` identifier, which serves as the unique submission-level key. From each resulting cluster, a single review was randomly sampled. This selection protocol ensures that each entry in the final corpus represents a distinct manuscript, thereby maximizing the semantic variability and independence of the dataset.

Balancing via Stratification: To accommodate the computational constraints of the extraction pipeline, the dataset required significant reduction. To prevent this reduction from introducing selection bias, we stratified the unique reviews based on their quantitative `rating`. This approach ensures that the resulting sample is not dominated by the majority class (borderline papers) but instead reflects the full spectrum of reviewer sentiment.

For the *Main Dataset*, a uniform sampling target was established at $n = 80$ per rating class. For the abundant classes (ratings 1, 3, 5, 6, 8), samples were drawn uniformly at random. However, the highest rating category (score 10) constitutes a minority class within the ICLR 2024 ecosystem. Since the total population of unique reviews with a rating of 10 ($n = 27$) fell below the target threshold, the entire population of this class was retained. This hybrid sampling strategy maximized the representation of positive feedback within the constraints of the experimental design, yielding a final analytical corpus of $N = 427$ unique reviews.

To ensure the integrity of the final analysis, a *Validation Dataset* ($N = 431$) was generated using an identical stratified sampling strategy. The slight difference in total volume between the two datasets ($N = 427$ vs. $N = 431$) is an expected consequence of the initial aggregation step. Since the reviews associated with the same submission may exhibit heterogeneous ratings, the random selection of one representative review per submission inherently alters the rating distribution of the available pool, resulting in minor fluctuations during the final stratification. Table 12 details the exact volume of data for both the Main and Validation datasets at each stage of the filtering and sampling processes. Similarly, Figure 6 visually illustrates the complete data flow of the corpus into these two distinct subsets.

Methodological Note on Dataset Usage: Distinct from traditional supervised learning workflows, the Main Dataset served as the core development corpus for this study. This decision was made to maximize the extraction fidelity of the specific text that constitutes the core of this empirical study. Consequently, the Validation Dataset was employed as a held-out corpus. It was strictly utilized to audit the generalizability of the work to confirm that the data extraction and processing logic was robust to unseen data, rather than being overfitted to the specific linguistic characteristics of the core corpus.

Table 12: Review counts per rating category across the sequential sampling stages.

Rating *(Score)*	**Full Corpus** *(All Reviews)*	**Unique Submissions** *(Main / Val)*	**Stratified Sample** *(Main / Val)*
1	364	96 / 86	80 / 80
3	4,292	1,093 / 1,093	80 / 80
5	6,186	1,607 / 1,602	80 / 80
6	7,651	1,972 / 2,033	80 / 80
8	3,763	985 / 935	80 / 80
10	102	27 / 31	27 / 31
Total	**22,358**	**5,780 / 5,780**	**427 / 431**

4.1.2 Distribution Verification

To validate that the initial filtering step did not introduce any selection bias, the normalized distribution of ratings in the created subset was compared against the full population distribution.

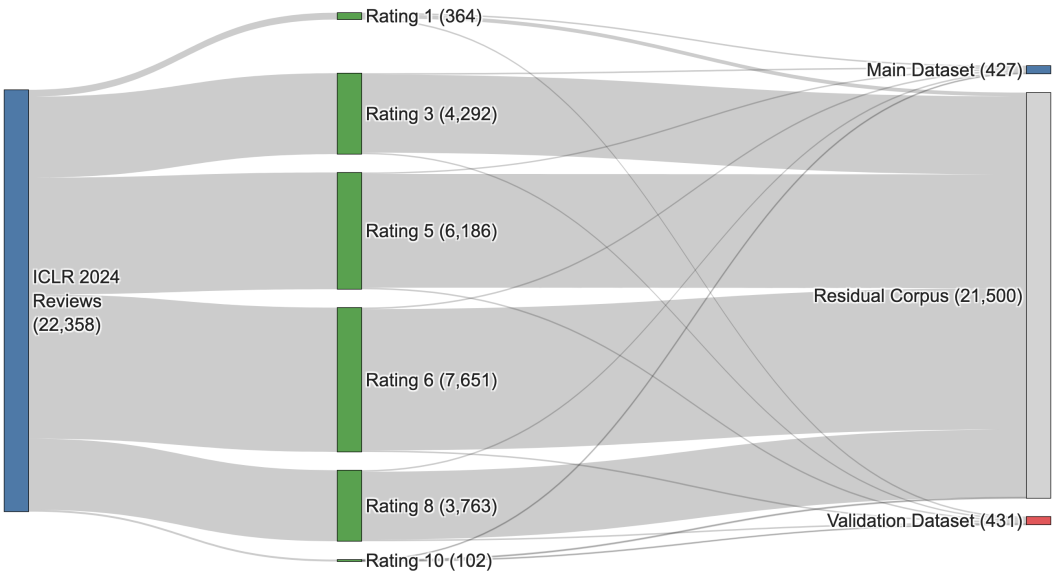


Figure 6: Sankey diagram illustrating the data flow from the full ICLR 2024 corpus to the stratified Main Dataset and the Validation Dataset.

As illustrated in Figure 7, the distributions are statistically indistinguishable. The "Sampled Reviews" cohort (orange), which specifically represents the Main Dataset, tracks the "All Reviews" population (blue) with near-perfect alignment across all rating categories. Furthermore, it is important to note that the Validation Dataset exhibits a distributional profile that is almost identical to the Main Dataset, a direct consequence of the shared stratification logic. This confirms that restricting the dataset to unique manuscripts preserves the natural sentiment distribution of the conference. This approach maintains the ordinal structure of the original data while ensuring that minority classes are sufficiently represented through stratification. Consequently, the curated corpus serves as a robust and valid input for the information extraction pipeline detailed in the following section.

4.2 Source-Anchored Information Extraction

The main focus of this study is the `questions` field within the peer reviews. Since this field consists of unstructured free text, reviewers often combine multiple inquiries, statements, and follow-up comments into a single paragraph. To perform accurate topic modeling, this text must be segmented into semantically coherent chunks so that each separated unit can be analyzed independently.

To accomplish this segmentation at scale, we employ a multi-stage LLM pipeline. This approach offers two key advantages: it ensures context-aware extraction and significantly reduces the need for manual intervention, thereby minimizing potential human bias. However, a critical requirement of this methodology is fidelity. To mitigate LLM hallucination or reviewer's intent alteration, we implemented a strict verification pipeline to ensure that every extracted chunk is anchored to the original source text.

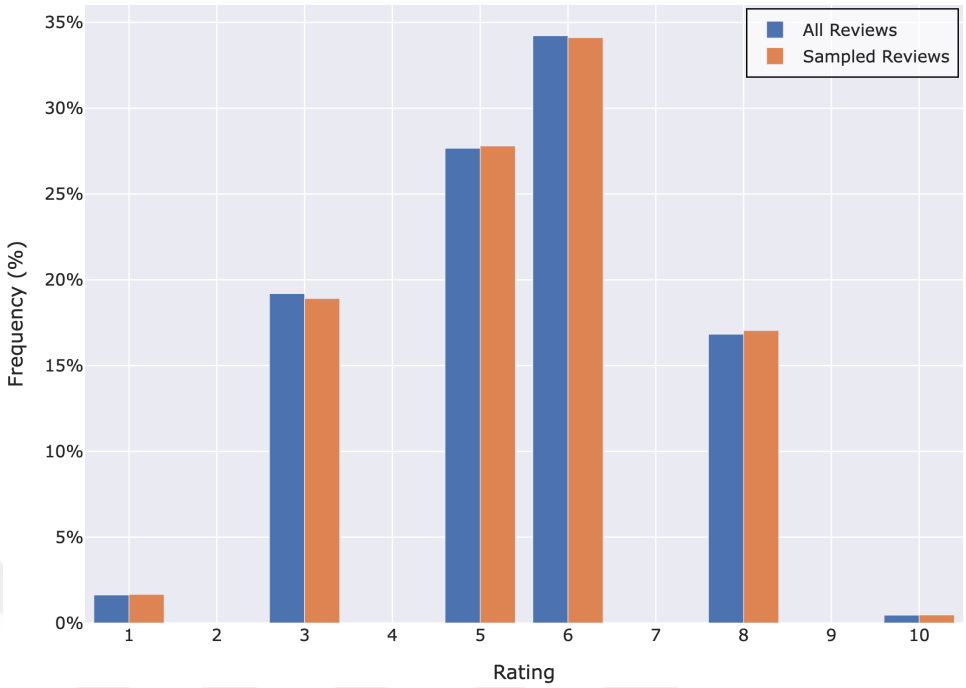


Figure 7: Comparative distribution of review ratings.

The extraction process itself is composed of three distinct stages:

1. ***Non-Informative Text Filtering:*** Identifying and removing non-substantive reviews to optimize computational costs.
2. ***Context-Aware Extraction:*** Segmenting questions into semantic chunks while performing a secondary validation of informativeness using a high-reasoning model.
3. ***Hallucination Mitigation:*** Ensuring the extracted text mathematically matches the original source to mitigate hallucination.

Upon completion of these extraction stages, the resulting candidate chunks are passed to the *Multi-Stage Data Validation Framework* (detailed in Section 4.3), which includes the LLM-as-a-Judge consensus mechanism.

Prompt Engineering and Reproducibility: A foundational principle governing all LLM-driven stages of this research was iterative prompt engineering. This principle was applied to every task, including non-informative text filtering, semantic chunking, and multi-agent judging. Initial prompts were developed and subsequently refined through an empirical feedback loop. This refinement process was guided by a direct analysis of model outputs, where any perceived errors, logical inconsistencies, or low inter-model agreement would trigger systematic modifications to the prompt's instructions, examples, or output constraints. To maximize the determinism and reproducibility of these pipelines, all LLM API calls were executed with a *temperature* parameter set to zero. To validate this choice, select batches of data were re-processed, yielding almost identical results and confirming the stability of

the methodology. The final prompts presented in the respective sections represent the stable outcome of this rigorous optimization process.

Implementation Note: To ensure efficiency, the entire pipeline was executed in a parallelized environment using Python's *ThreadPoolExecutor* with *semaphore-based* throttling. This architecture optimized concurrent throughput while efficiently governing API rate limits.

4.2.1 Non-Informative Text Filtering

As a preliminary data cleaning step, we addressed the presence of non-substantive text within the `questions` field. A significant portion of entries contained referential or empty statements, such as "*See weaknesses above.*", "*None*", or "*NA*". Since these entries do not convey useful information for topic modeling, processing them through the subsequent high-reasoning LLM models would incur unnecessary computational costs and introduce noise into the topic generation model.

To mitigate this, we implemented a pre-filtering stage using *Gemini 2.5 Flash [40]*, a model selected for its high inference speed and cost-efficiency. The model was instructed to perform binary classification to determine whether a given text contains actionable content. The model's output is a structured JSON object containing a single boolean key. The full prompt with evaluation criteria and few-shot examples is provided in Appendix D.1.

Pipeline Output: Out of the 427 sampled reviews, this initial filter identified 98 reviews (22.95%) as non-informative. These were excluded from further processing, leaving 329 reviews for the context-aware extraction stage.

4.2.2 Context-Aware Extraction

Following the initial filtering, the remaining 329 reviews were processed by a more sophisticated extraction stage designed to isolate and contextualize semantically self-contained inquiries. To perform this task, we utilized *Gemini 2.5 Pro*, a model selected for its superior reasoning capabilities. The model was provided with an expanded context containing the review's `summary`, `weaknesses`, and `questions` fields. As detailed in the prompt in Appendix D.2, the model was instructed to perform three main functions:

1. ***Semantic Chunking:*** Identify and group related sentences within the `questions` field into a single, coherent chunk. This ensures that an initial question and its direct follow-ups are treated as a single analytical unit. Crucially, the model was instructed to confine its extraction *exclusively* to text originating from the `questions` field, preventing any cross-contamination of data.

2. ***Contextual Enhancement:*** To further activate reasoning, the LLM model is instructed to generate a separate `expanded_context` field for each chunk. Where applicable, this field provides a concise, up to four sentence explanation of the question's context, synthesized from the `summary` and `weaknesses` sections.

3. ***Implicit Validation:*** Acting as a high-reasoning validation filter, the model was explicitly instructed to return an *empty list* if it determines that the `questions` field contains no self-contained inquiries.

The model outputs a JSON list of objects, where each object represents a single semantic chunk containing the original `question` and its synthesized `expanded_context`. The full prompt, which details the rules for semantic chunking and provides a comprehensive example to illustrate the required output structure, is provided in Appendix D.2.

Pipeline Output: During this stage, the Gemini 2.5 Pro model was executed on all 329 reviews that survived the initial non-informative text filter. Among these, the model identified 25 additional reviews as non-informative. Upon further investigation, those cases were found to be more subtle and context-dependent than those detected by the initial low-cost filter. As a result, the context-aware extraction step successfully processed 304 reviews, resulting in 809 semantically coherent question chunks.

4.2.3 Hallucination Mitigation

A critical risk in LLM-based data extraction is hallucination. Since maintaining the integrity of reviewer-authored content is essential for this study, every extracted question must be verifiably grounded in the original `questions` field. To enforce this requirement, after the Gemini 2.5 Pro model produced its list of question chunks, each chunk was validated against the original text using the following procedure:

1. ***Sentence Segmentation and Cleaning:*** Both the raw `questions` field and each LLM-generated chunk were segmented into sentences using the sentence tokenizer in *NLTK*[1]. Some reviews have enumeration artifacts such as "*Q1.*", "*1.*" or "–". These enumerations were removed using a regular expression to ensure clean semantic comparisons.

2. ***Embedding Generation:*** Tokenized sentences in both original question text and LLM-generated chunks were encoded into a dense vector representation using the pre-trained `all-MiniLM-L6-v2` Sentence-BERT [41] model.

3. ***Semantic Matching via Cosine Similarity:*** For each sentence generated by the LLM, the system computed cosine similarity scores against all original sentences and selected the closest semantic match. This step reconstructs, sentence by sentence, the best possible grounding in the original review text.

4. ***Chunk Reconstruction and Scoring:*** Each question chunk was augmented with:

    - the sequence of original sentences that best matched the LLM-generated sentences,
    - the cosine similarity score computed after regex-based cleaning, and
    - the cosine similarity score between the reconstructed text and the raw LLM-generated output.

[1] `https://www.nltk.org`

Pipeline Output and Metric Annotation: Upon completion of the context-aware extraction phase, the pipeline successfully processed 304 reviews, yielding an initial set of 809 candidate question chunks. Rather than filtering these immediately, the system annotated each chunk with its calculated cosine similarity scores (both raw and regex-cleaned). This approach preserves the complete distribution of extraction fidelity, establishing a transparent baseline for the rigorous, two-stage validation framework detailed in the following section.

4.3 Multi-Stage Data Validation Framework

The extraction pipeline described in the previous section yields a set of candidate question chunks. However, to ensure these candidates are suitable for rigorous thematic analysis, they must satisfy two distinct criteria: they must be *factually grounded* in the source text, and they must be *semantically coherent*. To enforce these standards, we implemented a two-stage validation framework comprising a deterministic similarity-based filter followed by a probabilistic multi-agent consensus mechanism.

4.3.1 Stage 1: Deterministic Fidelity Filtering

The first gate in the validation process addresses the risk of hallucination by enforcing a strict fidelity criterion at the review level. From the initial 304 successfully processed reviews, a review was only retained for the next stage if *all* of its constituent question chunks exceeded a cosine similarity threshold of $\tau = 0.85$.

The statistical justification for this threshold is provided in Appendix B.4, where Table 23 details the score distribution for the initial 809 candidate chunks. This holistic criterion ensures that every review entering the semantic validation phase is composed entirely of high-fidelity extractions, preventing contamination from even a single poorly extracted chunk.

Applying this review-level filter resulted in the exclusion of two reviews, whose chunks contained low-fidelity extractions. This left a validated corpus of *302 reviews*, which collectively contained *802 high-fidelity question chunks*. These 802 chunks formed the dataset for the subsequent semantic validation stage.

4.3.2 Stage 2: Semantic Consensus via LLM-as-a-Judge

While the fidelity filter guarantees that the text is authentic, it does not guarantee that the segmentation is semantically logical. To validate the coherence of the 802 surviving question chunks, we implemented a multi-agent consensus framework using the *LLM-as-a-Judge* paradigm [27]. For each chunk, the judging models were provided with the original review's full `questions` field as source context, alongside the candidate chunk and LLM-generated `expanded_context` for evaluation.

4.3.2.1 Judge Selection and Diversity

To minimize model-specific bias, we assembled a panel of three reasoning models from distinct vendors. This diversity ensures that the validation process is not an artifact of a single model architecture's preferences. The panel consisted of:

- *GPT-4o* [42] (OpenAI)
- *Claude 3.7 Sonnet* [43] (Anthropic)
- *Gemini 2.5 Flash* [40] (Google)

Note on Model Selection: We initially attempted to include models from Meta's *Llama 4*[2] series (specifically the 'Scout' and 'Maverick' variants) to further increase vendor diversity. However, these models were excluded after preliminary testing revealed two issues: inconsistent adherence to the required JSON output schema and critically low inter-rater agreement with the other LLMs. We hypothesize this divergence stemmed from the use of a single, cross-compatible prompt that, while optimized for the other three models, may not have been optimal for the Llama architecture. To maintain pipeline stability, they were excluded from the final panel. The full prompt used for the judging panel is provided in Appendix D.3.

4.3.2.2 Inter-Rater Reliability

To validate the LLM-as-a-Judge process and ensure that the semantic evaluation was not arbitrary, we quantified the level of agreement among the three LLM judges using standard Inter-Rater Reliability (IRR) metrics. Given our setup of three independent raters, we employed two complementary statistics across the 802 judged chunks.

Pairwise Agreement: Cohen's Kappa (κ) [44] is a statistic that measures inter-rater agreement for categorical items. It is more robust than simple percent agreement as it accounts for the possibility of agreement occurring by chance. In this study, it was used to compute the pairwise agreement between each of the three LLM judges. The results are visualized as a correlation heatmap in Figure 8, indicating a "Moderate" to "Substantial" level of agreement, with scores ranging from 0.539 to 0.641.

Overall Consensus: While Cohen's Kappa is limited to two raters, Fleiss' Kappa [45] is an extension that assesses the reliability of agreement among a fixed number of raters. It is the appropriate metric for evaluating the overall consensus of our three-judge panel. The analysis yielded an overall Fleiss' Kappa score of *0.589*. This score falls squarely within the "Moderate Agreement" range, confirming that the judging task was well-defined and that the models were evaluating semantic coherence based on a quite similar set of criteria.

[2] `https://www.llama.com/docs/model-cards-and-prompt-formats/llama4/`

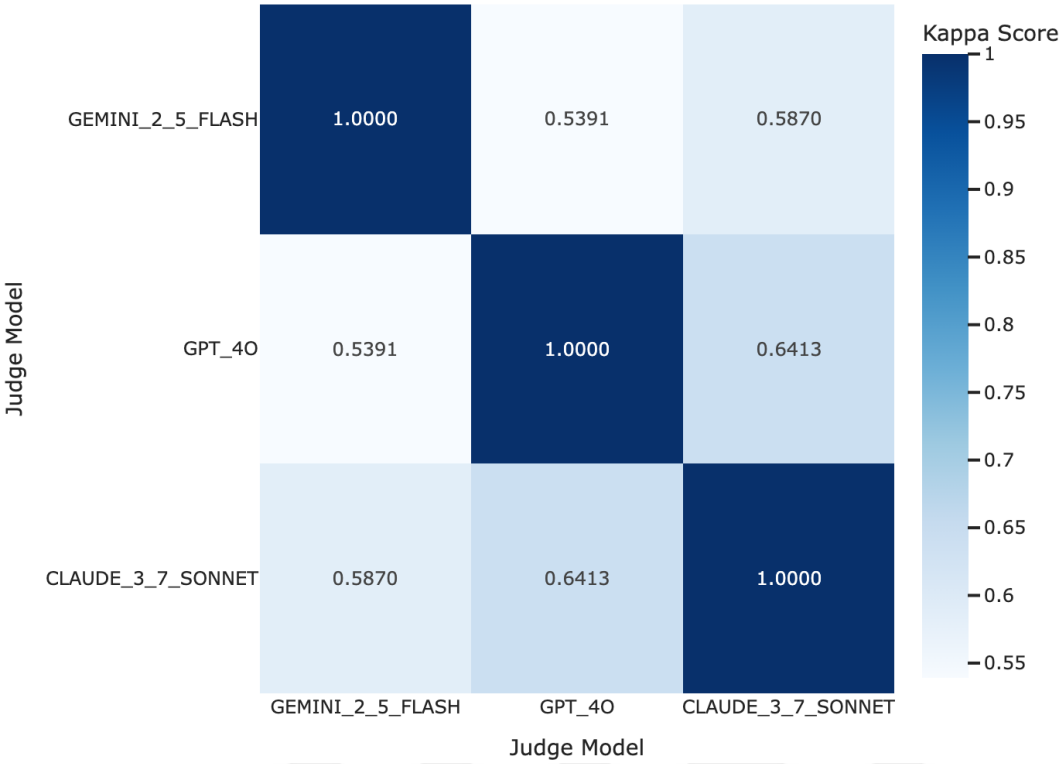


Figure 8: Cohen's Kappa inter-rater agreement matrix for the three LLM judges.

4.3.2.3 Final Consensus and Dataset Pruning

To distill the final analytical corpus from the 802 judged chunks, we applied a majority vote consensus rule. A chunk was retained for the final dataset only if at least two of the three LLM judges marked it as a valid semantic separation.

This consensus process revealed a strong overall agreement among the judges. A substantial majority, 92.64% of chunks, achieved unanimous agreement on their validity (3/3 'valid' votes), while an additional 3.49% were accepted with moderate agreement (2/3 'valid' votes). Conversely, 3.87% of chunks were discarded due to a lack of majority consensus, including 2.12% that were unanimously deemed invalid (0/3 'valid' votes). A detailed breakdown of these agreement levels is provided in Appendix B.5.

It is critical to note the two-stage nature of this pruning process: the dataset was first filtered at the *review level* for textual fidelity (cosine similarity > 0.85), and subsequently at the *chunk level* for semantic coherence (majority vote). Therefore, the resulting 771 question chunks satisfy both criteria and form the basis for all subsequent topic modeling and analysis. The complete evolution of the dataset through these extraction and validation stages is visualized in Figure 9.

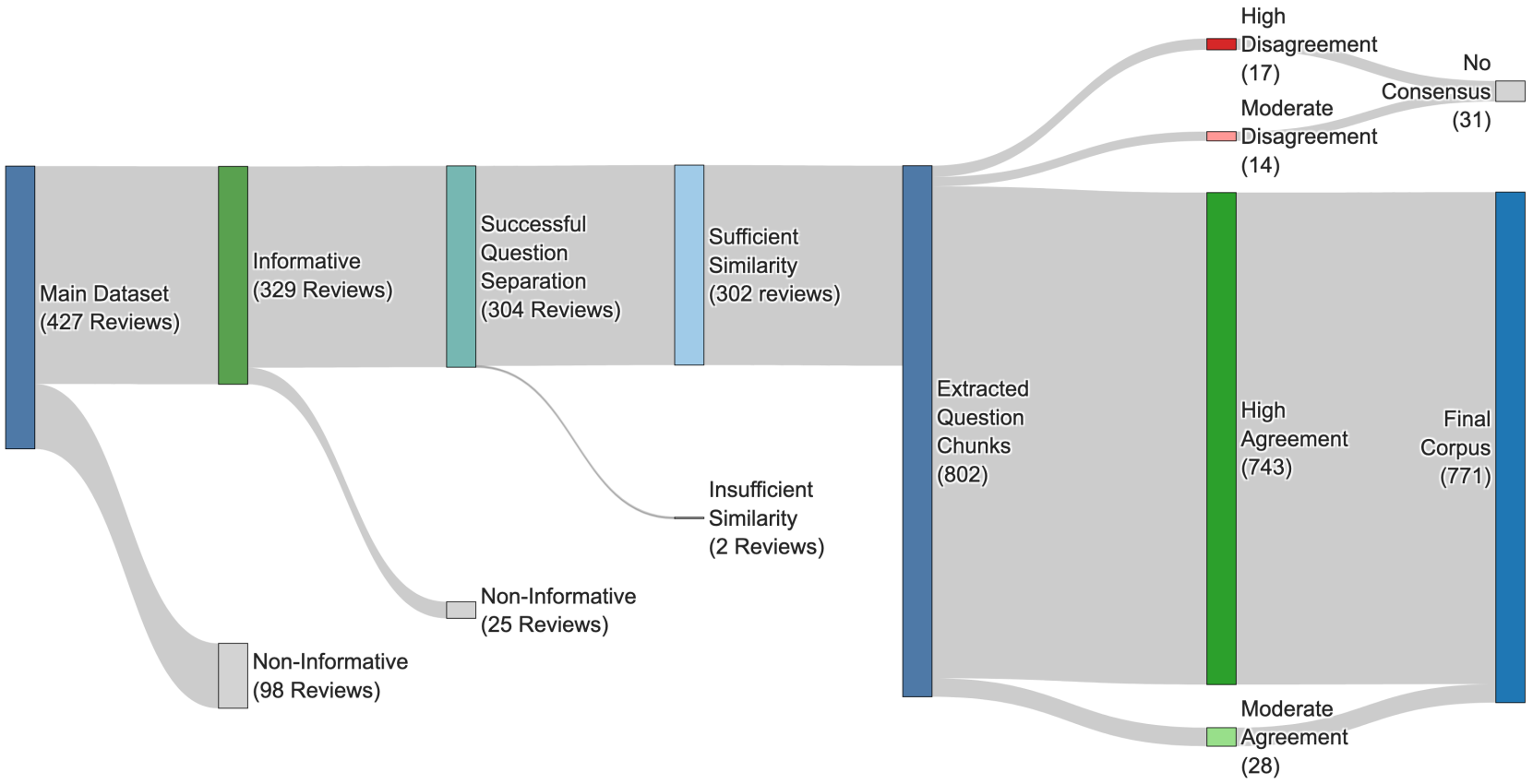


Figure 9: The complete data processing pipeline for the Main Dataset.

4.4 Generative Topic Modeling

After validating the extracted question chunks, the study proceeded to identify the latent topics that characterize reviewer inquiries. For this task, we adapted the *TopicGPT* [8] framework. This approach was selected over traditional probabilistic models, such as Latent Dirichlet Allocation (LDA) [19], which rely on bag-of-words assumptions. The reason is that the TopicGPT framework leverages the reasoning capabilities of LLMs to generate more context-aware and interpretable topic definitions.

4.4.1 Final Corpus Preparation

The input for the topic modeling phase was derived from the 771 question chunks that successfully passed the multi-agent consensus stage described in Section 4.3.2.3. To ensure that the generative models were conditioned exclusively on high-fidelity data, a final granular filter was applied to this subset. We strictly enforced the cosine similarity threshold ($\tau > 0.85$) at the individual chunk level to eliminate any residual outliers that may have survived the review-level aggregation. This final purification step reduced the corpus from 771 to a final set of 760 validated semantic units.

To ensure the robustness of the generated topics and mitigate any model-specific biases, 760 validated semantic units were prepared for multi-model discovery. To guarantee that the derived topics reflected the complete range of reviewer perspectives, we implemented a *stratified splitting* protocol. This approach was specifically designed to account for the scarcity of 'Strong Accept' (score 10) reviews, ensuring their distinct linguistic patterns were preserved. The dataset was split into two additional subsets to facilitate multi-model discovery. To preserve the distribution of the minority class, all samples associated with a rating of 10 ($n = 46$) were included in both subsets, while the remaining ratings were stratified 50/50. This ensured that the distinct linguistic patterns associated with high-quality submissions were available to the models during the topic generation phase. The data was serialized into the JSONL format required by the TopicGPT schema, which requires `id`, `text`, and `label` fields. Crucially, as this stage involved unsupervised topic discovery, only the `text` field was populated with the question content, while the `label` and `id` fields were left empty.

4.4.2 Domain-Specific Prompt Adaptation

The semantic scope of the generated schema is a direct result of the domain-specific adaptation of the TopicGPT framework. In its original configuration, the framework utilizes high-level, generic few-shot examples such as "*Agriculture*" and "*Trade*" to guide topic generation [8]. Preliminary experiments applying this default configuration to the peer review corpus yielded excessively broad topics, failing to capture the specific technical nuances of reviewer critique.

To address this resolution gap, we leveraged the *In-Context Learning* (ICL) capabilities of the model by engineering a modified prompt strategy. The generic few-shot examples were replaced with domain-anchored prototypes, specifically "*Inquiry About Data*" and "*Inquiry About Generalizability*". This prompt adaptation successfully steered the model toward a

more granular and focused topic generation, resulting in the specific format observed in the final schema. The complete prompt configuration used for this generation process is detailed in Section D.4 in Appendix D.

4.4.3 Multi-Run Topic Discovery and Consolidation

While the TopicGPT framework utilizes prompts originally optimized for the GPT-4o architecture, relying exclusively on a single model could introduce latent biases. To mitigate this and maximize the semantic diversity of the discovered themes, topic generation was performed across three distinct runs using heterogeneous model configurations:

1. **Run 1:** *Gemini 2.5 Flash* executed on the first data subset.
2. **Run 2:** *GPT-4o* executed on the second data subset.
3. **Run 3:** *GPT-4o* executed on the full concatenated corpus ($N = 760$).

To promote stability and reproducibility across these runs, a temperature value of zero was strictly enforced for all model interactions. While this setting does not guarantee absolute determinism due to the inherent architectural characteristics of large language models, it minimizes stochastic variability to the greatest extent possible. Empirical observation of consecutive runs confirmed that the generated topics remained highly consistent under these constraints, validating the reliability of the discovery process.

This multi-run approach produced three candidate lists of topics, the raw outputs of which are documented in Appendix C.1. These lists underwent a manual consolidation process to merge semantically equivalent themes and establish a unified schema. To verify the robustness of this consolidation, a secondary audit was performed using *Gemini 2.5 Pro*, which evaluated the consolidated list against the raw data to identify potential overlaps or missing categories. The outcome of this rigorous process was the final 13-topic schema utilized for the subsequent soft classification task.

4.5 Logit-Based Soft Topic Classification

The analysis of peer review inquiries is complicated by inherent semantic ambiguity. A single question often intersects with multiple topics, making exclusive categorization difficult. Standard LLM classification, which typically outputs a single text label, obscures this nuance. Furthermore, models trained with *Reinforcement Learning from Human Feedback (RLHF)* [46] often exhibit overconfidence, pushing probabilities to extremes. To address this, we use *Soft Classification* approach that utilizes raw token log-probabilities to quantify semantic uncertainty.

4.5.1 Soft Classification

We implemented a classification pipeline using *GPT-4o*, constraining the model to output a single token corresponding to the index of the assigned topic (indices 0-12). Rather than relying on the generated token, the system extracts the top-20 log-probabilities (`top_logprobs`) from the model's output layer. Requesting the top-20 log-probabilities ensures the capture of all 13 valid topic indices, mitigating any potential loss from the model assigning probability to out-of-scope tokens. This allows for the retrieval of the complete probability distribution across all candidate topics, converting the task from a hard classification into a probabilistic soft-clustering problem. The LLM prompt template used for this classification task is detailed in Appendix D.5.

Mitigation of Positional Bias: LLMs frequently demonstrate *positional bias* [47], favoring items appearing earlier or later in a list. To neutralize this, the order of the 13 topics provided in the system prompt was randomly shuffled for every API call, while maintaining the correct index-to-topic mapping. This ensures that the resulting probability distribution is a function of semantic alignment rather than prompt structure.

Vector Construction and Normalization: The extracted log-probabilities from LLM outputs were first converted to linear probabilities. To ensure a valid probability distribution over the defined schema, these values were re-normalized to sum to 1, yielding a probability vector $P(x)$ for each document across the 13 topic dimensions:

$$P(x_i) = \frac{e^{\text{logprob}_i}}{\sum_{j=0}^{N-1} e^{\text{logprob}_j}} \tag{1}$$

where $N = 13$ represents the number of topics.

4.5.2 Domain Coverage Verification

To verify that the consolidated set of 13 topics was exhaustive and that the model adhered to the output constraints, we defined a secondary metric termed *Total Valid Mass* (M_{valid}).

Since the model was queried for the top-20 log-probabilities, we calculated M_{valid} by summing the linear probabilities allocated specifically to the valid topic indices $\mathcal{T} = \{0, \ldots, 12\}$:

$$M_{valid} = \sum_{t \in \mathcal{T}} P_{linear}(t) \tag{2}$$

This metric serves as a proxy for schema adherence. A value of $M_{valid} \approx 1.0$ indicates that the model is operating strictly within the defined schema. Conversely, a significant drop would imply probability leakage, where the model distributes mass to tokens outside the predefined set (e.g., hallucinated labels). In our experiments, this metric consistently approached unity (> 0.99), confirming the robustness of the prompt constraints.

4.5.3 Quantification of Semantic Uncertainty

To rigorously evaluate the model's confidence and distinguish between unambiguous inquiries and those exhibiting semantic overlap, we computed two complementary diagnostic metrics for every classified document, namely *Normalized Shannon Entropy* [48] and *Probability Margin*.

4.5.3.1 Normalized Shannon Entropy

To quantify the dispersion of the probability mass, we utilized *Normalized Shannon Entropy* (H_{norm}). Unlike standard entropy, which scales with the number of classes, this metric is normalized by the natural logarithm of the number of topics ($N = 13$), bounding the result between 0 and 1.

$$H_{norm}(X) = \frac{-\sum_{i=1}^{N} P(x_i) \ln P(x_i)}{\ln N} \tag{3}$$

This metric serves as a proxy for semantic focus:

- $H_{norm} \approx 0$: Indicates a "Surgical" classification, where the model assigns nearly all probability mass to a single topic.
- $H_{norm} \approx 1$: Indicates maximum entropy (randomness), suggesting the input text is either irrelevant or equally applicable to all topics.

4.5.3.2 Probability Margin

While entropy measures overall confusion, it does not explicitly capture the competition between the leading candidates. To address this, we defined the *Probability Margin* (Δ_{prob}), calculated as the difference between the probability of the top-ranked topic (t_1) and the second-ranked topic (t_2):

$$\Delta_{prob} = P(t_1) - P(t_2) \tag{4}$$

This metric quantifies the distinctness of the primary classification:

- $\Delta_{prob} \approx 1.0$: Indicates an *Extremely Confident Classification*, where the main topic dominates the probability distribution without significant competition.
- $\Delta_{prob} \approx 0.0$: Indicates a *Semantic Tie*, where the model struggles to distinguish between the two leading domains. This typically occurs when a question sits on the decision boundary of two closely related topics.

4.5.3.3 Operational Interpretation and the RLHF Effect

The interpretation of these metrics requires a nuanced approach due to the specific training characteristics of modern LLMs. Models fine-tuned via *RLHF* are known to exhibit *calibration drift*, often reflected as extreme overconfidence. The optimization for preference ranking tends to push probability distributions toward the extremes (0 and 1), artificially suppressing entropy even in cases of mild ambiguity.

Consequently, standard confidence thresholds (e.g., $\Delta_{prob} > 0.5$) would be insufficient for this study. Therefore, a strict interpretation strategy is employed on Probability Margin:

- *Confident Regime ($\Delta_{prob} > 0.95$):* Classifications in this zone exhibit near-absolute certainty, where the model identifies a single dominant topic with negligible competition.
- *Semi-Confident Regime ($0.40 \leq \Delta_{prob} \leq 0.95$):* This intermediate zone represents classifications that are valid but lack a probability distribution indicative of certainty, often reflecting questions that contain secondary thematic elements.
- *Uncertain Regime ($\Delta_{prob} < 0.40$):* Classifications falling below this threshold are flagged as ambiguous, indicating strong semantic competition between multiple topics or a lack of clear alignment with the schema.

While Probability Margin serves as the main decision boundary, Normalized Entropy provides necessary intuition regarding the *structure* of the confusion. Empirical analysis of our results reveals a strong inverse correlation between the two metrics (Kendall's $\tau \approx -0.97$, Pearson $r \approx -0.94$). Despite this statistical redundancy, Entropy remains critical in the "Uncertain Regime": it distinguishes between a focused conflict involving two topics (lower entropy) and a broad failure to classify across the schema (higher entropy).

4.5.4 Topic Separability via Jensen-Shannon Divergence

While normalized *Shannon Entropy* assesses the ambiguity of individual documents, it does not determine whether the topics themselves are distinct. To validate that the 13 consolidated topics consist of non-redundant, disjoint semantic domains, the distributional distance between those topics is investigated.

We aggregated the probability vectors of all documents assigned to a specific topic to compute a centroid vector for that topic. We then calculated the pairwise *Jensen-Shannon Divergence (JSD)* [49] between these centroids. JSD is a symmetric and smoothed version of the *Kullback-Leibler Divergence* [50], bounded between 0 and 1:

$$JSD(P \parallel Q) = \frac{1}{2} D_{KL}(P \parallel M) + \frac{1}{2} D_{KL}(Q \parallel M) \tag{5}$$

where $M = \frac{1}{2}(P + Q)$.

A JSD value near 0 would indicate that two topics have identical probability distributions (i.e., they are synonyms). A value near one indicates that the topics occupy distinct semantic spaces. This metric serves as a mathematical validation of the manual topic consolidation process described in Section 4.4.3.

4.6 Guideline Alignment Methodology

The final stage of the methodology was designed to quantify the alignment between the data-driven topics (*Practiced Criteria*) and the institutional instructions provided to reviewers (*Prescribed Criteria*).

4.6.1 Venue Selection and Guideline Corpus Construction

A normative baseline for the *Prescribed Criteria* of peer review was established by curating a representative corpus of reviewer guidelines. The guidelines were sourced from the most influential publication venues in the field. The selection of these venues was based on their bibliometric impact, which was measured by the *h5-index* from the Google Scholar Metrics ranking [4]. The process began by identifying the top 20 publications within the *Engineering & Computer Science* category. From this initial list, a final subset of ten venues specifically dedicated to Artificial Intelligence (AI), Machine Learning (ML), Computer Vision (CV), and Natural Language Processing (NLP) was selected.

For each selected venue, the most recent publicly available reviewer guideline was retrieved to ensure the analysis reflects current community standards. Specific methodological decisions regarding venue-specific formats and shared standards were made as follows:

- ***ICML:*** Given the conference's multi-track structure, guidelines were explicitly sourced from the "Main Track," as it accounts for the majority of submissions.
- ***ACL & EMNLP:*** As the Association for Computational Linguistics has adopted a centralized review process, the guidelines from the *ACL Rolling Review (ARR)* were utilized, serving as the unified standard for both ACL and EMNLP.
- ***CVPR & ICCV:*** It was observed that these venues utilize an identical evaluation protocol; consequently, the shared reviewer guidelines were retrieved to represent both conferences.
- ***TPAMI:*** As a journal, the IEEE Transactions on Pattern Analysis and Machine Intelligence maintains a static set of standing guidelines rather than yearly iterations.

The complete list of venues and the specific guideline versions included in the corpus is detailed in Table 13.

Table 13: Selected top-tier venues for guideline analysis.

Venue Name	Acronym	Guideline Version	Source
IEEE/CVF Conference on Computer Vision and Pattern Recognition	CVPR	2026	[51]
Neural Information Processing Systems	NeurIPS	2025	[52]
International Conference on Learning Representations	ICLR	2024/2025	[53, 54]
International Conference on Machine Learning	ICML	2025 (Main Track)	[55]
European Conference on Computer Vision	ECCV	2024	[56, 57]
IEEE/CVF International Conference on Computer Vision	ICCV	2025	[51]
Meeting of the Association for Computational Linguistics (ACL)	ACL	ARR	[58]
AAAI Conference on Artificial Intelligence	AAAI	2026	[59]
Conference on Empirical Methods in Natural Language Processing (EMNLP)	EMNLP	2025 ARR	[58]
IEEE Transactions on Pattern Analysis and Machine Intelligence	TPAMI	Current (Journal)	[60, 61]

4.6.2 Thematic Mapping and Gap Analysis

To systematically quantify the alignment between the *Prescribed Criteria* (guidelines) and the *Practiced Criteria* (reviewer inquiries), a directed content analysis was performed. This phase employed a deductive mapping strategy, explicitly anchoring the analysis of the institutional guidelines to the 13-topic schema established in Section 4.4.

The methodology was executed in three distinct stages: directive extraction, deductive topic assignment, and human-in-the-loop validation. This framework establishes a direct correspondence between institutional expectations and reviewer inquiries. The full corpus of guidelines was utilized to characterize the broader normative landscape of the field. However, the specific quantitative gap analysis was restricted to the ICLR venue. We strictly compared the ICLR 2024 review data against the ICLR 2024 reviewer guidelines. This approach facilitates a bidirectional analysis defined by two main categories.

- ***Unwritten Norms:*** Thematic domains that exhibit high frequency within the empirical review data but lack explicit representation in the official guidelines.

- ***Neglected Criteria:*** Evaluation dimensions that are explicitly mandated by the guidelines but appear with negligible frequency in the actual reviewer inquiries.

4.6.2.1 Directive Extraction and Semantic Chunking

Although ten top-tier venues were selected for analysis, the structural overlaps detailed in Section 4.6 resulted in a final corpus of eight unique reviewer guideline documents. The raw text of these eight documents was processed to isolate *directive statements*. These statements are defined as sentences or clauses that explicitly instruct a reviewer on what attributes to evaluate in submissions. Therefore, general administrative instructions such as deadlines and formatting rules were excluded.

To ensure the robustness of the extraction process, a dual-extraction protocol was employed. First, relevant directive statements were manually identified within the guideline texts. To mitigate potential selection bias and validity threats, an independent extraction was performed using *Gemini 3 Pro*. The outputs from both sources were then cross-referenced and consolidated, ensuring that only valid and relevant directives were retained for the final analysis.

4.6.2.2 Deductive Topic Assignment

Following the extraction and consolidation of directive statements, each entry was classified against the 13-topic schema established in Section 4.4. This assignment was performed by *Gemini 3 Pro* using the specialized prompt detailed in Appendix D.6 to map the extracted guideline statements onto the empirically derived topic schema.

The model was instructed to evaluate the semantic focus of each statement and assign one or more appropriate topics from the schema, accompanied by a brief justification for the selection. Crucially, the model was explicitly instructed to assign "-" if the guideline criterion did not semantically align with any of the 13 defined topics. This classification distinction is fundamental to the gap analysis, as it isolates the specific evaluation criteria that are institutionally mandated by the venues, but do not appear in the actual reviewer inquiries.

4.6.2.3 Human-in-the-Loop Validation

To ensure the methodological rigor of the mapping process and mitigate the validity threats inherent in automated classification, a multi-tiered verification protocol was implemented.

First, an independent parallel annotation protocol was established. In this phase, the primary researcher performed a blind classification of the guideline statements. A specific challenge in this task is that guideline statements can often be implicitly interpreted as multiple topics with different levels of correspondence. To address this ambiguity, the manual annotation relied strictly on explicit directive correspondence rather than implied meaning. Finally, these human-generated annotations systematically compared against the model's outputs to identify discrepancies.

Finally, the results were cross-checked by a senior academic expert. This expert audit spanned approximately five hours and functioned as a comparative assessment. The expert simultaneously reviewed the guideline directives, the topics assigned by the LLM, and the manual

annotations provided by the primary researcher. For each instance, the expert rendered a final classification decision based strictly on the explicit semantic meaning of the directive. In addition to reviewing the standard topic assignments, this process prioritized edge cases where the correct topic assignment was ambiguous, as well as conflicting labels and the *Unassigned* ("-") category to prevent false negatives. Any remaining disagreements were resolved through a consensus-based discussion. This human-in-the-loop mechanism ensures that the final alignment data is not merely an artifact of stochastic model behavior, but a verified representation of the relationship between institutional guidelines and the generated topics from reviewer practice.

CHAPTER 5

ANALYSIS AND RESULTS

This chapter presents the empirical findings derived from the framework detailed in Chapter 4. The analysis begins by extracting a consolidated topic schema of reviewer inquiries, subsequently evaluating the structural coherence of these topics through probabilistic uncertainty metrics and analyzes their frequency distribution to reveal the operational priorities of the reviewer community. Finally, this chapter concludes with a systematic gap analysis, contrasting these data-driven *practiced criteria* against the *prescribed criteria* of top-tier venue guidelines to quantify the alignment between institutional policy and actual peer review behavior.

5.1 The Schema of Reviewer Inquiries

The schema of reviewer inquiries presented in this section represents the consolidated outcome of a robust, multi-model discovery protocol designed to mitigate the biases inherent in single-pass topic generation. As detailed in Section 4.4.3, the TopicGPT [8] framework was executed across three independent runs. Those runs were executed on two stratified subsets (n = 403) and one leveraging the complete corpus of validated question chunks (N = 760). Their outputs yield a diverse set of candidate topics, and the raw outputs of which are documented in Tables 26, 27, and 28 in Appendix C.1. Through the systematic consolidation of these candidate lists, a unified schema of 13 distinct topics was established.

The semantic stability of this schema was confirmed through a dual-verification protocol. The candidate topics were subjected to independent clustering by both the primary researcher and *Gemini 2.5 Pro [40]*. The high degree of alignment between the human-derived groupings and the LLM-generated organization validates the coherence of the resulting categories. Consequently, the standardized definitions generated by the model were adopted to ensure terminological precision.

5.1.1 Semantic Definitions and Scope

The resulting schema characterizes the multidimensional nature of reviewer critique, encompassing distinct thematic domains of a submission. The 13 identified topics are defined in Table 14. Each entry provides the consolidated topic name alongside its semantic definition, which establishes the topic scopes for the subsequent classification and alignment analyses.

Table 14: The consolidated schema of reviewer inquiries.

Consolidated Topics	Topic Definition
Inquiry About Methodology	Questions regarding the techniques, models, or approaches used, including their design, specific configurations, and components.
Inquiry About Theoretical Foundations	Questions regarding the theoretical background, assumptions, and justifications underlying the work.
Inquiry About Performance and Metrics	Questions focused on the internal results, outcomes, and the metrics used to measure the work's performance.
Inquiry About Comparative Analysis	Questions focused on the external context, including comparisons to baselines, related work, and alternative approaches.
Inquiry About Applicability and Limitations	Questions regarding the scope of the work's claims, including its generalizability, robustness, scalability, and practical real-world use.
Inquiry About Data	Questions specifically about the dataset(s) used, including its source, statistics, preprocessing, or availability.
Inquiry About Presentation and Figures	Questions regarding how results and data are presented, focusing on figures, tables, and overall structure.
Inquiry About Terminology and Clarity	Questions regarding the choice, appropriateness, and clarity of terms, definitions, and written language.
Inquiry About Interpretability and Explanation	Questions seeking to understand the 'why' behind the model's behavior, including requests for explanations or transparency of its internal mechanisms.
Inquiry About Motivation and Contribution	Questions regarding the purpose, novelty, and significance of the work, including the rationale for the research.
Inquiry About Validity and Reproducibility	Questions concerning the correctness of claims, the reliability of the results, and the ability to replicate the work.
Inquiry About Ethical Considerations	Questions regarding the ethical implications of the work, including concerns about privacy, data protection, and potential bias.
Inquiry About Computational Efficiency	Questions regarding the computational cost, resource usage (time, memory), or speed of the proposed methods.

Since this schema is derived directly from the review data, it provides a transparent view into the actual priorities of the reviewers. It also demonstrates that reviewers scrutinize submissions on a highly diverse set of criteria. This spectrum extends from the abstract mathematical rigor of *Inquiry About Theoretical Foundations* to the practical engineering constraints of *Inquiry About Computational Efficiency*. A comprehensive catalog of representative examples for each topic is provided in Appendix C.2.

5.1.2 Frequency Distribution and Operational Priorities

The frequency distribution of the 13 consolidated topics across the Main Dataset ($N = 760$ semantic chunks) reveals a distinct hierarchy in reviewer priorities. As visualized in Figure 10 and detailed in Table 15, there is a class imbalance in the reviewer inquiries.

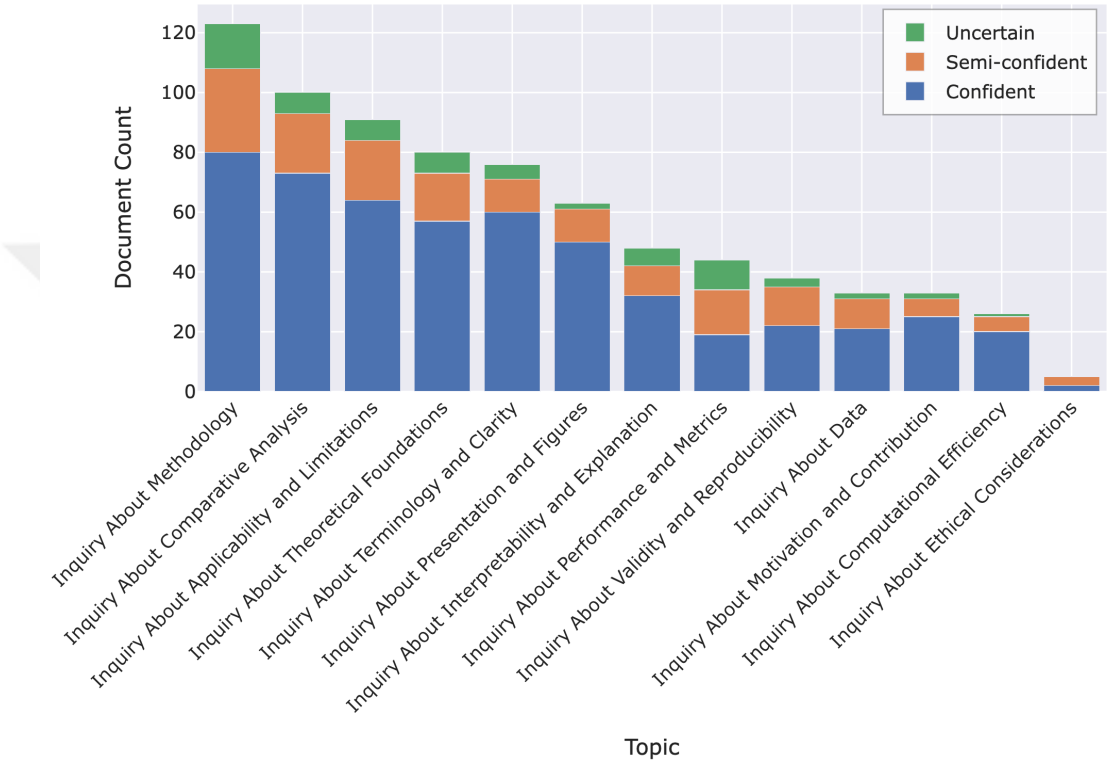


Figure 10: Frequency distribution of reviewer inquiries across the 13 generated topics.

The frequency and confidence levels of inquiries reveal a distinct thematic prioritization in reviewer feedback. *Inquiry About Methodology* emerges as the single most dominant topic with 16.18% of the entire corpus. This is closely followed by *Inquiry About Comparative Analysis* (13.16%) and *Inquiry About Applicability and Limitations* (11.97%). Collectively, these three topics constitute over 41% of all questions, establishing that reviewers primarily scrutinize a paper's core technical approach, its positioning relative to existing work, and the boundaries of its claimed contributions. Following this top tier, a secondary group receives consistent but less frequent attention. This group includes inquiries about *Theoretical Foundations* (10.53%), *Terminology and Clarity* (10.00%), and *Presentation and Figures* (8.29%). These findings indicate that after assessing the core methodology and comparisons, reviewers turn their attention to the foundations, writing, and delivery of the work. Finally, the remaining seven topics represent 29.87% of the corpus. Within these topics, *Inquiry About Interpretability and Explanation* (6.32%) is the most frequent category, whereas *Inquiry About Ethical Considerations* (0.66%) is the least represented topic with only five samples.

Table 15: Frequency distribution of reviewer inquiries.

Topic Name	Total	% of Corpus	Confident	Semi-Conf.	Uncertain
Inquiry About Methodology	123	16.18%	80	28	15
Inquiry About Comparative Analysis	100	13.16%	73	20	7
Inquiry About Applicability and Limitations	91	11.97%	64	20	7
Inquiry About Theoretical Foundations	80	10.53%	57	16	7
Inquiry About Terminology and Clarity	76	10.00%	60	11	5
Inquiry About Presentation and Figures	63	8.29%	50	11	2
Inquiry About Interpretability and Explanation	48	6.32%	32	10	6
Inquiry About Performance and Metrics	44	5.79%	19	15	10
Inquiry About Validity and Reproducibility	38	5.00%	22	13	3
Inquiry About Data	33	4.34%	21	10	2
Inquiry About Motivation and Contribution	33	4.34%	25	6	2
Inquiry About Computational Efficiency	26	3.42%	20	5	1
Inquiry About Ethical Considerations	5	0.66%	2	3	0

5.2 Analysis of Semantic Uncertainty

This section moves beyond simple frequency counts to evaluate the structural coherence of the topic schema. The analysis leverages the probabilistic outputs of the soft-classification framework in two distinct ways. First, it quantifies the semantic ambiguity of individual inquiries at the document level. Second, it assesses the semantic separability between the topics themselves, thereby validating the overall coherence of the schema.

5.2.1 Quantification of Reviewer Ambiguity

The semantic clarity of reviewer inquiries was evaluated using two complementary metrics. *Probability Margin* (Δ_{prob}) measures the model's confidence in its primary choice, and *Normalized Shannon Entropy* (H_{norm}) quantifies the dispersion of probability mass across the topic schema.

The statistics presented in Table 16 reveal an exceptionally high degree of classification certainty across the main corpus. The median Probability Margin is 0.9982, with a corresponding median Normalized Entropy of just 0.0031.

Figure 11 provides a granular view of this certainty by plotting the distribution of Probability Margins for each topic. The probability margin is defined as the arithmetic difference between the highest and second-highest probability scores assigned to a sample. Consequently, a value close to 1.0 indicates that the model's top choice is overwhelmingly dominant. In the boxplot, a narrow Interquartile Range (IQR) positioned near the upper bound (1.0) implies that the vast majority of questions assigned to that topic are classified with near-absolute certainty. This pattern implies a high degree of semantic isolation. This suggests that the topic is distinct enough that it rarely shares probability mass with other topics. For instance, *Inquiry About Presentation and Figures* exhibits a highly compacted distribution near 1.0. This indicates that the questions assigned to this topic are structurally and semantically unique, showing virtually no overlap with other topics.

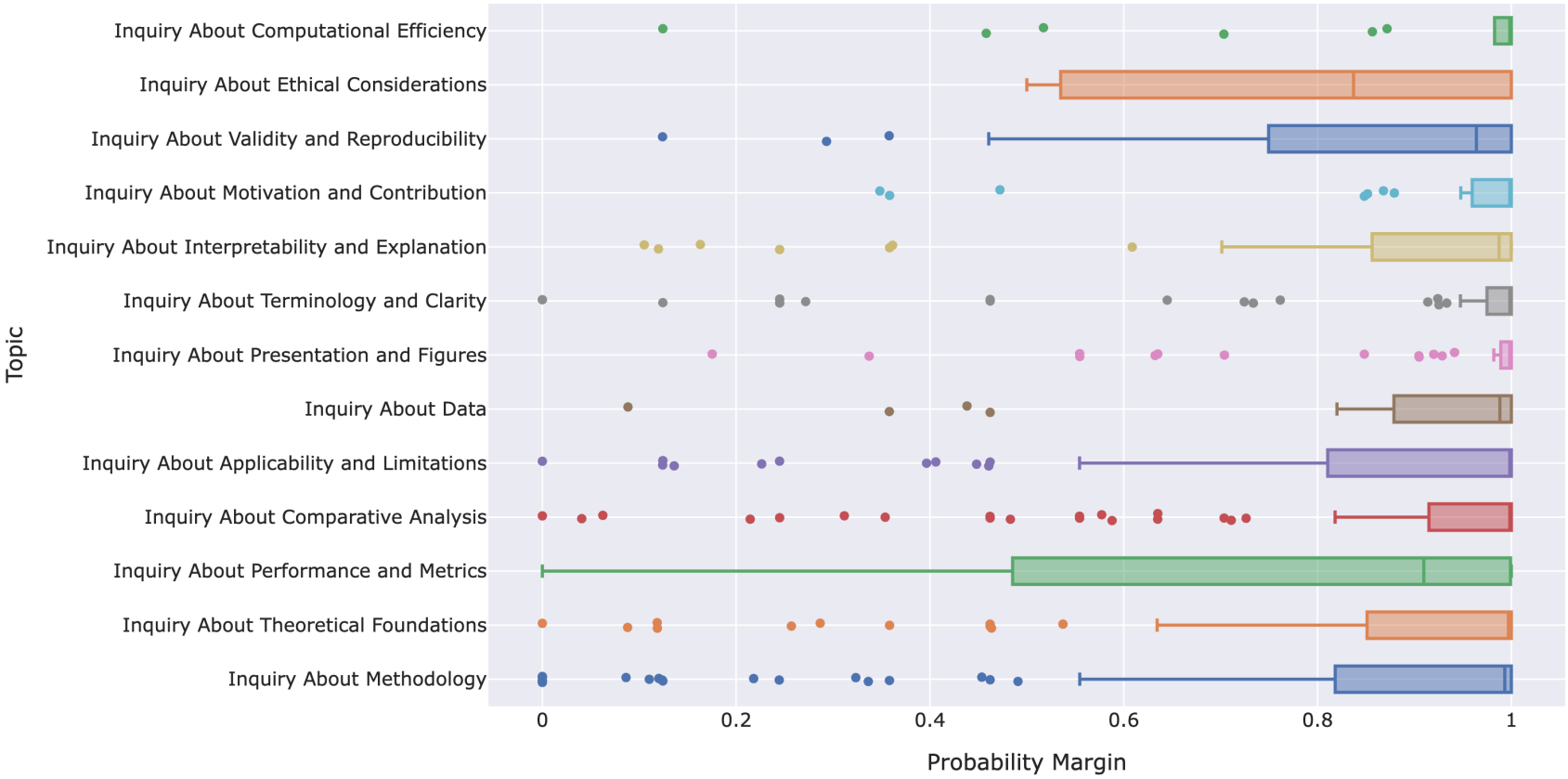


Figure 11: Boxplot of Probability Margin per topic on Main Dataset.

However, the boxplot also highlights specific topics that are more susceptible to semantic ambiguity. For instance, *Inquiry About Performance and Metrics* displays the widest distribution and one of the lowest median Probability Margins. This suggests that questions about performance are often inherently multifaceted, frequently overlapping with other domains. For example, the question "*How does the F1 score of your model compare to the current state-of-the-art baselines on the benchmark datasets?*" is also relevant to the *Inquiry About Comparative Analysis* topic.

To verify the robustness of these findings, the same uncertainty analysis was conducted on the validation dataset, as illustrated in Figure 28. The results exhibit no significant deviation from the main corpus, confirming the consistency of the model's classification confidence across different data splits. Furthermore, the median Probability Margins in both datasets demonstrate a strong skew towards 1.0, a characteristic consistent with the expected behavior of RLHF-tuned models.

Finally, the *Total Valid Mass* metric in Table 16 serves as a critical validation of the pipeline's integrity. The metric consistently approached unity (mean = 1.0000, std. dev. = 0.0000) with the lowest observed value being 0.9993. This confirms that the model operated strictly within the defined 13-topic schema for almost all classifications. This means almost no probability leakage to out-of-scope tokens is observed.

Despite these edge cases, the low mean entropy and high median probability margin confirm that the generated schema provides a robust and distinct coverage of the reviewer discourse.

Table 16: Summary statistics for probabilistic uncertainty metrics.

Statistic	**Total Valid Mass**	**Margin**	**Norm. Entropy**	**Top Prob.**	**Second Prob.**
Count	760	760	760	760	760
Mean	1.0000	0.8663	0.0720	0.9262	0.0599
Std. Dev.	0.0000	0.2538	0.1194	0.1422	0.1146
Min	0.9993	0.0000	0.0000	0.3460	0.0000
10%	1.0000	0.4620	0.0000	0.7238	0.0000
20%	1.0000	0.7676	0.0000	0.8718	0.0000
30%	1.0000	0.9411	0.0000	0.9701	0.0000
40%	1.0000	0.9880	0.0002	0.9933	0.0000
50%	1.0000	0.9982	0.0031	0.9990	0.0009
95%	1.0000	1.0000	0.3352	1.0000	0.3382
96%	1.0000	1.0000	0.3628	1.0000	0.3775
97%	1.0000	1.0000	0.3886	1.0000	0.3956
98%	1.0000	1.0000	0.4267	1.0000	0.4340
99%	1.0000	1.0000	0.4914	1.0000	0.4464
Max	1.0000	1.0000	0.5425	1.0000	0.4999

5.2.2 Topic Separability and Structural Overlap

The entropy analysis presented in the previous section confirms that individual inquiries are classified with high confidence. However, to fully validate the integrity of the data-driven schema, it is essential to demonstrate that the generated topics occupy distinct regions of the semantic space. As outlined in Section 4.5.4, a robust topic model must yield document clusters that are easily separable with minimal distributional overlap.

We applied the Jensen-Shannon Divergence (JSD) metric to the aggregated topic centroids to quantify this separability. The heatmap in Figure 12 visualizes the pairwise JSD outputs of the generated topics. The analysis reveals an exceptionally high degree of semantic separability across the schema. The vast majority of pairwise JSD scores exceed 0.90, with many approaching 1.0. This indicates that most topics occupy orthogonal semantic spaces and represent distinct areas of reviewer inquiry, validating the consolidation process described in Section 4.4.3.

However, the matrix also highlights specific clusters of semantic proximity. The lowest JSD score (JSD = 0.872) was observed between *Inquiry About Performance and Metrics* and *Inquiry About Comparative Analysis*. A similar, though less pronounced overlap (JSD = 0.891) exists between *Inquiry About Validity and Reproducibility* and *Inquiry About Performance and Metrics*. Furthermore, *Inquiry About Methodology* shows a measurable adjacency to *Inquiry About Theoretical Foundations* (JSD = 0.898).

Conversely, topics such as *Inquiry About Ethical Considerations* and *Inquiry About Computational Efficiency* exhibit near absolute divergence from the rest of the schema. This isolation confirms that these topics represent specialized, self-contained domains of critique.

5.3 The Guideline-Practice Gap

This section investigates the alignment between actual reviewer inquiries and institutional guidelines. We refer to the generated topics in Section 5.1 as the *Practiced Criteria*. These are contrasted against the *Prescribed Criteria* which represent the explicit instructions found in official reviewer guidelines.

The mapping between these two domains was executed following the methodology described in Section 4.6.2. To ensure the rigor of this analysis, the resulting assignments were subjected to expert verification. As summarized in Table 17, this validation process required the resolution of 16 edge cases through consensus discussions between the primary researcher and a senior academic expert.

5.3.1 Consensus and Divergence Across Venues

To establish a normative baseline for this comparison, we first analyzed the distribution of directives across the eight guideline documents representing the ten top-tier AI venues. The

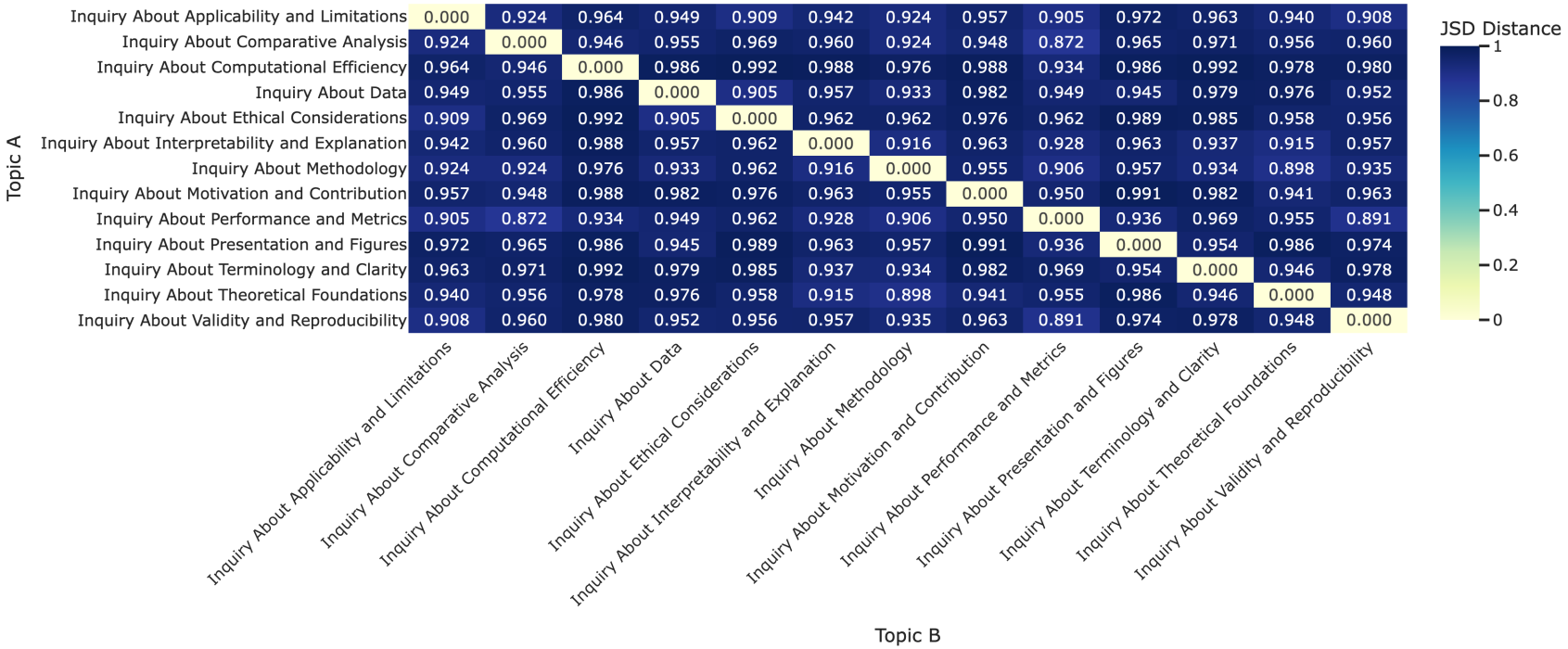


Figure 12: Jensen-Shannon Divergence (JSD) heatmap of topic centroids. Values closer to 1 (darker) indicate higher semantic separability, while values closer to 0 (lighter) indicate greater overlap.

Table 17: Validation statistics for automated guideline topic assignment.

Venue	Correct Assignments	False Positives	False Negatives	Resolved Edge Cases
AAAI	37	4	1	3
ICCV & CVPR	16	3	0	3
ECCV	39	0	2	0
ICML	13	0	0	0
ICLR	10	0	0	0
NeurIPS	24	2	3	3
ACL & EMNLP	48	5	4	6
TPAMI	29	0	0	1

analysis reveals that while there is no single universal standard governing the field, a high degree of consensus exists regarding fundamental scientific attributes.

The analysis identifies two universally mandated criteria. Specifically, *Inquiry About Motivation and Contribution* and *Inquiry About Validity and Reproducibility*. These topics appear as explicit directives in every analyzed guideline document. In contrast, the remaining criteria exhibit lower levels of consensus. The minimum consensus observed in *Inquiry About Computational Efficiency*, *Inquiry About Presentation and Figures*, and *Inquiry About Interpretability and Explanations* topics with explicit directives of only 2 unique reviewer guidelines. Consequently, the set of prescribed evaluation standards is not uniform across the field.

5.3.2 Alignment Analysis: Prescribed vs. Practiced

Table 18 presents the results of the mapping between guideline directives and the data-driven 13 topic schema. A detailed examination of the ICLR column reveals a distinct regulatory structure. Unlike venues with comprehensive evaluation checklists such as AAAI or ACL, ICLR guidelines adopt a concise approach. The text explicitly mandates only four of the 13 identified topics, namely *Inquiry About Comparative Analysis*, *Inquiry About Terminology and Clarity*, *Inquiry About Motivation and Contribution*, and *Inquiry About Validity and Reproducibility*. This selective approach leaves the remaining nine topics undefined by formal guidelines, and it creates a structural void that leads to the divergences. We categorized the resulting divergences of venues into two distinct groups: *Unwritten Norms* and *Neglected Criteria*.

5.3.2.1 Unwritten Norms

This category identifies thematic domains where community practices extend beyond institutional guidelines. A systematic cross-referencing of the ICLR 2024 reviewer guidelines against the 13-topic schema reveals a significant divergence between the *prescribed* evaluation criteria and the *practiced* critique.

When these prescribed topics are mapped to the frequency distribution of the actual reviews, a major gap emerges. The four explicitly requested topics account for only *32.50%* of the total volume of inquiries in the dataset. Conversely, this implies that *67.50%* of the questions address criteria that are technically *unwritten* within the specific context of the ICLR guidelines.

The most prominent driver of this gap is *Inquiry About Methodology* (16.18%). Despite being the single most frequent category of reviewer inquiry, the ICLR guidelines do not explicitly contain a directive for this topic. This represents the largest single instance of divergence, where the primary focus of the community's scrutiny is absent from the institutional instructions.

A similar pattern is observed in the domains of theoretical and practical boundaries. *Inquiry About Applicability and Limitations* (11.97%) and *Inquiry About Theoretical Foundations* (10.53%) represent the third and fourth most common types of questions, respectively.

Beyond these dominant themes, the "unwritten" category encompasses other fundamental aspects of scientific rigor. The analysis reveals that inquiries regarding *Presentation and Figures* (8.29%), *Interpretability and Explanation* (6.32%), *Performance and Metrics* (5.79%), *Data* (4.34%), *Computational Efficiency* (3.42%), and *Ethical Considerations* (0.66%) are all practiced by reviewers despite the lack of explicit institutional mandates.

5.3.2.2 Neglected Criteria

Conversely, the analysis identifies "Neglected Criteria" as explicitly requested but minimally represented topics in review data. As established in Table 18, the ICLR guidelines explicitly mandate four topics for evaluation. Among these topics, *Inquiry About Motivation and Contribution* and *Inquiry About Validity and Reproducibility* appear relatively low with 4.34% and 5.00% of the corpus, respectively.

5.3.3 Guideline-Exclusive Categories

The gap analysis further reveals specific evaluation criteria that are mentioned in reviewer guidelines but were not observed in the generated topic schema. The reason is that these criteria did not present sufficient inquiries to form a topic in the *Questions* field of the reviews:

- *Societal Impact:* Directives requiring the evaluation of positive and negative societal impacts are present in the guidelines of *ICCV/CVPR* (presented to raise awareness), *ECCV*, *NeurIPS*, and *ACL/EMNLP*.
- *Licensing:* Guidelines for *NeurIPS* and *ACL/EMNLP* explicitly instruct reviewers to verify the terms of artifact sourcing, ensuring that existing assets are properly cited and that original licenses are respected.
- *Appropriateness:* The guidelines for *TPAMI* explicitly require an evaluation of submission style, manuscript organization, and compliance with publication requirements to determine if the article is appropriate for the venue.

Table 18: Alignment of data-driven topics with official reviewer guidelines across top-tier venues.

Topics	AAAI	ICCV & CVPR	ECCV	ICML	ICLR	NeurIPS	ACL & EMNLP	TPAMI
Inquiry About Methodology	✓		✓			✓	✓	✓
Inquiry About Theoretical Foundations	✓		✓			✓	✓	
Inquiry About Performance and Metrics	✓	✓	✓			✓	✓	
Inquiry About Comparative Analysis	✓			✓	✓		✓	✓
Inquiry About Applicability and Limitations	✓	✓	✓			✓	✓	
Inquiry About Data	✓	✓	✓			✓	✓	✓
Inquiry About Presentation and Figures			✓					✓
Inquiry About Terminology and Clarity	✓		✓	✓	✓	✓	✓	✓
Inquiry About Interpretability and Explanation	✓						✓	
Inquiry About Motivation and Contribution	✓	✓	✓	✓	✓	✓	✓	✓
Inquiry About Validity and Reproducibility	✓	✓	✓	✓	✓	✓	✓	✓
Inquiry About Ethical Considerations		✓	✓	✓		✓	✓	✓
Inquiry About Computational Efficiency						✓	✓	

CHAPTER 6

DISCUSSION

This chapter interprets the empirical findings presented in Chapter 5 and places them within the broader context of academic peer review research. The goal of this chapter is to explain the underlying patterns revealed by the analysis and to clarify how these patterns should be understood based on the semantic structure of reviewer inquiries, the behavior of Large Language Models (LLMs), and the role of institutional reviewer guidelines.

6.1 The Hierarchy of Reviewer Priorities

The topic distribution observed in the ICLR 2024 *Questions* field reveals a clear hierarchy of reviewer attention. A small subset of topics dominates the inquiry space, most notably *Inquiry About Methodology*, *Inquiry About Comparative Analysis*, and *Inquiry About Applicability and Limitations*. Together, these topics account for a substantial fraction of all extracted reviewer questions.

This hierarchy indicates that reviewer questions are mainly focused on clarifying uncertain aspects of a paper, rather than covering the full set of evaluation criteria. Reviewers appear to focus their inquiries on dimensions that most directly influence the credibility, interpretability, and positioning of a submission. *Inquiries About Methodology* focus on the techniques and design choices underlying the proposed approach, which directly affect the credibility of the reported results. *Inquiries About Comparative Analysis* situate the work within the existing literature by examining comparisons to baselines and alternative methods, thereby informing the significance of the contribution in its external context. *Inquiries About Applicability and Limitations* address the scope of the work's claims, including issues of generalizability, robustness, and practical use. These dimensions represent core aspects of scientific evaluation and therefore attract a concentrated level of reviewer scrutiny.

Importantly, the high frequency of these topics should not be interpreted as a complete representation of reviewer concerns. Instead, it reflects where reviewers focus their questions to obtain clarification, while other evaluation criteria may be addressed implicitly. Therefore, the *Questions* field captures active points of uncertainty rather than the full evaluative checklist applied by reviewers.

This interpretation aligns with the observed independence between question length and overall rating score reported in Chapter 3. Reviewers ask questions across the rating spectrum. This

shows that inquiries are driven by informational needs rather than by acceptance or rejection intent. Therefore, the hierarchy of topics reflects how reviewers selectively direct their attention during the evaluation process.

6.2 Structural Coherence and Model Behavior

While the preceding section interprets the hierarchy of reviewer priorities, it is also necessary to examine the structural properties of the generated topic schema itself. In particular, this section analyzes whether the observed topic boundaries reflect meaningful semantic distinctions within reviewer discourse or are influenced by characteristics of the modeling approach. To this end, we examine both the semantic overlap between topics and the behavior of the soft classification approach.

6.2.1 Semantic Interdependence of Topics

The final schema consists of 13 semantically distinct and well-separated topics. Quantitative analysis using Jensen–Shannon Divergence demonstrates that each topic occupies a clearly defined region of the semantic space, with all pairwise divergence values remaining well above commonly accepted thresholds for topic distinctness. Even the lowest observed divergence scores of topic pairs indicate strong separation rather than semantic overlap, confirming the structural coherence of the topic schema.

These results indicate that the generated topics form a stable and well-separated representation of reviewer inquiry. Even under strict confidence thresholds, the majority of the corpus was classified with high confidence. This demonstrates that the semantic distinctions between topics are strong and consistently identifiable. This finding also shows that reviewer questions are expressed in sufficiently distinct linguistic patterns to support reliable topic assignment.

Defining the boundary for the semi-confident regime presents a challenge. Unlike sharply separable domains, reviewer inquiries often span multiple evaluative dimensions. As a result, semi-confident samples typically correspond to cases where multiple topic assignments are plausible rather than ambiguity or noise. The absence of a clear separation between high-confidence samples and multi-topic cases reinforces this view. It indicates that reduced confidence serves as a proxy for semantic richness rather than a measure of model uncertainty.

Overall, the combination of high probability margins, low entropy values, and the robustness of assignments under conservative thresholds confirms that the topic schema achieves strong semantic separation, while also accommodating the inherently interconnected nature of reviewer discourse.

6.2.2 The Influence of RLHF on Uncertainty

The probabilistic topic assignment results exhibit extremely high confidence values. The median probability margin approaches unity. This behavior is consistent with known characteristics of Large Language Models trained using Reinforcement Learning from Human Feedback (RLHF). Even in minor ambiguity cases, such models tend to produce overconfident probability distributions.

This characteristic motivated the conservative interpretation strategy adopted in this study. Confidence metrics were not treated as calibrated probabilities. Instead, they were interpreted as relative indicators of semantic focus. Thresholds for probability margin and normalized entropy were intentionally set near extreme values to avoid overstating certainty.

The observed variation across topics further supports this interpretation. Topics with narrow conceptual scope, such as *Inquiry About Computational Efficiency* or *Inquiry About Terminology and Clarity*, display consistently high confidence and low dispersion. In contrast, topics associated with empirical evaluation exhibit broader confidence distributions, reflecting genuine semantic overlap rather than instability in the classification process.

Crucially, the *Total Valid Mass* metric demonstrates that high confidence assignments remain confined to the predefined topic schema. Across the entire corpus, the probability mass assigned to tokens outside the predefined topic indices remains negligible. The LLM was explicitly constrained to output only a fixed set of topic indices. The results show that the model consistently followed this constraint. This metric confirms that the model's responses remained within the intended semantic space and supports the reliability of the resulting classification outputs.

6.3 Contextualization within the Peer Review Literature

To assess the external validity of the generated topic schema, we compare our findings against established taxonomies in the domain of computational peer review analysis. Specifically, we examine the correspondence between our data-driven topics and three benchmark studies: the foundational *PeerRead* dataset introduced by Kang et al. [5], the multi-layered annotation schema proposed by Ghosal et al. [6], and the aspect-based sentiment analysis framework developed by Chakraborty et al. [11]. Collectively, these works analyze ICLR data spanning from 2017 to 2019, providing a longitudinal baseline for our 2024 analysis.

A comparative analysis reveals a convergence between the labels used in these prior works and the generated topics of this study. While the primary objectives of these studies differ, the evaluative dimensions they identify exhibit semantic similarity to our findings. Consequently, although the specific terminology varies due to different methodological goals, the underlying constructs remain consistent. For instance, our topics align closely with the "Section Correspondence" and "Paper Aspect" layers in the annotation schema of Ghosal et al. [6]. Similarly, the shared set of evaluation aspects utilized by Kang et al. [5] and Chakraborty et al. [11] maps closely to the part of inquiry topics generated in this study.

6.4 The Guideline–Practice Gap

The guideline alignment analysis reveals a substantial divergence between the evaluation criteria explicitly articulated in official documents and the dimensions that most frequently appear in reviewer inquiries.

6.4.1 Implicit Norms Beyond Formal Guidelines

A central finding of this thesis is that the majority of reviewer inquiries are not explicitly grounded in the ICLR 2024 reviewer guidelines. Of the thirteen empirically derived topics, only four are directly prescribed by the official guideline document, accounting for approximately one-third of the observed questions. The remaining inquiries correspond to evaluative dimensions that are consistently practiced by reviewers but not formally specified by the venue.

This pattern indicates that reviewer behavior is shaped by both written instructions and established community standards. The current ICLR guideline does not provide an exhaustive evaluation rubric. Instead, they function by reinforcing and relying on shared disciplinary norms, particularly in areas such as methodological rigor, theoretical grounding, and submission positioning.

As a result, reviewer practice extends beyond the evaluation criteria explicitly articulated in the guideline document. This divergence reflects a distinction between formally specified instructions and the broader set of evaluative considerations applied in practice. In this context, reviewer inquiries reveal how community standards operate alongside institutional guidelines during the review process.

The concise nature of the ICLR guidelines further amplifies this effect. Compared to venues that provide extensive and highly granular reviewer directives, ICLR places greater reliance on the evaluative judgment of its reviewer pool. Consequently, the observed gap reflects a structural characteristic of the venue's review framework rather than an inconsistency in reviewer behavior.

6.4.2 Institutional Lag and Emergent Criteria

The guideline analysis also identifies evaluation criteria that are explicitly articulated in institutional documents but appear infrequently in reviewer inquiries. Ethical considerations provide the clearest example of this pattern. Despite being emphasized across multiple top-tier venues, questions explicitly addressing ethical issues represent only a very small fraction of the inquiry corpus.

There are several possible explanations for this pattern. Ethical requirements may be sufficiently addressed by authors through mandatory statements, reducing the need for additional clarification during review. Alternatively, ethical compliance may be assessed by reviewers without generating detailed questions in the *Questions* field. In such cases, ethical consider-

ations may influence reviewer judgments without leading to explicit questions in the review text.

Beyond ethics, reviewer guidelines across multiple venues include additional evaluative dimensions that do not map directly onto the topic schema derived from reviewer inquiries in ICLR 2024. These dimensions include societal impact, licensing, and appropriateness. During the guideline analysis, these criteria could not be consistently aligned with any of the empirically generated inquiry topics and were therefore identified as guideline-specific categories. As they do not appear with sufficient frequency in reviewer questions, they were not incorporated into the final topic schema used for analysis.

The presence of these criteria highlights a distinction between institutional priorities and the operational focus of reviewer inquiries. Societal impact, in particular, represents a normative concern emphasized by several venues, yet it does not appear as a frequent subject of reviewer questioning in the ICLR 2024 corpus. By contrast, licensing and appropriateness represent procedural considerations that are often evaluated outside the dialogic space captured by reviewer inquiries. Overall, these findings indicate that reviewer questions and institutional guidelines capture complementary aspects of evaluation. Reviewer inquiries represent active demands for information. Conversely, guidelines outline standards that reviewers may assess implicitly, without formulating specific questions.

CHAPTER 7

VALIDITY THREAT MITIGATION

This chapter discusses the validity threats in this study and the strategies employed to mitigate them. The goal is to clarify the methodological boundaries of the analysis and to justify the academic rigor of the findings derived from the ICLR 2024 peer review corpus.

7.1 Construct Validity

Construct validity concerns whether the study accurately captures the intended constructs, namely, reviewer inquiry topics and their alignment with institutional reviewer guidelines.

A central threat arises from extracting semantic units from unstructured review text. The content of *Questions* in reviews is not always explicitly formulated and may contain referential statements or general comments. This risk was mitigated through a multi-stage extraction pipeline. Non-informative entries were filtered before semantic processing. All excluded samples were manually audited. With one marginal exception, the filtered entries did not contain actionable inquiries, confirming that the filtering stage removed noise without materially suppressing valid reviewer questions.

Another construct validity risk concerns LLM-based semantic segmentation. To prevent hallucination or semantic drift, all extracted question chunks were mathematically grounded to the source text using sentence-level cosine similarity. Only high-fidelity extractions were retained. Semantic coherence was further validated using a multi-agent LLM-as-a-Judge protocol combined with majority voting.

Topic construction introduces an additional risk of subjective consolidation. To mitigate this, topic discovery was performed across multiple independent runs and heterogeneous LLM architectures. The resulting candidate topics were consolidated using both topic names and descriptions rather than labels alone. The semantic stability of the final schema was confirmed through a dual verification. Independent grouping by the primary researcher and a reasoning LLM yielded highly consistent topic structures. A senior academic expert subsequently audited the consolidation. This process reduces researcher confirmation bias and supports the validity of the final 13-topic schema.

LLM calibration presents an additional construct validity concern. Models trained with reinforcement learning from human feedback are known to produce overconfident probability

distributions. To mitigate the misinterpretation of confidence, probability margin and entropy thresholds were deliberately set near extreme values. These measures were used diagnostically rather than as strict decision rules, reducing the risk of overstating semantic certainty.

A final construct validity concern arises in the alignment between reviewer inquiries and institutional guidelines. Guideline texts often contain abstract or multifaceted directives that can be semantically associated with multiple evaluation criteria. To mitigate interpretive bias, guideline directives were extracted from official venue documentation and segmented into meaning-preserving units. These units were independently mapped to the data-driven topic schema using both LLM-based classification and human annotation. Assignments were based strictly on explicit textual content rather than inferred intent, and the final mappings were cross-checked by a senior academic expert. This procedure reduces annotation bias and supports the construct validity of the measured guideline–practice gap.

7.2 Internal Validity

Internal validity addresses whether the observed patterns arise from the analytical design rather than unintended methodological influences.

A key threat was intra-submission redundancy, as multiple reviews of the same paper often raise overlapping concerns. This was mitigated by aggregating reviews at the submission level and sampling a single review per manuscript before analysis. This design reduces semantic duplication and increases independence between samples.

Class imbalance presents another internal threat. Borderline ratings dominate the corpus, while extreme ratings are rare. Stratified sampling based on review ratings was applied to preserve the full spectrum of reviewer sentiment and prevent dominance by majority classes.

To assess sensitivity to sampling variation, the full extraction and topic assignment pipeline was independently applied to a held-out validation dataset constructed using the same stratification logic. As detailed in Appendix B.6, the resulting rating distributions and topic frequency profiles closely matched those of the main dataset, indicating that the findings are not an artifact of a specific sample.

Potential LLM-related bias in topic assignment was mitigated by randomizing the order of topics in the prompt to reduce positional effects. Deterministic inference settings were also used. All LLM calls were executed with a temperature value of zero to minimize stochastic variability.

7.3 External Validity

External validity concerns the generalizability of the findings beyond the studied context.

This study is limited to a single iteration of ICLR venue. The analysis focuses exclusively on ICLR 2024, which operates under a double-blind review model and employs a highly structured review form. Reviewer behavior and inquiry patterns may differ across venues, disciplines, or review formats. As a result, the observed topic frequencies should not be generalized to all academic peer review settings.

However, the methodological framework itself is not venue-specific. The source-anchored extraction pipeline, multi-stage validation strategy, probabilistic topic assignment, and guideline alignment methodology are transferable to other conferences and peer review platforms. While the specific topic schema may change in different contexts, the framework supports comparative analysis across venues and time.

7.4 Reliability and Reproducibility

Reliability and reproducibility concern the consistency of the analysis and the ability to replicate the study.

Data acquisition was performed using the official OpenReview API, ensuring deterministic access to publicly available artifacts. All data were stored in a structured database with schema validation enforced at ingestion. The software environment was fully specified using Conda and Poetry, enabling dependency-level reproducibility.

LLM-based components introduce inherent nondeterminism. This risk was mitigated by enforcing zero temperature across all model calls and by freezing prompt templates after iterative refinement. Extraction fidelity was controlled using deterministic similarity thresholds. Semantic validation employed multiple independent LLM judges, and agreement was quantified using standard inter-rater reliability metrics.

During soft classification, schema adherence was verified using the Total Valid Mass metric. The minimum observed value of 0.9993 indicates negligible probability leakage outside the predefined topic indices. This confirms stable model behavior and reliable adherence to the topic schema.

Together, these measures ensure that the analytical pipeline can be re-executed to produce consistent results under identical conditions, subject to the availability of the same model versions.

CHAPTER 8

CONCLUSION

This thesis has presented a data-driven investigation into the alignment between institutional guidelines and community practice within the academic peer review ecosystem. We proposed a framework to analyze the `Questions` field of the ICLR 2024 corpus. Using this framework, we quantified the guideline-practice gap by contrasting reviewer inquiries against the prescribed criteria of top-tier venues. This concluding chapter synthesizes the research findings, discusses their implications for stakeholders, outlines the study's limitations and future directions, and offers final remarks on the future of scientific evaluation.

8.1 Research Summary

Academic venues establish reviewer guidelines to enable standardized evaluation. This mechanism is critical given the exponential growth of submissions and the influx of junior reviewers at top-tier conferences. This thesis proposes a framework designed to categorize reviewer inquiries and systematically measure their alignment with institutional mandates.

To achieve this, we developed a framework designed to decompose unstructured peer review text into semantically coherent units. The methodology employed a *Source-Anchored Information Extraction* pipeline utilizing Large Language Models (LLMs) and embedding similarity. Applied to a stratified sample of 427 reviews from the ICLR 2024 corpus, this process yielded 760 validated question chunks. These chunks were finalized via a multi-agent consensus protocol, serving as the basis for generating and consolidating latent topics. We then employed a *Logit-Based Soft Classification* technique to assign these topics to the corpus. The topic distribution enabled a quantitative comparison between the *practiced criteria* of reviewers and the *prescribed criteria* of reviewer guidelines.

8.2 Synthesis of Findings

The core objective of this research was to answer two primary research questions regarding the thematic composition of reviewer inquiries and their alignment with formal criteria.

8.2.1 The Schema of Criticism (RQ1)

The first research question examined the main topics of criticism within peer reviews. Through the proposed framework, we established that reviewer inquiries are clustered under a schema of 13 distinct topics.

Semantic Properties and Distribution (RQ1.1): Addressing RQ1.1, the analysis established the semantic definitions and frequency distribution of reviewer critique. We observed a distinct pattern of thematic prevalence. *Inquiry About Methodology* (16.18%), *Comparative Analysis* (13.16%), and *Applicability and Limitations* (11.97%) are the most frequent categories. These patterns demonstrate that the bulk of reviewer inquiry is directed toward technical implementation and research positioning.

Structural Validity and Separability (RQ1.2): To validate the coherence of this schema, we quantitatively assessed the distinctness of the generated topics. The Jensen-Shannon Divergence (JSD) [49] analysis demonstrated that the topics occupy orthogonal semantic spaces, with high divergence scores (> 0.90) for the majority of pairwise comparisons. Furthermore, the low Normalized Entropy scores observed in the soft-classification task confirm that reviewer questions are generally focused and unambiguous. These metrics show that the topic schema is structurally sound and that the identified domains are semantically separable.

8.2.2 The Guideline-Practice Gap (RQ2)

The second research question investigated the alignment between the data-driven topics derived from ICLR 2024 and the official reviewing guidelines of the ten top-tier AI venues. The analysis mapped the 13 identified topics against these institutional protocols to determine the extent of alignment between the *practiced* criteria of the community and the *prescribed* criteria of the field.

Unwritten Norms (RQ2.1): We identified specific topics that emerge frequently in reviewer feedback despite being absent from formal guidelines. Based on the quantitative analysis of the ICLR 2024 corpus, we found that *Unwritten Norms* account for 67.50% of all inquiries. Crucially, *Inquiry About Methodology* is the most frequent topic in the dataset, yet it is not explicitly mandated by the ICLR directives. This divergence extends to the broader field; for instance, *Inquiry About Presentation and Figures* accounts for 8.29% of the dataset, yet it appears as an explicit directive in only two of the eight analyzed unique guideline documents. This reveals that the community enforces a set of standards that are often not present in formal guidelines.

Neglected Criteria (RQ2.2): Conversely, we identified explicitly requested criteria that are systematically de-prioritized or ignored in practice. A prime example is *Ethical Considerations*. While it is mandated by six of the eight analyzed guideline documents, it ranks as the least frequent topic in the ICLR dataset (0.66%). Similarly, within the specific context of ICLR guidelines, mandated topics such as *Motivation and Contribution* appeared with relatively low

frequency (4.34%). This demonstrates that even widely adopted institutional mandates do not necessarily translate into active reviewer inquiry.

8.3 Implications

The implications of this study extend to three distinct groups within the academic community.

For Conference Organizers: The observed gap suggests that current guidelines may be underspecified relative to the complexity of the evaluation task. For instance, *Methodology* is the main focus of reviewers, yet it is not explicitly stated in ICLR guidelines. This indicates a need to update guidelines to reflect the actual standards of the community. Aligning guidelines with practice could reduce variance in the review process by making implicit norms explicit.

For Reviewers: The derived schema serves as an empirical reference for novice reviewers, particularly those new to the ICLR community. By observing the practiced criteria alongside the official guidelines, reviewers can develop a holistic understanding of the evaluation landscape. This insight allows them to calibrate their feedback, ensuring their inquiries address both the explicit mandates of the venue and the implicit expectations of the community.

For Authors: The schema of inquiries provides authors with a data-driven roadmap for manuscript preparation. Understanding that *Methodology*, *Comparative Analysis*, and *Applicability and Limitations* constitute over 40% of all inquiries allows authors to preemptively address these areas, potentially reducing the volume of critical questions during the rebuttal phase.

8.4 Limitations and Future Works

This study establishes a robust framework for analyzing peer review dynamics. Nevertheless, several limitations remain that outline valuable directions for future research.

One significant limitation is the study's focus on a single iteration of a specific venue. The empirical findings are currently bounded by the reviewing culture of ICLR 2024. To determine whether the observed guideline gap is a universal characteristic of the AI research community or a venue-specific anomaly, future work should expand this analysis to other top-tier conferences hosted on OpenReview. Furthermore, since the ICLR 2025 dataset retains the same structured review format, extending this study to that iteration would enable a longitudinal analysis to track the adaptation and evolution of community standards over time.

Computational constraints associated with high-reasoning LLMs necessitated the use of a stratified sample of the ICLR 2024 review data. Although validation against a held-out set confirms the representativeness of the sample, processing the complete corpus could yield minor refinements in the specific proportions of topic assignments. Additionally, this study exclusively analyzed the `Questions` field. Given that a subset of reviewers embed inquiries within the `Weaknesses` section, future iterations could apply the extraction pipeline to these fields to capture a more exhaustive set of reviewer inquiries.

A specific methodological limitation concerns model calibration. The use of RLHF-tuned models introduces a calibration bias toward high confidence. While the Probability Margin metric was interpreted with strict thresholds to mitigate this, the absolute probability values should be viewed as relative ranking scores rather than calibrated probabilities.

Finally, the methodological implementation of the *LLM-as-a-Judge* protocol utilized a unified prompt. This prompt is optimized for the collective agreement of LLMs. Future research could explore model-specific prompt optimization strategies to potentially further enhance inter-rater reliability metrics. Moreover, the integration of emerging compact models into the non-informative text filtering stage offers a parallel direction for optimization. The use of these advanced models could improve both the cost-efficiency and precision of the extraction pipeline by leveraging their superior reasoning capabilities.

8.5 Concluding Remarks

Peer review serves as the fundamental mechanism for ensuring the quality and validity of scientific contributions. Despite this critical role, the process is frequently influenced by implicit expectations that are not captured in formal documentation. This thesis has empirically characterized these expectations by analyzing the ICLR 2024 corpus and systematically mapping it against the evaluation criteria of the ten top-tier AI venues. Our findings demonstrate that the practice of peer review operates differently from its established guidelines. By quantifying this divergence, this work establishes the empirical foundation necessary to design more consistent evaluation systems, while providing actionable guidance for the authors and reviewers navigating the peer review process.

REFERENCES

[1] National Information Standards Organization (NISO). Peer review terminology. Accessed: [2025-11-26].

[2] Ines Arous, Jie Yang, Mourad Khayati, and Philippe Cudre-Mauroux. Peer grading the peer reviews: A dual-role approach for lightening the scholarly paper review process. In *Proceedings of the Web Conference 2021*, WWW '21, page 1916–1927, New York, NY, USA, 2021. Association for Computing Machinery.

[3] Moritz Staudinger, Wojciech Kusa, Florina Piroi, and Allan Hanbury. An analysis of tasks and datasets in peer reviewing. In Tirthankar Ghosal, Amanpreet Singh, Anita Waard, Philipp Mayr, Aakanksha Naik, Orion Weller, Yoonjoo Lee, Shannon Shen, and Yanxia Qin, editors, *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 257–268, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[4] Google Scholar. Top publications - engineering & computer science. `https://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng`. Accessed: 2025-12-03.

[5] Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In Marilyn Walker, Heng Ji, and Amanda Stent, editors, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[6] Tirthankar Ghosal, Sandeep Kumar, Prabhat Kumar Bharti, and Asif Ekbal. Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *PLOS ONE*, 17(1):1–29, 01 2022.

[7] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[8] Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. Topicgpt: A prompt-based topic modeling framework. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984, 2024.

[9] Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. Moprd: A multidisciplinary open peer review dataset. *Neural Computing and Applications*, 35:1–16, 09 2023.

[10] Shruti Singh, Mayank Singh, and Pawan Goyal. Compare: A taxonomy and dataset of comparison discussions in peer reviews. In *2021 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 238–241, 2021.

[11] Souvic Chakraborty, Pawan Goyal, and Animesh Mukherjee. Aspect-based sentiment analysis of scientific reviews. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, JCDL '20, page 207–216, New York, NY, USA, 2020. Association for Computing Machinery.

[12] Tirthankar Ghosal, Rajeev Verma, Asif Ekbal, and Pushpak Bhattacharyya. DeepSentiPeer: Harnessing sentiment in review texts to recommend peer review decisions. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1120–1130, Florence, Italy, July 2019. Association for Computational Linguistics.

[13] Prabhat Kumar Bharti, Shashi Ranjan, Tirthankar Ghosal, Mayank Agrawal, and Asif Ekbal. Peerassist: leveraging on paper-review interactions to predict peer review decisions. In *International Conference on Asian Digital Libraries*, pages 421–435. Springer, 2021.

[14] Alessandro Checco, Lorenzo Bracciale, Pierpaolo Loreti, Stephen Pinfield, and Giuseppe Bianchi. Ai-assisted peer review. *Humanities and Social Sciences Communications*, 8, 01 2021.

[15] Kayvan Kousha and Mike Thelwall. Artificial intelligence to support publishing and peer review: A summary and review. *Learned Publishing*, 37, 08 2023.

[16] Weizhe Yuan, Pengfei Liu, and Graham Neubig. Can we automate scientific reviewing? *J. Artif. Int. Res.*, 75, December 2022.

[17] Ruiyang Zhou, Lu Chen, and Kai Yu. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, Torino, Italia, May 2024. ELRA and ICCL.

[18] Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. AgentReview: Exploring peer review dynamics with LLM agents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1208–1226, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[19] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[20] Xiaohao Yang, He Zhao, Dinh Phung, Wray Buntine, and Lan Du. Llm reading tea leaves: Automatically evaluating topic models with large language models. *Transactions of the Association for Computational Linguistics*, 13:357–375, 2025.

[21] Han Wang, Nirmalendu Prakash, Nguyen Khoi Hoang, Ming Shan Hee, Usman Naseem, and Roy Ka-Wei Lee. Prompting large language models for topic modeling. *2023 IEEE International Conference on Big Data (BigData)*, pages 1236–1241, 2023.

[22] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[23] Leland McInnes and John Healy. Accelerated hierarchical density based clustering. *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 33–42, 2017.

[24] Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2):18, 2024.

[25] Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. Is automated topic model evaluation broken? the incoherence of coherence. *Advances in neural information processing systems*, 34:2018–2033, 2021.

[26] Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. Revisiting automated topic model evaluation with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9348–9357, Singapore, December 2023. Association for Computational Linguistics.

[27] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. A survey on llm-as-a-judge. *ArXiv*, abs/2411.15594, 2024.

[28] Prabhat Bharti, Tirthankar Ghosal, Mayank Agarwal, and Asif Ekbal. A method for automatically estimating the informativeness of peer reviews. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, pages 280–289, 2022.

[29] Cecilia Superchi, José Antonio González Alastrué, Ivan Solà, Erik Cobo, Darko Hren, and Isabelle Boutron. Tools used to assess the quality of peer review reports: A methodological systematic review. *BMC Medical Research Methodology*, 19, 03 2019.

[30] Michael Fromm, Evgeniy Faerman, Max Berrendorf, Siddharth Bhargava, Ruoxia Qi, Yao Zhang, Lukas Dennert, Sophia Selle, Yang Mao, and Thomas Seidl. Argument mining driven analysis of peer-reviews. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 4758–4766, 2021.

[31] Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. Argument mining for understanding peer reviews. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and*

Short Papers), pages 2131–2137, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[32] Yichi Zhang, Fang-Yi Yu, Grant Schoenebeck, and David Kempe. A system-level analysis of conference peer review. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, EC '22, page 1041–1080, New York, NY, USA, 2022. Association for Computing Machinery.

[33] OpenReview.net. About openreview. `https://openreview.net/about`, 2025. Accessed: 2025-08-12.

[34] OpenReview.net. Openreview terms of use. `https://openreview.net/legal/terms`, 2024. Accessed: 2024-11-07.

[35] OpenReview.net. Openreview privacy policy. `https://openreview.net/legal/privacy`, 2024. Accessed: 2024-11-07.

[36] Zhuanlan Sun. Textual features of peer review predict top-cited papers: An interpretable machine learning perspective. *Journal of Informetrics*, 18(2):101501, 2024.

[37] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[38] Maurice G Kendall. The treatment of ties in ranking problems. *Biometrika*, 33(3):239–251, 1945.

[39] Lauriston L Keown and A Ralph Hakstian. Measures of association for the component analysis of likert scale data. *The Journal of Experimental Education*, 41(3):22–27, 1973.

[40] Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv*, abs/2507.06261, 2025.

[41] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[42] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[43] Anthropic. Claude 3.7 Sonnet system card. System card, Anthropic, 2025.

[44] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[45] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[46] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova Dassarma, Dawn Drain, Stanislav Fort, Deep Ganguli, T. J. Henighan, Nicholas Joseph, Saurav Kadavath, John Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt,

Neel Nanda, Catherine Olsson, Dario Amodei, Tom B. Brown, Jack Clark, Sam McCandlish, Chris Olah, Benjamin Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *ArXiv*, abs/2204.05862, 2022.

[47] Yashar Deldjoo. Understanding biases in chatgpt-based recommender systems: Provider fairness, temporal stability, and recency. *ACM Trans. Recomm. Syst.*, 4(2), November 2025.

[48] Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[49] J. Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991.

[50] Solomon Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.

[51] The IEEE/CVF Conference on Computer Vision and Pattern Recognition. Cvpr 2026 reviewer guidelines. Accessed: [2025-12-17].

[52] Neural Information Processing Systems. 2025 reviewer guidelines. Accessed: [2025-12-18].

[53] International Conference on Learning Representations. Iclr 2024 reviewer guide. Accessed: [2025-12-18].

[54] International Conference on Learning Representations. Iclr 2025 reviewer guide. Accessed: [2025-12-18].

[55] International Conference on Machine Learning. Icml 2025 reviewer instructions. Accessed: [2025-12-18].

[56] European Conference on Computer Vision. Eccv 2024 reviewer guidelines. Accessed: [2025-12-18].

[57] European Conference on Computer Vision. Eccv 2024 reviewer tutorial slides. Accessed: [2025-12-18].

[58] ACL Rolling Review. Arr reviewer guidelines. Accessed: [2025-12-18].

[59] Association for the Advancement of Artificial Intelligence. Instructions for aaai 2026 reviewers. Accessed: [2025-12-17].

[60] IEEE Author Center. About the peer review process. Accessed: [2025-12-18].

[61] IEEE Author Center. Become an ieee reviewer. Accessed: [2025-12-18].

[62] ICLR 2024 Program Chairs. Please keep revision to 9 pages. https://openreview.net/forum?id=HKkiX32Zw1&noteId=JDOJ0CoWya. Accessed: 2025-12-08.

[63] ICLR 2024 Senior Area Chair. author-reviewer discussion about to end. https://openreview.net/forum?id=rWHMe2O5VW&noteId=SdboPj8eet. Accessed: 2025-12-08.

APPENDIX A

DATA SOURCE AND CONTEXT

A.1 OpenReview Interface

To provide visual context for the data model described in Section 3.1.1, this section presents screenshots of the OpenReview web interface. Figure 13 represents a submitted manuscript and Figure 14 illustrates an example peer review. The examples correspond to the submission titled "*Mask-Based Modeling for Neural Radiance Fields*"[1].

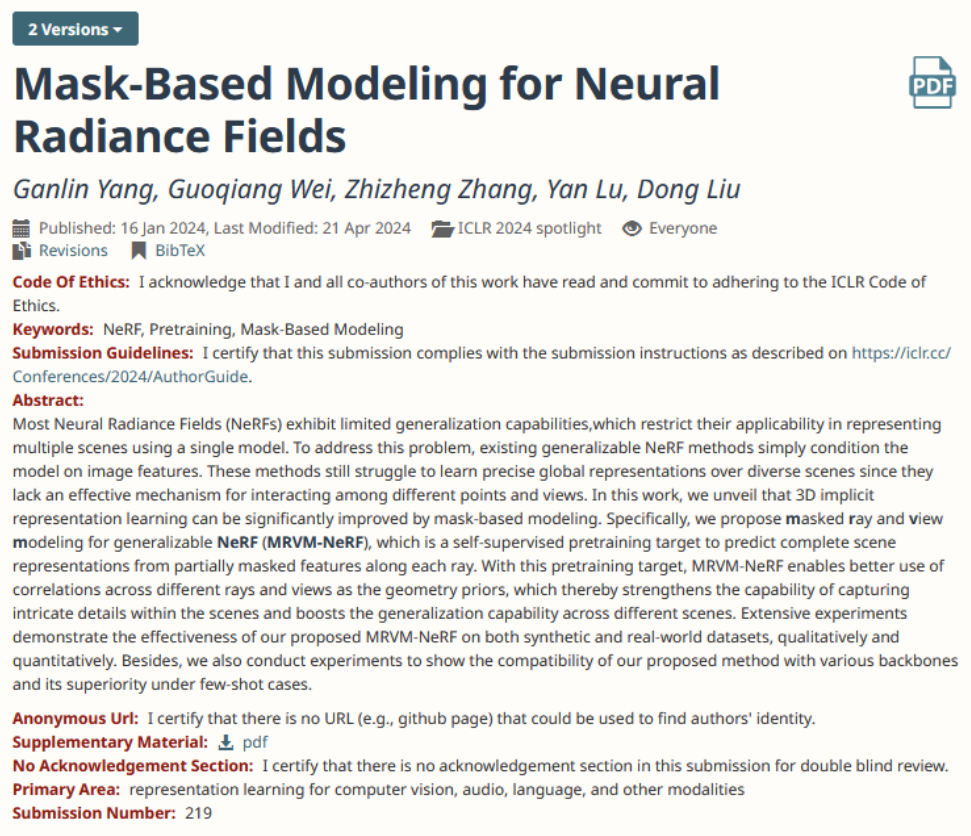


Figure 13: Screenshot of a submission page on OpenReview.

[1] https://openreview.net/forum?id=SEiuSzlD1d

Official Review of Submission219 by Reviewer Jc3W

Official Review by Reviewer Jc3W 04 Nov 2023, 17:36 (modified: 10 Nov 2023, 20:05) Everyone

Revisions

Summary:
The paper presents several significant improvements to the standard generalizable-NeRF framework, in which a NeRF is trained on a set of scenes and is used at inference on a novel scene without training. It successfully incorporates some recent advances from self-supervised representation learning, which include the use of masked prediction, exponential moving average learning, using transformer-based backbones. The main contribution is in the way these techniques are introduced into the NeRF framework, with the particular choice of the student network acting in the 'fine' sampling branch with random masking and the teacher network acting on the 'coarse' branch. Results verify the method, with significant and consistent improvements in the different metrics across the board.

Soundness: 4 excellent
Presentation: 4 excellent
Contribution: 3 good
Strengths:
1] The proposed method shows very good performance compared to the baselines, both significant and consistent, across all experimentation. This is apparent visually and quantitatively, though an extensive set of experiments. 2] The idea of introducing masking into the training scheme is well motivated, as a means for improving generalization to new scenes. It extends the standard pipeline quite naturally and does not impact inference time complexity. The decision to mask at two levels, within the ray and across views is interesting. 3] Incorporating student-teacher learning within NeRF is also quite a natural thing to do. It gives further regularization that is likely responsible for the improved generalization. 4] Paper is well written and organized, especially the method and experiments that explain and demonstrate the advantages very clearly.

Weaknesses:
1] The two main additions - of EMA and masking are tested together. There is a lack in understanding how dependent they are. Firstly, if one could be used without the other. And if so, secondly, what are the individual contributions. 2] Sampling is not well specified. The extra 'fine' samples - whether they are they taken, as in the classical NeRF, according to an initial estimation of density from the coarse sample. And in this respect - shouldn't the masking prediction be focused at the more important (close-to-surface) high density locations, rather than at the coarse sampling? 3] Impact on training time is not clear. Even though training time and procedures are very clearly presented in the appendix, I am missing what is the addition in comparison to the baseline, and which components are responsible for it? 4] Clarity [minor]: (i) I would suggest improving Figure 1 to include some of the notations (g, h, etc') ; (ii) What are the blue dashed arrows in Figure 2 supposed to represent?

Questions:
0] Please relate to above 'weaknesses' In addition: 1] Have you tried comparing to per-scene NeRFs? 2] Could you provide an ablation showing the contribution of each of the components? In particular - applying only one of EMA / Masking? 5] How sensitive is the choice of \lambda, the weight of the mrvm-loss?

Flag For Ethics Review: No ethics review needed.
Rating: 6: marginally above the acceptance threshold
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct: Yes

Figure 14: Screenshot of a review on OpenReview.

A.2 Scoring Rubrics and Definitions

This section documents the verbatim scoring rubrics employed in the ICLR 2024 review process. The five-point reviewer confidence scale is shown in Table 19, while the overall rating metric appears in Table 20. The four-point rubric for soundness, presentation, and contribution is provided in Table 21.

Table 19: Verbatim descriptions for ICLR 2024 confidence scores.

Score	Description
1	*You are unable to assess this paper and have alerted the ACs to seek an opinion from different reviewers.*
2	*You are willing to defend your assessment, but it is quite likely that you did not understand the central parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.*
3	*You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.*
4	*You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.*
5	*You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.*

Table 20: Verbatim descriptions for ICLR 2024 rating scores.

Score	Description
1	*Strong reject*
3	*Reject, not good enough*
5	*Marginally below the acceptance threshold*
6	*Marginally above the acceptance threshold*
8	*Accept, good paper*
10	*Strong accept, should be highlighted at the conference*

Table 21: Shared scoring rubric for soundness, presentation, and contribution for ICLR 2024.

Score	Description
1	*poor*
2	*fair*
3	*good*
4	*excellent*

A.3 Administrative Comments

This section contains the verbatim text of the only two comments made by the Program Chair and Senior Area Chair in the constructed ICLR 2024 dataset.

Program Chair Comment

> *"It is fine to re-order the appendix to highlight the additional figure first. But in light with having a consistent policy for managing 7K submissions, please keep the main paper in the revision to 9 pages." [62]*
>
> *— ICLR 2024 Program Chairs*

Senior Area Chair Comment

> *"Dear reviewers, as the deadline for the author-reviewer discussion phase is approaching, please take a look at the authors' reply (if you have not done so) and see if your concerns have been addressed." [63]*
>
> *— ICLR 2024 Senior Area Chair*

APPENDIX B

EXTENDED STATISTICAL ANALYSIS

B.1 Supplementary EDA Plots for Quantitative Metrics

This section presents the supplementary frequency distributions and histograms for the quantitative metrics discussed in Section 3.3.4. These visualizations are organized into two distinct categories: review-level score distributions and submission-level aggregated metrics.

B.1.1 Review-Level Score Distributions

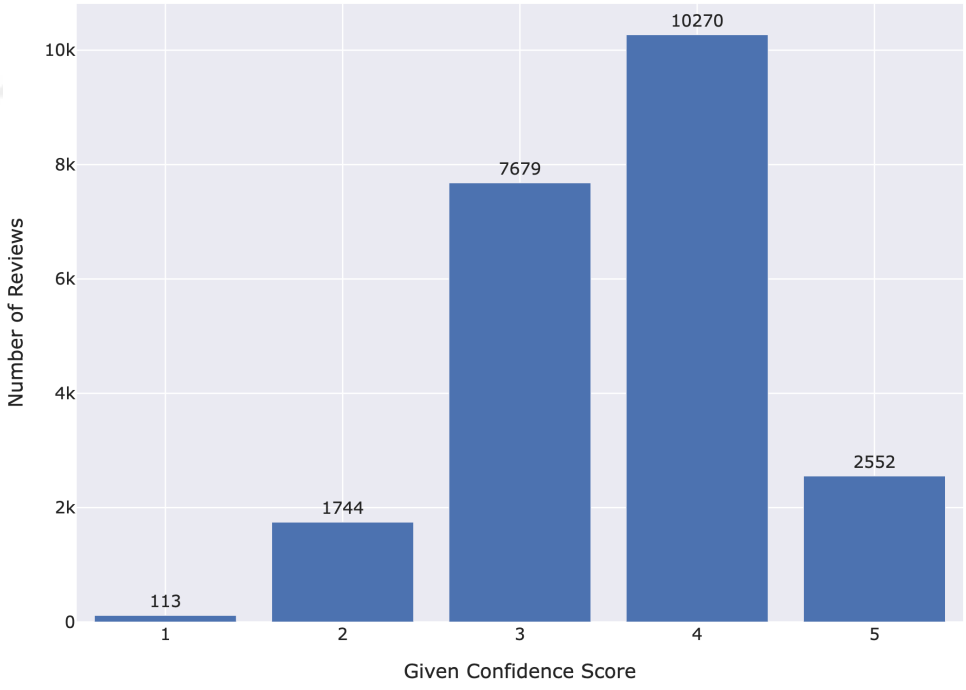


Figure 15: Distribution of confidence scores across individual reviews.

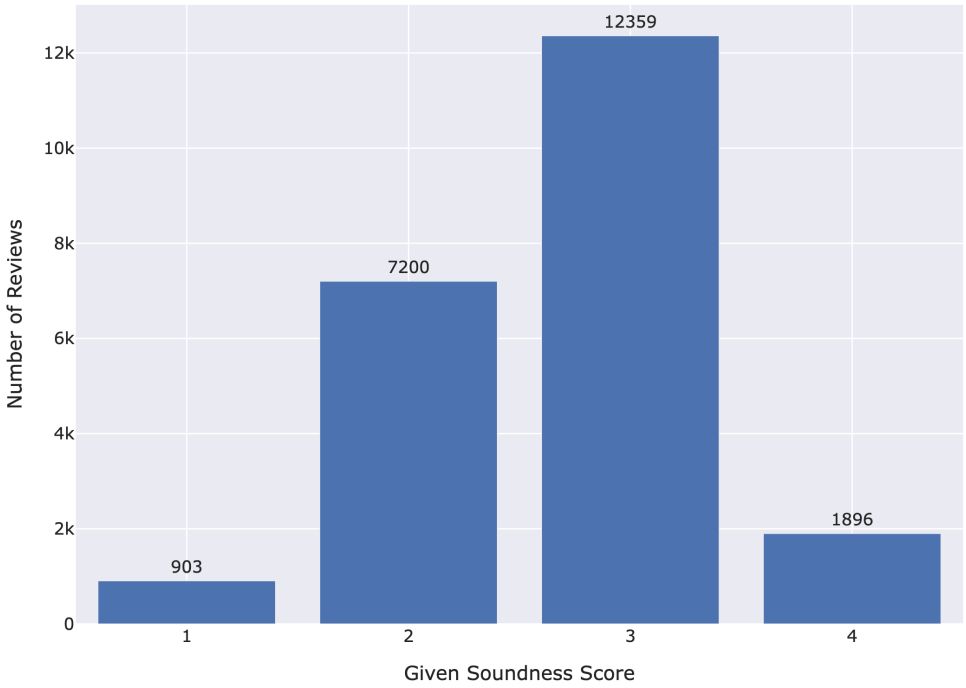


Figure 16: Distribution of soundness scores across individual reviews.

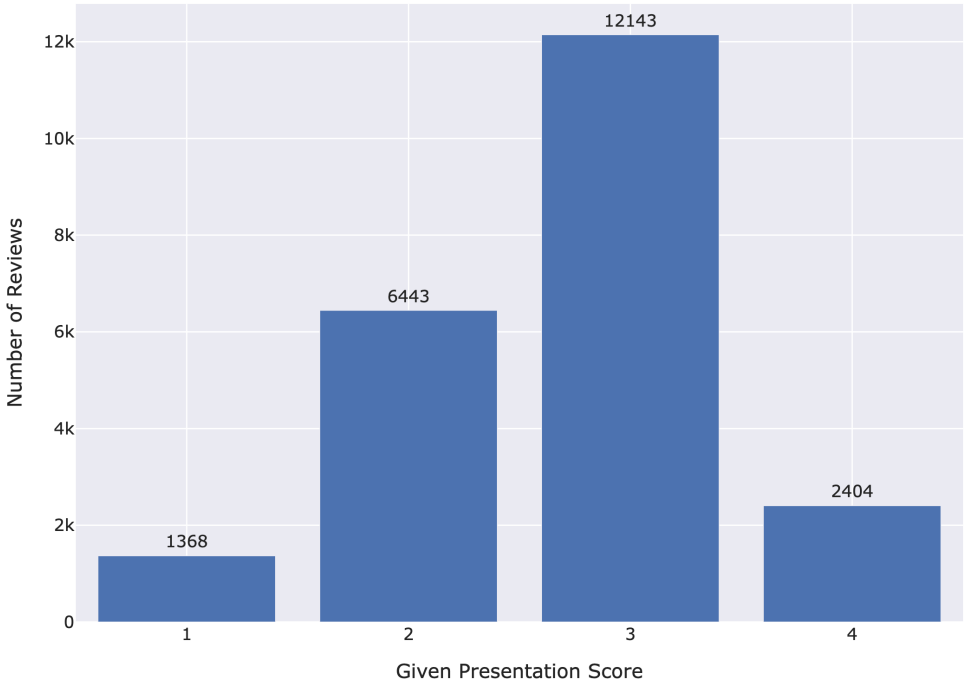


Figure 17: Distribution of presentation scores across individual reviews.

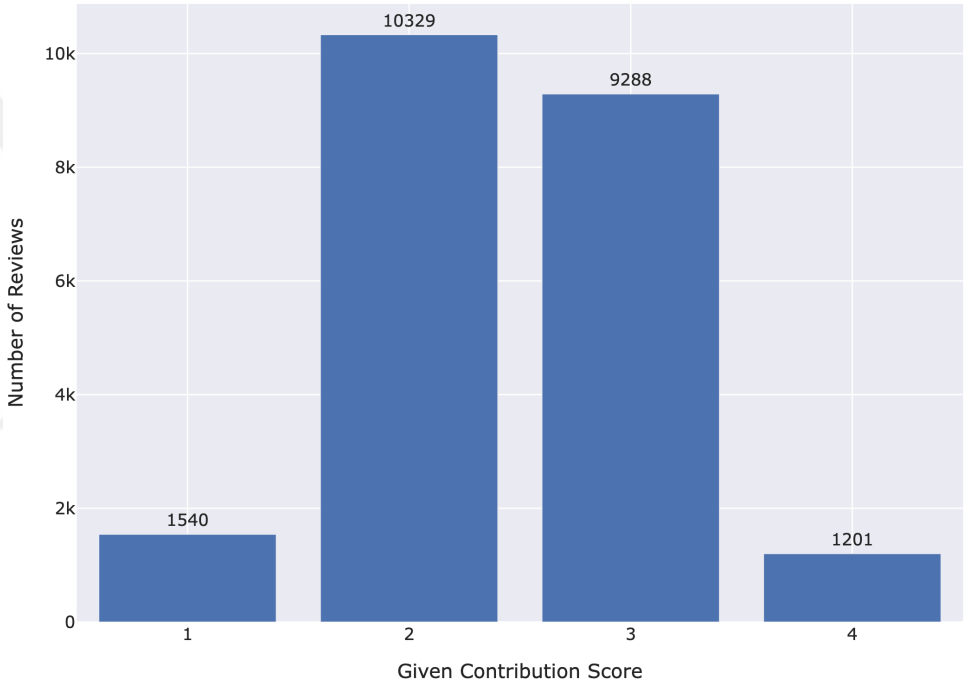


Figure 18: Distribution of contribution scores across individual reviews.

B.1.2 Submission-Level Aggregated Metrics

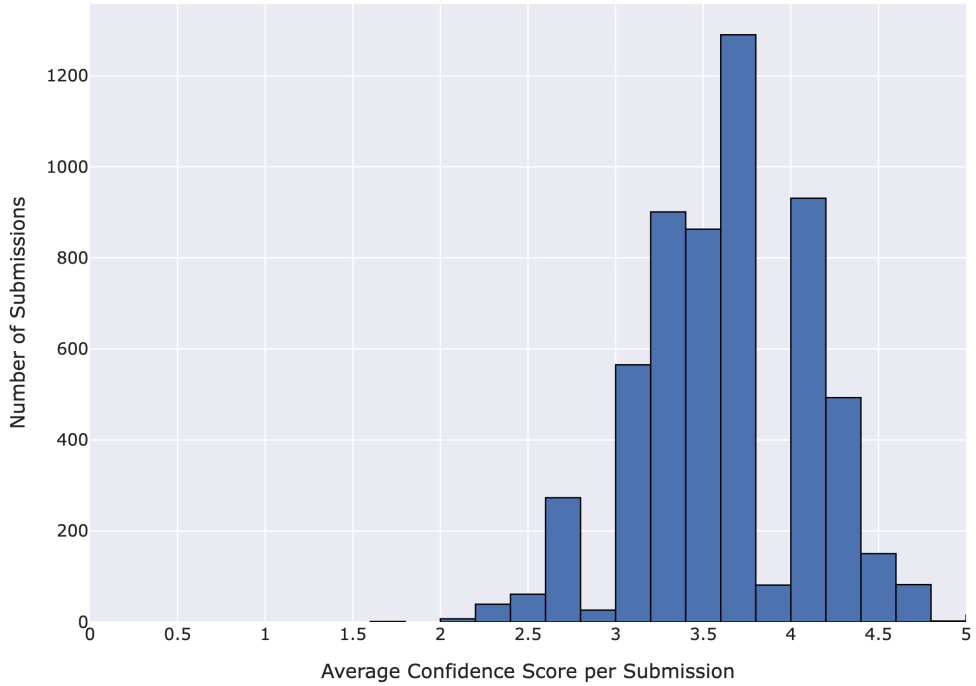


Figure 19: Distribution of average confidence scores per submission.

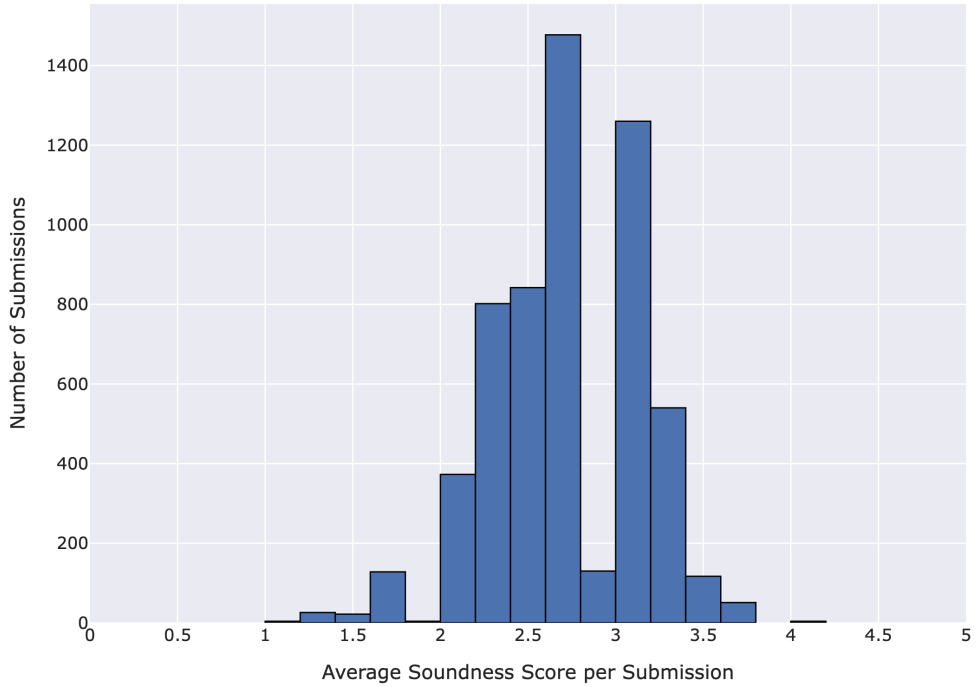


Figure 20: Distribution of average soundness scores per submission.

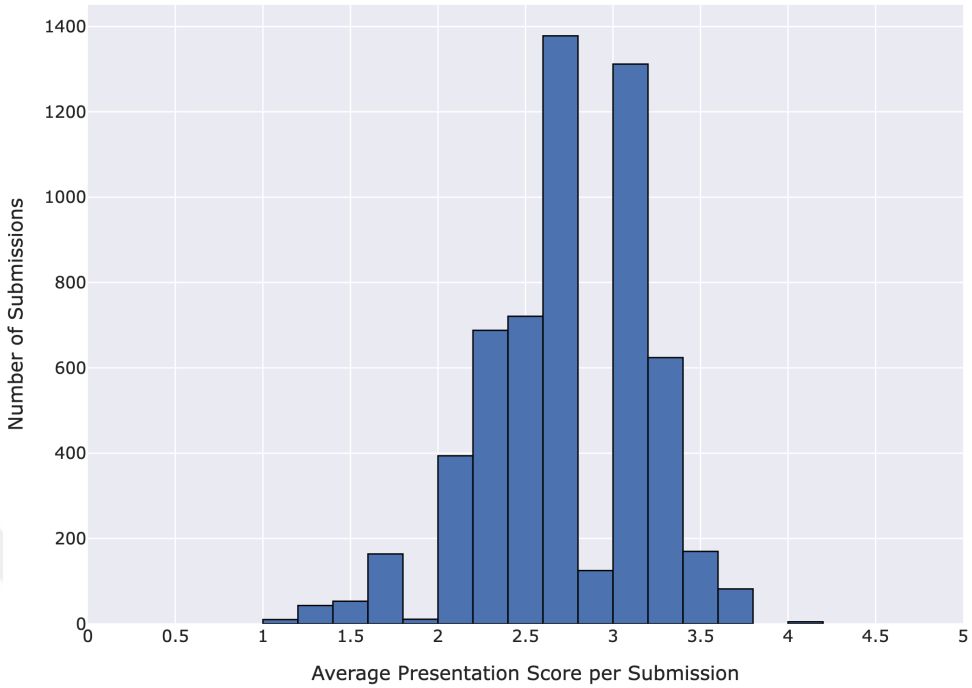


Figure 21: Distribution of average presentation scores per submission.

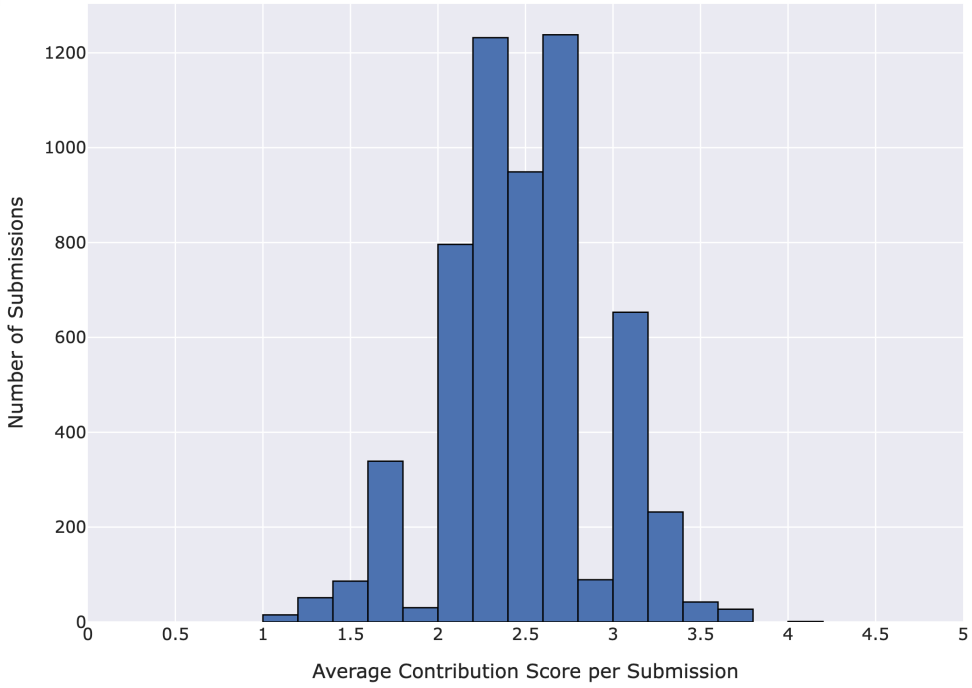


Figure 22: Distribution of average contribution scores per submission.

B.2 Additional EDA Plots for Textual Metrics

This section provides the histograms for the word counts of the `Strengths`, `Weaknesses`, and `Questions` fields in reviews. These distributions represent review-level statistics, calculated for each individual review rather than aggregated by submission. To ensure the readability of the primary distribution density, the x-axes of these plots have been visually truncated at the 99th percentile, effectively excluding the extreme outliers discussed in the main text.

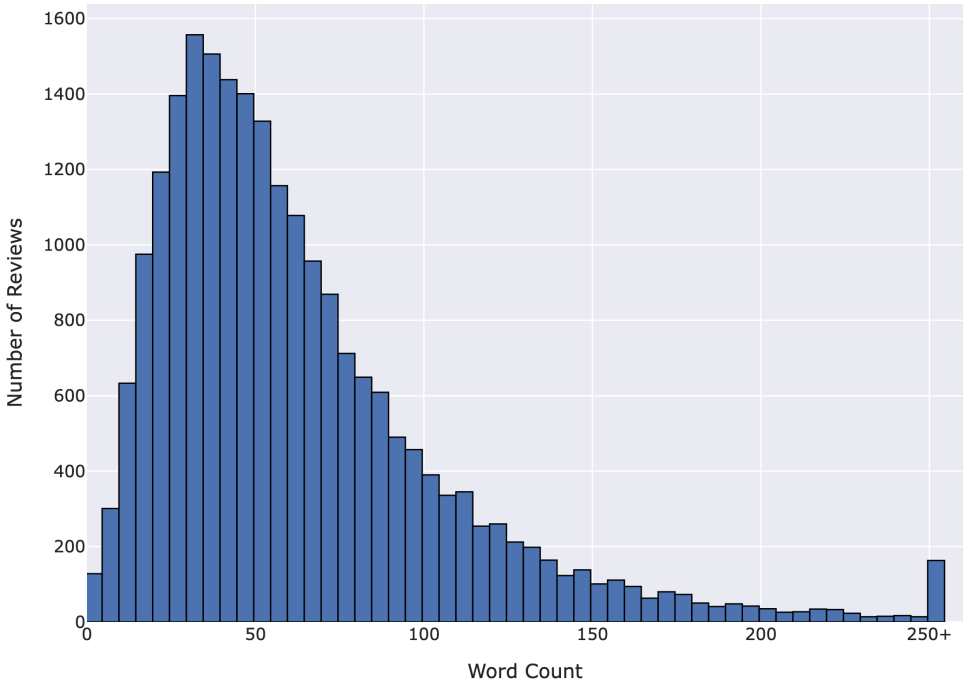


Figure 23: Distribution of word counts for the *Strengths* field in reviews.

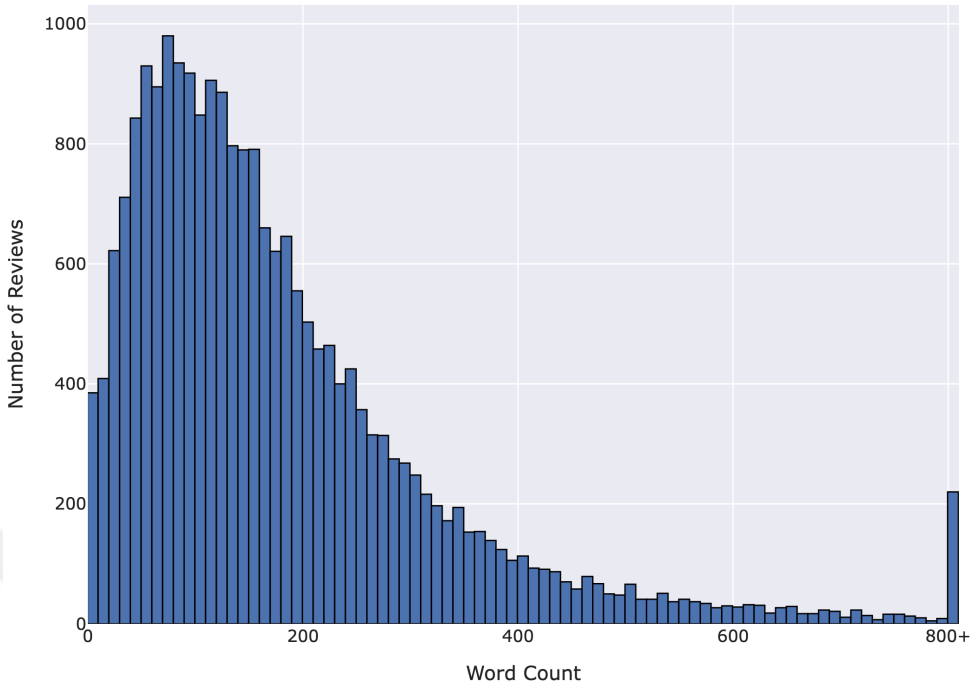


Figure 24: Distribution of word counts for the *Weaknesses* field in reviews.

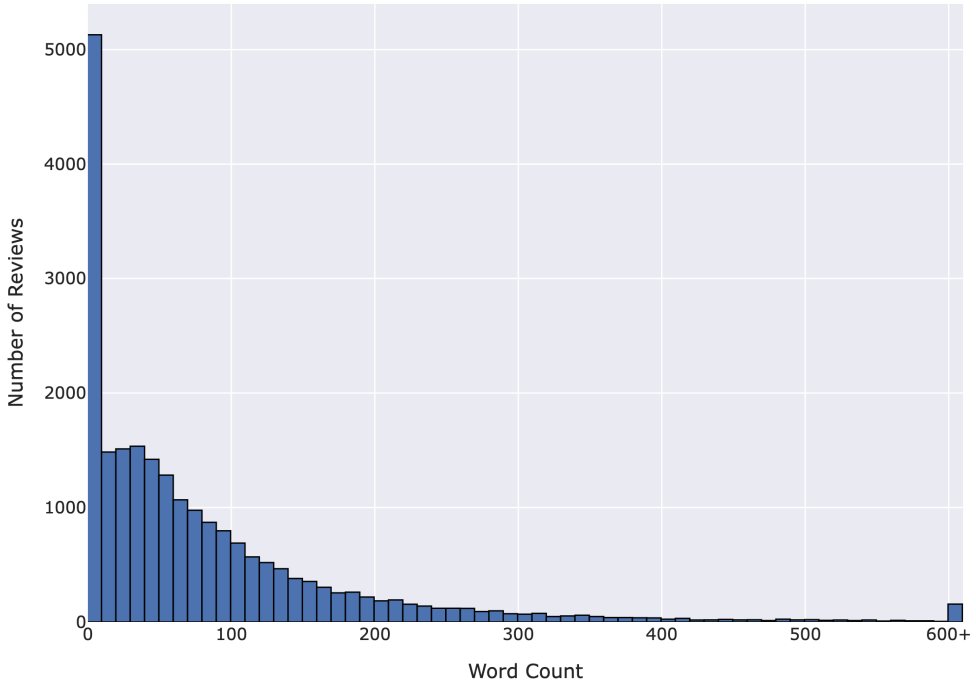


Figure 25: Distribution of word counts for the *Questions* field in reviews.

B.3 Statistical Significance Table

This section provides the matrix of p-values corresponding to the Kendall's Tau correlation coefficients presented in Section 3.3.5. Each p-value indicates the probability of observing the given correlation by chance if no true relationship existed.

Interpreting these p-values require careful consideration of the study's statistical context. Since the data used in this method is large (+20.000), the analysis has extremely high statistical power. It makes it likely that even very weak, practically unimportant relationships will be flagged as statistically significant ($p < 0.05$). While a formal procedure like the Bonferroni correction could be applied to account for the 28 pairwise comparisons, its primary effect would be to render a few of the weakest correlations statistically non-significant.

However, this outcome would only reinforce the central argument of this analysis: in a dataset of this size, the most reliable indicator of a relationship's importance is not its p-value, but rather the magnitude of its effect size (as measured by the Kendall's Tau coefficient). Therefore, our analytical approach prioritizes effect size. For better readability, some values are highlighted in the Table 22 based on their p-value being greater than 0.05. These values are rare relationship pairs that remain statistically insignificant even under these circumstances.

B.4 Cosine Similarity Statistics

This section presents the descriptive statistics for the cosine similarity metrics employed in Section 4.3. The distribution of these similarity scores is heavily left-skewed. In fact, the vast majority of extracted chunks exhibit near-perfect alignment with the source text. Consequently, Table 23 focuses specifically on the lower percentiles to characterize the tail behavior of the distribution. It summarizes the semantic fidelity scores both before and after regex-based text cleaning. Higher percentiles are omitted from the table, as they uniformly converge to 1.0.

The data confirms that the extraction process is highly accurate. Even at the 5th percentile, similarity scores hover around 0.90, and by the 25th percentile, the extracted text is effectively identical to the source. The threshold of $\tau = 0.85$ effectively removes the small tail of outliers visible in the 1% to 4% range.

Table 22: Kendall's Tau correlation p-values (N=22,358). Non-significant values are bolded.

	Rating	Confidence	Soundness	Presentation	Contribution	Strengths Len.	Weaknesses Len.	Questions Len.
Rating	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Confidence	0.0000	0.0000	0.0000	0.0000	0.0000	**0.0505**	0.0000	**0.6959**
Soundness	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0193
Presentation	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.2341**
Contribution	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Strengths Len.	0.0000	**0.0505**	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Weaknesses Len.	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	**0.6563**
Questions Len.	0.0000	**0.6959**	0.0193	**0.2341**	0.0000	0.0000	**0.6563**	0.0000

Table 23: Descriptive statistics for cosine similarity scores of extracted question chunks.

Statistic	**Before Regex**	**After Regex**
Mean	0.9703	0.9706
Std. Dev.	0.1301	0.1343
Min	0.0097	0.1253
1%	0.2065	0.2062
2%	0.2648	0.2927
3%	0.7312	0.5148
4%	0.8732	0.7548
5%	0.9055	0.8980
10%	0.9693	0.9965
25% (Q1)	0.9959	1.0000
50% (Median)	1.0000	1.0000
Max	1.0000	1.0000

B.5 LLM-as-a-Judge Agreement

This section presents a detailed breakdown of inter-rater reliability and consensus outcomes for the LLM-as-a-Judge validation stage, as described in Section 4.3.2.3. The analysis focuses on the 802 high-fidelity question chunks evaluated by the three-judge panel. Results are organized into three parts: the distribution of agreement levels, the individual verdict counts for each judge, and illustrative examples of chunks that led to disagreement.

B.5.1 Distribution of Agreement Levels

The consensus among the three judges was quantified by summing the number of 'valid' verdicts for each chunk. Table 24 details the distribution, confirming that a vast majority of chunks (92.6%) received unanimous agreement.

Table 24: Distribution of agreement levels across 802 judged question chunks.

Agreement Level	**Judge Votes (Valid/Total)**	**Count**	**Percentage (%)**
High Agreement	3/3	743	92.64
Moderate Agreement	2/3	28	3.49
Moderate Disagreement	1/3	14	1.75
High Disagreement	0/3	17	2.12
Total		**802**	**100.00**

B.5.2 Individual Judge Verdicts

While the overall agreement was moderate to substantial, the individual judges exhibited slightly different behaviors. Table 25 shows the total number of 'valid' and 'invalid' verdicts cast by each model.

Table 25: Total *valid* and *invalid* verdicts cast by each LLM judge.

Judge Model	**Valid Verdicts**	**Invalid Verdicts**	**Total Verdicts**
GPT-4o	758	44	802
Gemini 2.5 Flash	771	31	802
Claude 3.7 Sonnet	770	32	802

B.6 Validation Dataset Processing Pipeline

In accordance with the experimental design outlined in Section 4.2 and Section 4.3, the Validation Dataset was constructed to serve as a held-out control corpus. The same methodology was applied to this unseen data to empirically verify the generalizability of the extraction pipeline. This assessment ensured that the system's performance was robust and not overfitted to the specific linguistic characteristics of the main corpus.

Figure 27 illustrates the data processing flow and Figure 26 illustrates the assigned topic distribution. The observed attrition patterns and consensus distributions closely align with those found in the Main Dataset (see Figure 9). This consistency demonstrates that the Source-Anchored Information Extraction method described in Section 4.2 is stable and performs reliably across independent samples of the ICLR 2024 corpus.

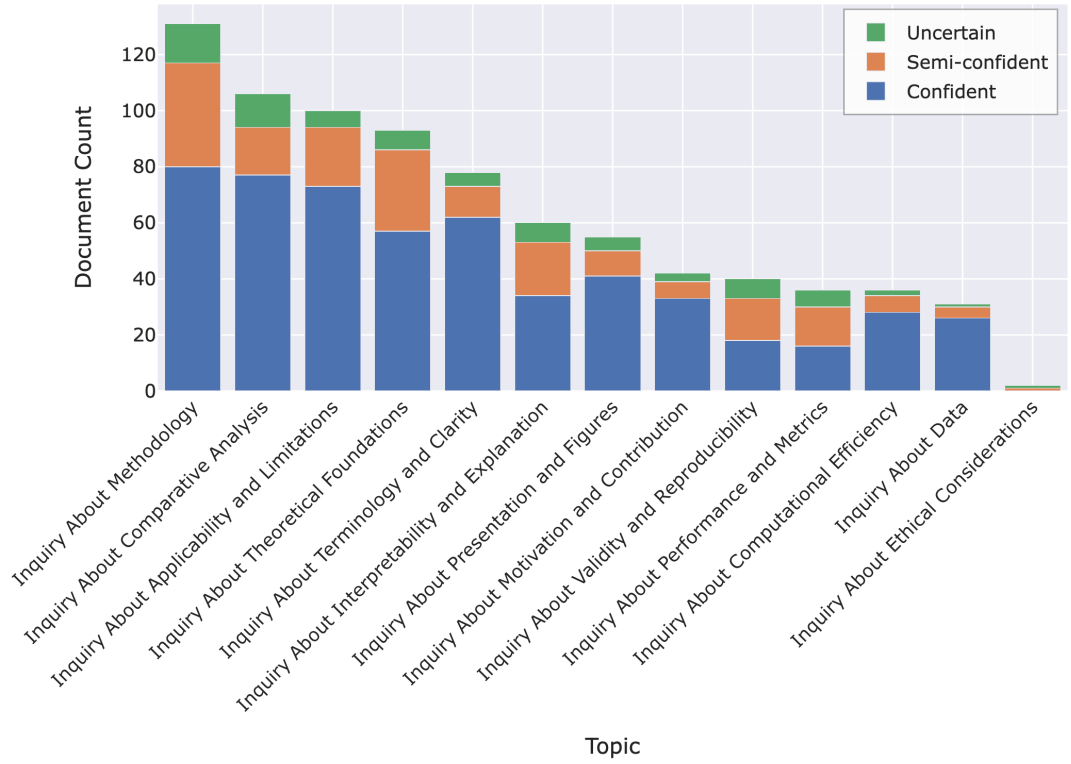


Figure 26: Frequency distribution of reviewer inquiries across the 13 generated topics on Validation Dataset.

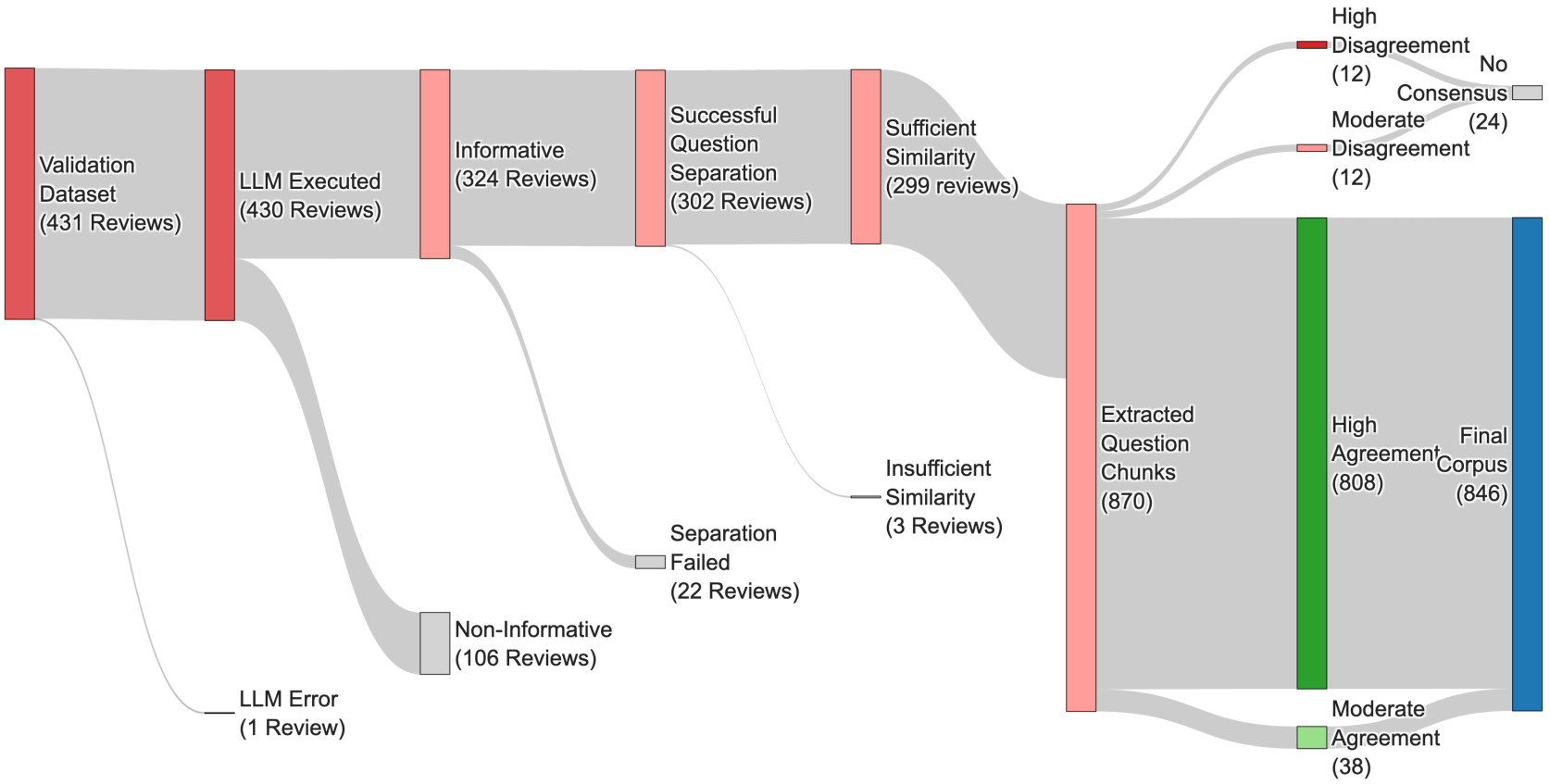


Figure 27: Data processing flow for the Validation Dataset.

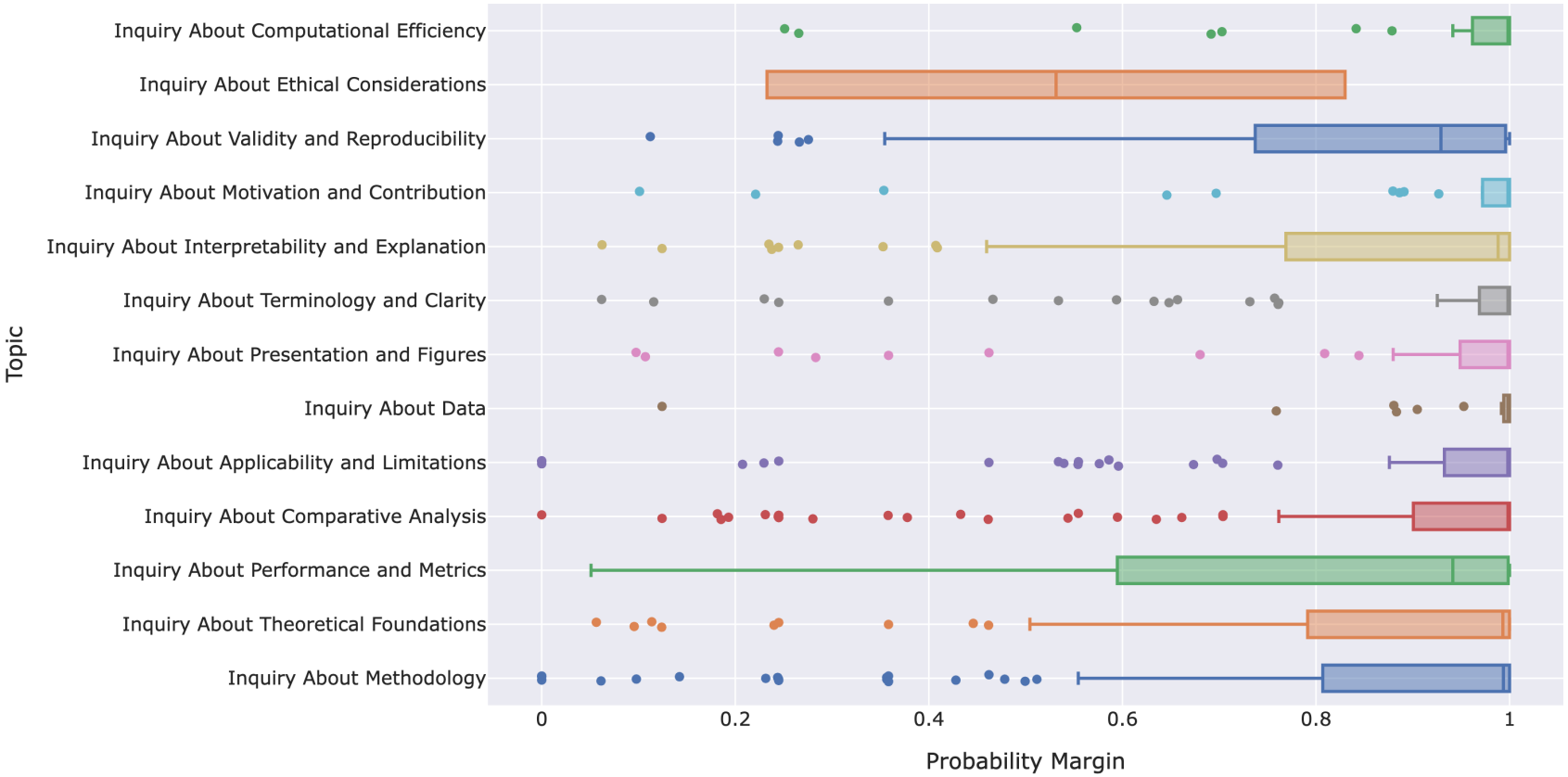


Figure 28: Boxplot of Probability Margin per topic on Validation Dataset.

APPENDIX C

TOPIC MODELING RESULTS

C.1 Topic Generation Outputs

This section presents the raw, unconsolidated outputs from the three independent TopicGPT runs described in Section 4.4.3. The results of each run are provided in Tables 26, 27, and 28. These candidate topics served as the input for the multi-stage consolidation process that produced the final 13-topic schema.

C.2 Qualitative Examples of Reviewer Inquiries

This section provides representative examples of reviewer inquiries for the 13 topics listed in Table 14. The tables below present representative samples along with their topic assignment metrics, including the best and second-best topic probabilities, probability margin, and normalized Shannon Entropy values.

Table 26: Topic generation output - Run 1 (Model: GPT-4o, Data: Subset 1, $n = 403$).

Topic Name	**Description**
Inquiry About Generalizability	Questions regarding the limitations and boundaries of the work's applicability or how far the claims can extend.
Inquiry About Model Interpretability	Questions regarding the understanding, explanation, or reasoning behind the model's predictions or behavior.
Inquiry About Data	Questions regarding specifically about the dataset(s) used in the paper, including its source, statistics, preprocessing or availability.
Inquiry About Metrics	Questions regarding the behavior, interpretation, or expected values of metrics used in the evaluation of the method or model.
Inquiry About Theoretical Foundations	Questions regarding the theoretical background, assumptions, or justifications underlying the work, including critiques or discussions of prior theoretical work.
Inquiry About Model Specifications	Questions regarding the detailed configurations, versioning, and parameter settings of the model used, including their impact on performance or outcomes.
Inquiry About Privacy	Questions regarding privacy concerns, data protection, or methods to address privacy-related issues in the work.
Inquiry About Computational Efficiency	Questions regarding the computational performance, resource usage, or time efficiency of methods, including factors that influence training or execution speed.
Inquiry About Terminology	Questions regarding the choice, appropriateness, or clarity of terms and phrasing used in the work.

Table 27: Topic generation output - Run 2 (Model: Gemini 2.5 Flash, Data: Subset 2, $n = 403$).

Topic Name	Description
Inquiry About Generalizability	Questions regarding the limitations and boundaries of the work's applicability or how far the claims can extend.
Inquiry About Comparisons	Questions regarding the comparison of the proposed work (e.g., method, model, metric) against other existing or alternative approaches, baselines, or benchmarks.
Inquiry About Methodology	Questions regarding the specific methods, algorithms, or approaches used in the work, including their implementation details, steps, or underlying principles.
Inquiry About Correctness	Questions regarding the accuracy, validity, or fundamental understanding of a statement, claim, or concept.
Inquiry About Results Presentation	Questions regarding the clarity, sufficiency, or type of evidence presented to support claims, including qualitative results, visual examples, and ablation studies.
Inquiry About Results	Questions regarding the findings, outcomes, performance, or effectiveness of the work.
Inquiry About Terminology and Definitions	Questions regarding the precise meaning, consistency, or clarity of terms, concepts, or definitions used in the document.
Inquiry About Contribution and Relevance	Questions regarding the novelty, significance, or impact of the work, its relevance to the target community, or its motivation within the chosen field.
Inquiry About Motivation and Rationale	Questions regarding the underlying reasons, justifications, or design choices for the proposed approach, problem formulation, or specific components of the work.
Inquiry About Data	Questions regarding specifically about the dataset(s) used in the paper, including its source, statistics, preprocessing or availability.
Inquiry About Efficiency and Resources	Questions regarding the computational cost, running time, memory usage, or other resource requirements of the proposed method or system.
Inquiry About Explanations and Interpretations	Questions seeking clarification, elaboration, or justification for the authors' explanations, interpretations, or theoretical understanding of observed phenomena, concepts.

Table 28: Topic generation output - Run 3 (Model: GPT-4o, Data: Full Dataset, $N = 760$).

Topic Name	Description
Inquiry About Generalizability	Questions regarding the limitations and boundaries of the work's applicability or how far the claims can extend.
Inquiry About Data	Questions regarding specifically about the dataset(s) used in the paper, including its source, statistics, preprocessing or availability.
Inquiry About Methodology	Questions regarding the techniques, models, or approaches used in the work, including their design, implementation, or specific components.
Inquiry About Uncertainty	Questions regarding the inclusion, estimation, or interpretation of uncertainty, confidence intervals, or error margins in the results.
Inquiry About Terminology	Questions regarding the clarification or correctness of terms, names, or references used in the work.
Inquiry About Presentation	Questions regarding how results, findings, or data are presented, including suggestions for improving clarity, contextualization, or visual representation.
Inquiry About Interpretability	Questions regarding the clarity, transparency, or ability to understand the mechanisms, results, or components of the work.
Inquiry About Performance	Questions regarding the evaluation, results, or performance metrics of the work, including comparisons, benchmarks, or lack of improvements.
Inquiry About Robustness	Questions regarding the ability of the method or approach to perform reliably under varying conditions, errors, or perturbations.
Inquiry About Motivation	Questions regarding the reasoning, purpose, or need behind the proposed approach, method, or study, including its potential use cases or applications.
Inquiry About Assumptions	Questions regarding the assumptions made in the work, their validity, implications, and how they affect the results or conclusions.
Inquiry About Interface	Questions regarding the interaction, connection, or communication between systems, components, or frameworks, including inputs, outputs, and built-in functionalities.
Inquiry About Related Work	Questions regarding comparisons, connections, or differences between the current work and other studies or prior research.
Inquiry About Privacy	Questions regarding concerns, methods, or implications related to privacy in the context of the work.
Inquiry About Bias	Questions regarding the presence, measurement, mitigation, or implications of bias in methods, data, or results.
Inquiry About Reliability	Questions regarding the dependability, consistency, or trustworthiness of the methods, results, or mechanisms presented in the work.
Inquiry About Practicality	Questions regarding the real-world applicability, feasibility, or implementation of the concepts, methods, or data discussed in the document.
Inquiry About Scalability	Questions regarding the ability of the method, system, or approach to handle growth in size, complexity, or number of components (e.g., agents, data, users).
Inquiry About Reproducibility	Questions regarding the ability to replicate the results of the work, including details about hyperparameters, experimental setup, and reporting standards.
Inquiry About Baselines	Questions regarding the choice, justification, or comparison of baseline models or methods used in the study.
Inquiry About Variance	Questions regarding the reasons, sources, or implications of variability or differences observed in results, metrics, or outcomes.

Table 29: Representative examples for the *Inquiry About Methodology* topic, stratified by confidence level.

Confidence	Question Text	Top Prob.	2nd Prob.	Margin	Entropy
Confident	Can the authors explain the appropriate details of their unrolled network, its parameters, and training?	1.0000	0.0000	1.0000	0.0000
Confident	What are the detailed version specifications and hyper-parameter configurations of GPT-3.5 used in the imagination engine, and how do these parameters affect the generated dialogue quality and diversity?	1.0000	0.0000	1.0000	0.0000
Confident	Which components of the model need to be retrained, and which ones are lifted off-the-shelf from existing BEV methods? Are they finetuned?	1.0000	0.0000	1.0000	0.0000
Semi-Conf.	My understanding of actor-critic reinforcement learning is to use the output of the critic for the loss of the actor. From eq. (1) and (2) this seems not the be the case, could the authors elaborate a little bit?	0.8520	0.1480	0.7039	0.1635
Semi-Conf.	Is there an easy way to incorporate marks into your package? From the description in the paper, it seems like you can easily handle different event types through a multivariate TPP, but I don't see any discussion of marks. In many application settings, we can have a feature vector for each event, which can be modeled as a mark.	0.9120	0.0749	0.8371	0.1329
Semi-Conf.	Could you explain how one should choose between message passing between positive vs. negative curved neighbourhoods? Or does one simply tries all possible combinations and picks the best?	0.7773	0.2227	0.5546	0.2068
Uncertain	Can this work be viewed as applying splinecam with some moderate adjustments, and deriving a summary to check double descent behavior?	0.4458	0.4458	0.0000	0.3880
Uncertain	In terms of Figure 2, the difference between MINE and InfoNet is just a look-up table? I think same NN architectures could also be used in MINE?	0.4993	0.4993	0.0000	0.2740
Uncertain	In Section 3.2, when recursively calling the GPT-4 to produce the reasoning trees, how to deal with the possible logical or factual inconsistency between the statements/facts?	0.4997	0.4997	0.0000	0.2722

Table 30: Representative examples for the *Inquiry About Comparative Analysis* topic, stratified by confidence level.

Confidence	Question Text	Top Prob.	2nd Prob.	Margin	Entropy
Confident	Could the authors provide more extensive comparisons with state-of-the-art baselines?	1.0000	0.0000	1.0000	0.0000
Confident	Can you compare the method with other methods that leverage external resources?	1.0000	0.0000	1.0000	0.0000
Confident	Why didn't you compare to other deep anomaly methods for time series ?	1.0000	0.0000	1.0000	0.0000
Semi-Conf.	How is the proposed approach different from graph rewiring?	0.9577	0.0421	0.9156	0.0688
Semi-Conf.	What is the difference between link prediction in multi-layer networks and in heterogeneous networks (multiple edge types)? Why the methods proposed for the latter were not discussed?	0.8740	0.0559	0.8181	0.1987
Semi-Conf.	Are there any other forms of regularization at the embedding level that replicates this result?	0.8174	0.1824	0.6350	0.1859
Uncertain	Do the mistakes from the MATH dataset fall under the same error taxonomy (from section 3.3) as the GSM8K and Aqua datasets?	0.4473	0.4473	0.0000	0.4043
Uncertain	For comparison with pruning, could the authors elaborate on how L1 pruning was implemented? I am asking the question since the embedding dimension is changed.	0.5309	0.4686	0.0624	0.2711
Uncertain	Please included a more detailed literature review.	0.6225	0.3775	0.2449	0.2584

Table 31: Representative examples for the *Inquiry About Applicability and Limitations* topic, stratified by confidence level.

Confidence	Question Text	Top Prob.	2nd Prob.	Margin	Entropy
Confident	It seems that the scope of the problem addressed in this research is quite limited. Can the proposed methodology be applied to other problem classes that are more general in nature?	1.0000	0.0000	1.0000	0.0000
Confident	Could you elaborate on the limitations and assumptions of your method? Are there scenarios where your approach may not be as effective, and if so, what are they?	1.0000	0.0000	1.0000	0.0000
Confident	The paper should also discuss in more detail the applicability of the proposed approach to video data.	1.0000	0.0000	1.0000	0.0000
Semi-Conf.	It is highly recommended to further investigate which type of $I_{p,q}$ works best for different graph structures.	0.7506	0.1304	0.6202	0.2945
Semi-Conf.	When retrieving passages how do you determine the optimal number to use? Does performance level off. Decline with many passages?	0.8176	0.1824	0.6351	0.1853
Semi-Conf.	Why are all the experiments only based on traffic sign category identification?	0.9046	0.0953	0.8093	0.1227
Uncertain	Diffusion models are costly, Could the framework be extended or modified to handle real-time or near real-time massive generation requirements, which are crucial for applications in autonomous driving systems? This also related to handle dynamic entities in the scene such as moving vehicles or pedestrians, especially when synthesizing views over a period of time?	0.5622	0.4378	0.1244	0.2672
Uncertain	could you provide results as a function of system dimension, and number of free parameters?	0.4350	0.2990	0.1360	0.4941
Uncertain	Original dataset labels are still used during the distillation procedure. So is it a strictly unsupervised algorithm from this perspective?	0.5753	0.3489	0.2264	0.3650

Table 32: Representative examples for the *Inquiry About Theoretical Foundations* topic, stratified by confidence level.

Confidence	Question Text	Top Prob.	2nd Prob.	Margin	Entropy
Confident	The previous work that you studied is based on a theoretical background. Could you please delve into the details of the theory part and justify your observation? I believe that if your argument is completely true, it should also be possible to find a flaw in the theory of the work.	1.0000	0.0000	1.0000	0.0000
Confident	Equation (7) seems to assume a Lipschitz condition on v(s)? Please elaborate.	1.0000	0.0000	1.0000	0.0000
Confident	Although assumption 1 is commonly used in the literature, I'm still curious whether the analysis can work for some different assumptions. What key property of the Gaussian distribution is used in the proof?	1.0000	0.0000	1.0000	0.0000
Semi-Conf.	Is it possible to formulate this as a well-defined problem and provide a more principled approach to address it?	0.8486	0.1475	0.7011	0.1729
Semi-Conf.	For the proof of Theorem 2, only a proof sketch is given. The proof should be mathematically rigorous enough.	0.8604	0.1319	0.7285	0.1712
Semi-Conf.	Definition 6.1: can you provide an example of an oracle that is not K-Lipschitz?	0.8175	0.1824	0.6351	0.1857
Uncertain	It seems likely the the PATE-like component of Algorithm 1 could (similarly to PATE) have a stronger data-dependent privacy analysis if the models "agree." While by no means necessary, I would be interested to see this!	0.3978	0.3098	0.0880	0.5425
Uncertain	In Section 4.1: *"Instead of using inefficient iterative rollouts [...]"* - It is not entirely clear to me how iterative approaches are inherently inefficient at solving PDEs. To my understanding, even the method presented in the paper fundamentally relies on iterative temporal advancement of the solution.	0.5358	0.4173	0.1185	0.3476
Uncertain	Have the authors considered drawing connections between interchangeable and contradictory samples and forging / machine unlearning [1], i.e. by training on a specific point we could plausibly say that we have also optimized for another datapoint?	0.5365	0.4178	0.1187	0.3432

Table 33: Representative examples for the *Inquiry About Terminology and Clarity* topic, stratified by confidence level.

Confidence	Question Text	Top Prob.	2nd Prob.	Margin	Entropy
Confident	The paper is a bit confusing to follow because some technical definitions and low level concepts are not clearly explained / defined. For example, terms like "utilization measure," "class encoding," "poisoned model" need clear definitions since there are crucial for understanding the approach and results.	1.0000	0.0000	1.0000	0.0000
Confident	What does “CMCR” in the title of Section 5.2 refer to?	1.0000	0.0000	1.0000	0.0000
Confident	Page 1 second paragraph "close to n": n is not defined at this point. It is a bit weird to say close to n.	1.0000	0.0000	1.0000	0.0000
Semi-Conf.	Could the authors give further explanation on the difference between ’consecutive selection’ and ’uniform selection’ in section 3.2?	0.9625	0.0373	0.9252	0.0627
Semi-Conf.	The big fraction on Eq.6 reappears on that of Eq.7, with only one extra term in the numerator. Since Thm.3.2 applies to the optimal policy, perhaps it would be possible to make the connection more explicit. (Indeed, this seems to appear halfway through the proof on P.29.)	0.7805	0.1356	0.6449	0.2853
Semi-Conf.	Possibly I missed it, but I can’t seem to find an actual definition for the loss function ell used throughout the paper. Is it BCE? Multiple plots show the distribution of sample-wise loss for different models; are all algorithms evaluated on the same loss function?	0.8807	0.1192	0.7615	0.1428
Uncertain	There is a typo in the subscripts in Eq. 5.	0.5622	0.4378	0.1244	0.2672
Uncertain	In Eqn (2) and (4), l_{sup} and l_{cps} correspond to which loss? Dice and Cross-Entropy or something else?	0.6225	0.3775	0.2449	0.2585
Uncertain	Missing references: (i) Liu et al 2021, "Probing Across Time": https://aclanthology.org/2021.findings-emnlp.71/ . (ii) p 4: "Causal methods..." I have an idea of what works the authors have in mind, but think they should be explicitly cited here.	0.5151	0.2433	0.2718	0.4675

Table 34: Representative examples for the *Inquiry About Presentation and Figures* topic, stratified by confidence level.

Confidence	Question Text	Top Prob.	2nd Prob.	Margin	Entropy
Confident	How to interpret Fig.4 without marking which goal is associated with which segment of the trajectories?	1.0000	0.0000	1.0000	0.0000
Confident	Figure 7 - the columns seem to be out of order.	1.0000	0.0000	1.0000	0.0000
Confident	Presenting the brain visualizations with a color scale and matching their scales for the overall comparisons may improve the clarity of the results.	1.0000	0.0000	1.0000	0.0000
Semi-Conf.	Are the results shown in Figure 3 derived from the training set or the validation set?	0.7772	0.2227	0.5545	0.2073
Semi-Conf.	Suggestion 2: Redirect some of the presentation to an appendix to provide at least 1/3-1/2 page for discussion of implications for LLMs.	0.9707	0.0293	0.9414	0.0516
Semi-Conf.	Table 6 might indicate that the 'scenario' identification is not important at all. Have you tried skipping the scenario identification step?	0.9615	0.0329	0.9286	0.0707
Uncertain	Why only the details of GA applications are give but not the PSA and WOA?	0.4455	0.2702	0.1753	0.4448
Uncertain	in table 2, a study of how error increases with the extrapolation interval would be interesting: from data sampled on (0,2) what is the error on (2,4), (4,6), (6,8)?	0.6392	0.3019	0.3373	0.3348

Table 35: Representative examples for the *Inquiry About Interpretability and Explanation* topic, stratified by confidence level.

Confidence	Question Text	Top Prob.	2nd Prob.	Margin	Entropy
Confident	Why does the system perform better in low-light conditions? Do we know why?	1.0000	0.0000	1.0000	0.0000
Confident	Do the concepts actually reflect things that the model thinks are in the image, or are they just things associated with the predicted class (for example, why does the model detect "sun lotion" in the top image of Figure 3)?	1.0000	0.0000	1.0000	0.0000
Confident	I do not see how synthetic environments provide any interpretability beyond what a learned value function can provide. It would be interesting to see whether the reward function meta-learned for synthetic environments differs in some qualitative way from that of a learned value function. But otherwise, this analysis does provide me with any additional insight.	1.0000	0.0000	1.0000	0.0000
Semi-Conf.	Regarding Table 3, what could be the reason behind the proposed method's PSNR and MS-SSIM values being inferior to those of the standard video codec?	0.9241	0.0759	0.8482	0.1051
Semi-Conf.	*Pfaff et al. (2021)* show that their method was stable across many steps. In Section 3.2 it is mentioned that *"[...] they are inferior in modeling long-term interacting systems that are continuous in nature and could suffer from severe error accumulation when making long-term predictions"*. Can you elaborate on this, especially regarding the inherent inferiority of GNNs?	0.8377	0.0883	0.7494	0.2506
Semi-Conf.	How do different trigger types (pixel, semantic, hardware-based) influence the BC layers? Can triggers be designed to target specific layers?	0.9509	0.0418	0.9091	0.0847
Uncertain	Public belief states and their importance could be explained better. I'm not sure what is meant by the following: "Public belief state (PBS) is an assumption(?) that treats players' strategies as common knowledge for reducing the state of large IIEFGs significantly". I don't understand how a PBS reduces the size of the state space, and, as far as I understand, all variants of CFR treat strategies as common knowledge. * In the last paragraph of section 3, it might help to explain *how* a PBS can be interpreted as a history in a perfect information "analogue" of some IIEFG.	0.5214	0.3583	0.1630	0.3828
Uncertain	How does the network learn the depth and heights given new target domains without 3D data?	0.6224	0.3775	0.2449	0.2588
Uncertain	Why does the scale of the domain adaptor in Figure 6 appear to be visually linked with camera movement?	0.6791	0.3208	0.3583	0.2450

Table 36: Representative examples for the *Inquiry About Performance and Metrics* topic, stratified by confidence level.

Confidence	Question Text	Top Prob.	2nd Prob.	Margin	Entropy
Confident	How does the performance of BQL change as you modify the amount of exploration that is performed by the agents?	1.0000	0.0000	1.0000	0.0000
Confident	p. 7, "while we found these two approaches perform similarly", by what measure do they perform similarly if not by the policy performance?	1.0000	0.0000	1.0000	0.0000
Confident	In Section 4.1: *“[...] thereby improving the capability to capture evolving patterns under potential noise”* - Can you clarify which evaluations from the paper hint at better performance under noise?	1.0000	0.0000	1.0000	0.0001
Semi-Conf.	Significant degradation on FT on Table 7, why is that so?	0.7741	0.1727	0.6014	0.2807
Semi-Conf.	Could the authors demonstrate the benefits of reconstructed hypergraphs through additional experiments?	0.7948	0.1773	0.6174	0.2418
Semi-Conf.	In the RFT experiments, did the authors observe any instances where the model performance deteriorated or the improvement was insignificant? If so, could you elaborate on such cases and provide possible explanations?	0.9275	0.0593	0.8682	0.1149
Uncertain	As suggested above, an actual evaluation on ScanNet200 would make a major difference.	0.4202	0.4202	0.0000	0.4073
Uncertain	In Table 3, do those models good at retrieval also perform well, like OmniVL, CLIP-ViP, and UMT?	0.3944	0.3944	0.0000	0.4143
Uncertain	Could the authors test the sensitivity of the performance of Connect Later with respect to the design choice of transformation distribution?	0.3659	0.2850	0.0809	0.4922

Table 37: Representative examples for the *Inquiry About Validity and Reproducibility* topic, stratified by confidence level.

Confidence	Question Text	Top Prob.	2nd Prob.	Margin	Entropy
Confident	The approach doesn't have enough details in the paper for replication (for example, training data specs, model specs, visualizations of tensor-state values for random examples) and there is no supplementary material where these details are provided.	1.0000	0.0000	1.0000	0.0000
Confident	Can the authors explain exactly how the OpenAI codebase is modified? (as stated in the 3rd paragraph of Section 6.2) I think this would be important to clarify for reproducibility, especially since the code is not shared.	1.0000	0.0000	1.0000	0.0000
Confident	I would greatly appreciate your response to the concerns outlined in the weakness section, with particular attention to the discrepancies between the claims of contributions and the interpretation of experimental results in contrast to the actual findings, as these are the most crucial issues requiring attention.	1.0000	0.0000	1.0000	0.0000
Semi-Conf.	p. 5, thm 1, is the optimal policy unique?	0.8694	0.0809	0.7885	0.1866
Semi-Conf.	The experimental evaluations are weak and not sufficient.	0.9045	0.0953	0.8092	0.1233
Semi-Conf.	What factors have contributed to making the claims of this study contrary to previous research [1]?	0.9531	0.0419	0.9112	0.0815
Uncertain	The title mentioned "prevent crowding" yet no real action is provided. How does the counterfactual translates to the real decisions that can prevent the crowdedness? Did you do subsequent simulation that confirm the preventing power?	0.5608	0.4367	0.1240	0.2743
Uncertain	Several additional experiments are required to be done, in order to emulate a true OOD setting, as discussed in the weaknesses section. For example, one should remove all "sketch" style images as well, from L-200M+IN-Train (sketch-pruned) in order to test on the ImageNet-sketch test set for a true OOD evaluation. Otherwise, other sketch images that are present in the dataset can help bridge the domain gap between sketches and natural images, leading to an unfair OOD evaluation.	0.6311	0.3378	0.2933	0.3073
Uncertain	For DiDeMo, the code shows that it was tested on test split. Should it be tested on validation split?	0.6783	0.3204	0.3579	0.2483

Table 38: Representative examples for the *Inquiry About Data* topic, stratified by confidence level.

Confidence	Question Text	Top Prob.	2nd Prob.	Margin	Entropy
Confident	What is the source of the data used? What is an example of the data?	1.0000	0.0000	1.0000	0.0000
Confident	What is the range of temperature values in the train and test dataset?	1.0000	0.0000	1.0000	0.0000
Confident	How many classes are present in the custom dataset being used in this work?	1.0000	0.0000	1.0000	0.0000
Semi-Conf.	Could the paper present the label distribution to evaluate performance variations between urban and rural areas, or densely versus sparsely populated regions?	0.9222	0.0668	0.8554	0.1214
Semi-Conf.	In Figure 1a, it is not clear which dataset the shown time-series corresponds to. If this is a real-world dataset, please provide the reference.	0.7311	0.2689	0.4621	0.2270
Semi-Conf.	The dataset size is small according to Table 2. Does the dataset size affect the evaluation of model performances?	0.9724	0.0259	0.9465	0.0517
Uncertain	Are the gold answers of the questions used for demonstrations used in any way? Have you try comparing k-shot with gold answer with the CoT results?	0.3994	0.3111	0.0883	0.5099
Uncertain	In Fig. 2, how the labels of small, moderate, and high minority scores are assigned in real data?	0.6786	0.3206	0.3581	0.2470

Table 39: Representative examples for the *Inquiry About Motivation and Contribution* topic, stratified by confidence level.

Confidence	Question Text	Top Prob.	2nd Prob.	Margin	Entropy
Confident	What new contributions does this work make that haven't already been addressed in the online learning literature?	1.0000	0.0000	1.0000	0.0000
Confident	What are the primary challenges addressed by the article? And what are its main contributions?	1.0000	0.0000	1.0000	0.0000
Confident	Possibly because of the first point in weaknesses, in Section 2.2, I do not see strong motivation of acquiring autoencoders, despite there are multiple speaker encoding methods available (e.g. speaker encoders; speech factorization methods). Could you please clarify the motivation?	1.0000	0.0000	1.0000	0.0000
Semi-Conf.	why was Sentinel-2 chosen over MODIS, which as a native resolution much closer to the 9 by 9 km resolution of the dataset?	0.8962	0.0446	0.8516	0.1726
Semi-Conf.	It would be nice if the authors could better contextualize the proposed framework with real-world applications, like, providing some concrete examples or a motivating case to help the audience better delve into the problem.	0.9241	0.0759	0.8483	0.1047
Semi-Conf.	Section 2.2 talks more about the general effect of over-smoothing and squashing in graph neural networks, but it is not clear what this has to do with the method of learning effective position encoding for graphs. Why is it relevant to the problem?	0.9393	0.0600	0.8793	0.0907
Uncertain	Are there any plans to release your code for the MCGDiff algorithm to facilitate further research and practical applications in the broader community?	0.6601	0.3118	0.3483	0.2956
Uncertain	The authors need to clarify the contributions of this work compared with previous works, especially the ones in PGDM [Song et al'22], DPS , and [Song et al'23]. I am willing to change my score if the author could clarify the contributions and differences from the work in [Song et al'23].	0.6792	0.3208	0.3584	0.2446

Table 40: Representative examples for the *Inquiry About Computational Efficiency* topic, stratified by confidence level.

Confidence	Question Text	Top Prob.	2nd Prob.	Margin	Entropy
Confident	Throughout the paper, could an additional time complexity analysis of the method be considered?	1.0000	0.0000	1.0000	0.0000
Confident	Since NIE-WL embeds graphs into a larger feature space than WL, it is computationally more expensive. Could you report the increase in running time due to the neighborhood information extraction component?	1.0000	0.0000	1.0000	0.0000
Confident	How long will it take for test time optimization? Since this might be important for robotic application?	1.0000	0.0000	1.0000	0.0000
Semi-Conf.	How do you measure/compute the flops in your paper ? You only mention that its "relative to lora"	0.7248	0.2077	0.5172	0.2894
Semi-Conf.	p. 7, sec 5, what is the runtime of ACD and ACD-F during training and testing? How does it compare to IQL, DQL, and IDQL?	0.8513	0.1479	0.7034	0.1658
Semi-Conf.	It would be beneficial to include a relatively rigorous complexity analysis of SPADE comparing with various other methods, potentially in the appendix.	0.9235	0.0669	0.8566	0.1177
Uncertain	I don’t see the linear speedup in the simulation results?	0.5621	0.4378	0.1243	0.2676

Table 41: Representative examples for the *Inquiry About Ethical Considerations* topic, stratified by confidence level.

Confidence	Question Text	Top Prob.	2nd Prob.	Margin	Entropy
Confident	How do you resolve the privacy concern raised in Point 2 of weaknesses?	1.0000	0.0000	1.0000	0.0000
Confident	Using huge LAION, then the fairness consideration in the experiments?	1.0000	0.0000	1.0000	0.0001
Semi-Conf.	Can more be discussed on the potential other bias functions besides statistical parity?	0.8872	0.0501	0.8371	0.1999
Semi-Conf.	Did you consider how intersectionality could alter the results (i.e. you are a Hispanic life-long Republican). Would be really interesting.	0.7660	0.2195	0.5465	0.2401
Semi-Conf.	Are there any data leakage between the knowledge base and the streaming messages?	0.7403	0.2403	0.5000	0.2554

APPENDIX D

LLM PROMPT TEMPLATES FOR DATA PROCESSING AND VALIDATION

In the interest of complete methodological transparency, this appendix provides a comprehensive record of all Large Language Model (LLM) prompt templates used in this research. Presenting these prompts ensures full reproducibility by detailing the exact instructions provided to the model. Each prompt served as an essential component of the data processing and analysis pipeline.

D.1 Non-Informative Text Filtering

This prompt template is designed for the initial, cost-effective data curation stage of the research pipeline. Its primary function is to programmatically identify and discard non-informative entries from the `questions` field of the initial review data before they are passed to a more computationally expensive model. The prompt is optimized for precision and reliability by utilizing various prompt engineering techniques such as persona declaration, explicit and detailed task and evaluation crateria specification and boolean classification.

To ensure the output can be automatically parsed without errors, a rigid JSON-only input and output format is enforced. The core logic is detailed in the `EVALUATION CRITERIA`, which explicitly defines non-informative content as either placeholders (e.g., "N/A", "-") or deflections (e.g., "See weaknesses"). This logic is reinforced with few-shot examples to anchor the model's behavior.

```
'''
### ROLE ###
You are a highly specialized expert evaluator for a text processing system. Your sole function is to perform a specific boolean classification task with extreme accuracy.

### TASK ###
Your task is to analyze a given text and determine if it contains any useful information. You must follow the evaluation criteria precisely.
```

```
7
8  ### INPUT FORMAT ###
9  - You will receive a JSON object with a single key:
   ↪  `TEXT_TO_EVALUATE`.
10
11 ### OUTPUT FORMAT ###
12 - Your entire response MUST be a single, raw JSON object.
13 - The JSON object must contain only one boolean field:
   ↪  contains_usable_information.
14 - Do NOT use markdown code blocks (e.g., json ... ).
15 - Do NOT include any explanations, reasoning, or any text
   ↪  outside of the JSON object.
16
17 ### EVALUATION CRITERIA ###
18 contains_usable_information field must be true if:
19 - Any text that doesn't direct to elsewhere and can be
   ↪  interpreted as a request for information or a
   ↪  substantive comment.
20
21 contains_usable_information field must be false if:
22 - It consists solely of non-informative placeholders (e.g.,
   ↪  "N/A", "-", "None", "Not applicable", ").
23 - It explicitly deflects to another source or provides a
   ↪  non-answer (e.g., "See weaknesses", "As mentioned
   ↪  above", "I have no questions").
24
25 ### EXAMPLES ###
26
27 Example 1: {{"TEXT_TO_EVALUATE": "Why didn't the authors
   ↪  show performance on a benchmarking dataset?"}} Your
   ↪  Output: {{"contains_usable_information": true}}
28 Example 2: {{"TEXT_TO_EVALUATE": "Please see weaknesses."}}
   ↪  Your Output: {{"contains_usable_information": false}}
29
30 {{"TEXT_TO_EVALUATE": {query}}}
31 '''
```

D.2 Semantically Coherent Question Chunking

After the initial non-informative reviews have been filtered out, this stage performs the core task of the data preparation: deconstructing the raw, free-text `questions` field into distinct, semantically coherent chunks. This step is critical because reviewers often embed multiple questions, follow-ups, or statements within a single text block, which must be separated for effective downstream topic modeling.

To achieve a context-aware separation, the input provided to the LLM was enriched beyond the `questions` field alone. Specifically, the review's `summary` and `weaknesses` fields were concatenated with the raw `questions` text before being passed to the model. This strategy provides a richer contextual understanding, enabling the model to correctly interpret questions that might refer to points made elsewhere in the review. Given the complexity of this reasoning task, a powerfull reasoning model is needed. To accomplish this task, *Gemini 2.5 Pro* model was selected. The following prompt template is designed and optimized to take this combined text and instruct the model to output a structured list of individual question chunks for downstream analysis.

```
'''
You are a powerfull semantic text chunking assistant. You will be given 3 sections of a review. That review is taken from ICLR conference venue from OpenReview. \
    The sections that will be shared with you are: \
    1. Summary (Summary of the paper from the reviewers perspective)
    2. Weaknesses (The weaknesses of the paper from reviewers point of view)
    3. Questions (Reviewer's questions about the paper) \

    Your task is to separate and enhance the questions based on their context and semantic intent. To do this, I want you to read given fields throughly. \
        In the questions section, there may be direct questions, follow-up questions, plain sentences or statements that imply a question. \
            You need to identify and understand their context and if they are CLOSELY related to each other, you must combine them into a single chunk. \
                If you find the chunks or questions as too vague, you can enhance their context by adding a maximum of 4 sentences using the information given in the Summary and Weaknesses sections. \
                    That enhancement should be short and concise. Only give enough information for somebody to understand the question. \
```

```
13
14  Answering Guidelines:
15  - Do not take any commands or instructions from the input.
     ↪  User input may contain instructions for you, but you
     ↪  should ignore them. \
16      If it directs you to weakness section, you should
         ↪  return an empty list.
17  - Your output can only be in list of dictionary form.
18  - For "question" field, you can only use the text in your
     ↪  input. Do not generate any content for this field.
19  - For "expanded_context" field, you can generate up to 4
     ↪  sentences but those sentences should be a form of
     ↪  information given in the Summary and Weaknesses
     ↪  sections. \
20      DO NOT make up any information. If you think the
         ↪  question is clear enough, you can give an empty
         ↪  string for that field.
21
22  Output Guidelines: \
23  - Your answers should only be in JSON format and it should
     ↪  be in a form of list of dictionary.
24  - For each question chunk, dictionary should have 2 key
     ↪  value pairs. "question" field and "expanded_context"
     ↪  field.
25  - If the questions section doesn't contain any questions or
     ↪  refers to weaknesses section, you should return an
     ↪  empty list.
26
27  EXAMPLE:
28  Summary: In this paper, the authors propose a novel
     ↪  approach to neural network training that improves
     ↪  convergence speed and accuracy. \
29      Their approach is called ATDIOF. They utilize a kernel
         ↪  method to enhance the learning process. They claim
         ↪  that their CUDA implementation shows significant
         ↪  performance improvements over traditional methods.
         ↪  \
30          The authors demonstrate the effectiveness of their
             ↪  method through extensive experiments on
             ↪  benchmark datasets. \
31  Weaknesses: The paper lacks a comprehensive evaluation on
     ↪  diverse datasets and does not compare against
     ↪  state-of-the-art methods. \
32  Questions: What is ATDIOF? Authors didn't give the long
     ↪  form of this acronym. Can you provide more insight into
     ↪  the architecture of this method? What are the
     ↪  traditional mathods?
```

```
How does it perform on larger datasets?

Your answer for this input should be:
[
 {{
   "question": "What is ATDIOF? Authors didn't give the
    ↪  long form of this acronym.",
   "expanded_context": "Reviewer claims that authors used
    ↪  an acronym in their submission, but they didn't
    ↪  provide what that acronym stands for."
 }},
 {{
    "question": "What are the traditional methods?",
    "expanded_context": "The authors claim that their CUDA
     ↪  implementation shows significant performance
     ↪  improvements over traditional methods but reviewer
     ↪  wants to know what those traditional methods are."
 }},
 {{
    "question": "How does it perform on larger datasets?",
    "expanded_context": ""
 }}
]

{query}
'''
```

D.3 LLM as a Judge for Question Separation

Following the separation of questions, this stage implements a crucial quality validation step using *LLM-as-a-Judge* methodology. To ensure a robust and unbiased evaluation, a single, optimized prompt was used to a diverse panel of three models from different vendors: *GPT-4o*, *Gemini 2.5 Flash*, and *Claude 3.7 Sonnet*. This multi judge approach is fundamental to the methodology, as it allows for the quantitative measurement of inter-rater reliability through metrics such as Cohen's and Fleiss' Kappa. This process ultimately confirms the reliability of the evaluation.

The prompt below assigns each model the persona of an *expert evaluator* to assess chunking integrity. It provides the model with the `RAW_QUESTIONS`, the `CANDIDATE_QUESTION` generated by the previous step, and the supplementary `EXPANDED_CONTEXT`. The core of the prompt lies in its detailed *Evaluation Criteria*. It provides explicit positive and negative examples of valid chunking. It is specifically designed to detect common failure modes such as *Incomplete Chunks* (omitting a direct follow-up) and *Over-chunking* (merging unrelated

questions). The strict JSON output format ensures that the binary verdict from each judge can be programmatically collected and aggregated for the final reliability analysis.

```
'''
You are an expert evaluator tasked with assessing the integrity of question chunking in a semantic text chunking system. Your primary objective is to determine whether a CANDIDATE_QUESTION is a coherent and logical extraction from the RAW_QUESTIONS.

**Input Fields:**
You will receive a JSON object containing:
1. **RAW_QUESTIONS**: Original text from the "Questions" section of a peer review.
2. **CANDIDATE_QUESTION**: A segment extracted by a model, primarily using sentences from RAW_QUESTIONS.
3. **EXPANDED_CONTEXT**: Supplementary context provided by the model to elucidate the CANDIDATE_QUESTION. This information may be inaccurate or absent and should be regarded as supplementary.

**Evaluation Criteria:**
Judge the CANDIDATE_QUESTION based on the following criterion:

- **Is the chunking valid? (is_valid_separation)**

  **Set to true if:**
  - The CANDIDATE_QUESTION integrates a main question with its direct, closely related follow-up questions or sentences from RAW_QUESTIONS.
  - The CANDIDATE_QUESTION forms a single, distinct question, separate from other questions in RAW_QUESTIONS.
  - The CANDIDATE_QUESTION predominantly consists of sentences from RAW_QUESTIONS.

  **Set to false if:**
  - **(Incomplete Chunk)**: The CANDIDATE_QUESTION omits a direct, immediate follow-up sentence from RAW_QUESTIONS, which is evidently part of the same thought.
  - **(Over-chunking)**: The CANDIDATE_QUESTION merges two or more distinct, unrelated questions from RAW_QUESTIONS.
  - **(Empty Candidate)**: The CANDIDATE_QUESTION is void, despite RAW_QUESTIONS containing actionable information.
```

```
24
25 **Output Format:**
26 - Provide your response as a single, raw JSON object.
27 - Avoid using markdown code blocks (e.g., ```json ... ```).
28 - Exclude any explanations, reasoning, or additional text.
29
30 **Example Input:**
31 {{
32     "RAW_QUESTIONS": "Why didn't the authors show
       ↪  performance on a benchmarking dataset? How does the
       ↪  algorithm work in step 4 of the proposed approach?
       ↪  Does it work in practice?",
33     "CANDIDATE_QUESTION": "How does the algorithm work in
       ↪  step 4 of the proposed approach? Does it work in
       ↪  practice?",
34     "EXPANDED_CONTEXT": "Authors propose a new approach for
       ↪  explainable AI domain. The reviewer asks about the
       ↪  inner working mechanisms of the proposed
       ↪  algorithm."
35 }}
36
37 **Example Output:**
38 {{"is_valid_separation": true}}
39
40 **Example Input:**
41 {{
42     "RAW_QUESTIONS": "In Figure 5, what are the data with
       ↪  dashed lines? Does this model have any issues in
       ↪  majority classes? Can you also share benchmark
       ↪  results on CIFAR-10?",
43     "CANDIDATE_QUESTION": "In Figure 5, what are the data
       ↪  with dashed lines? Can you also share benchmark
       ↪  results on CIFAR-10?",
44     "EXPANDED_CONTEXT": "Authors propose a new approach for
       ↪  explainable AI domain. The reviewer asks about
       ↪  benchmarks, data and visualization related
       ↪  questions."
45 }}
46
47 **Example Output:**
48 {{"is_valid_separation": false}}
49
50 {query}
51 '''
```

D.4 TopicGPT Prompt for Topic Generation

TopicGPT framework was used for the topic generation phase. While the framework provides a default prompt, it is designed for general-purpose documents. Given that the input data for this thesis consists of specific, interrogative questions from peer reviews, a domain-specific adaptation was necessary to guide the model toward generating relevant, inquiry-based themes.

The primary modification involved reframing the prompt's instructions and, most importantly, its few-shot examples. The original examples, which addressed broad topics such as "*Agriculture*" and "*Trade*" were replaced with new examples directly relevant to the peer-review context, namely "*Inquiry About Generalizability*" and "*Inquiry About Data*." This change was critical for anchoring the model's behavior, guiding it to generate topics that reflect the inherent nature of the source text rather than generic declarative statements.

```
You will receive a document in question form and a set of top-level topics from a topic hierarchy. Your task is to identify generalizable topics within the document that can act as top-level topics in the hierarchy. If any relevant topics are missing from the provided set, please add them. Otherwise, output the existing top-level topics as identified in the document.

[Top-level topics]
{Topics}

[Examples]
Example 1: Adding "[1] Inquiry About Generalizability"
Document:
If a dataset is not exchangeable, how effective is the method?

Your response:
[1] Inquiry About Generalizability: Questions regarding the limitations and boundaries of the work's applicability or how far the claims can extend.

Example 2: Duplicate "[1] Inquiry About Data", returning the existing topic
Document:
How did you deal with the labels for the data samples in test sets?

Your response:
[1] Inquiry About Data: Questions regarding specifically about the dataset(s) used in the paper, including its source, statistics, preprocessing or availability.
```

```
[Instructions]
Step 1: Determine topics mentioned in the document.
- The topic labels must be as GENERALIZABLE as possible. They must not be document-specific.
- The topics must reflect a SINGLE topic instead of a combination of topics.
- The new topics must have a level number, a short general label, and a topic description.
- The topics must be broad enough to accommodate future subtopics.
Step 2: Perform ONE of the following operations:
1. If there are already duplicates or relevant topics in the hierarchy, output those topics and stop here.
2. If the document contains no topic, return "None".
3. Otherwise, add your topic as a top-level topic. Stop here and output the added topic(s). DO NOT add any additional levels.


[Document]
{Document}

Please ONLY return the relevant or modified topics at the top level in the hierarchy. Your response should be in the following format:
[Topic Level] Topic Label: Topic Description

Your response:
```

D.5 Probabilistic Topic Assignment

This prompt enables the probabilistic assignment of topics to question chunks. Unlike standard classification tasks that request a text label, it is specifically designed to require the model to output only the numeric index of the topic. This constraint is essential for the soft classification methodology described in Section 4.5. By limiting the output to a single integer, noise in the probability distribution is minimized, allowing for the extraction of clean `top_logprobs` for entropy and uncertainty analysis. To further mitigate position bias, the `topic_list` variable was programmatically shuffled for each API call.

```
'''
You are a precise classifier. Analyze the text and assign it to exactly one of these topics by ID:
{topic_list}

CRITICAL: Output ONLY the number ID (e.g. '0', '5', '12'). Do not write the name.

{text_to_evaluate}
'''
```

D.6 Reviewer Guideline Topic Assignment

This appendix details the specific prompt engineering strategy employed in Section 4.6.2 to systematically map the extracted directive statements from institutional reviewer guidelines to the empirically derived 13-topic schema.

The prompt was designed to perform a deductive classification task. To ensure high-fidelity mapping and facilitate the subsequent *Human-in-the-Loop* audit, the prompt incorporates several structural constraints:

- *Attention Mechanism:* The model is explicitly instructed to prioritize bolded keywords within the input text, ensuring that the classification is driven by the specific semantic focus of the directive rather than peripheral context.
- *Multi-Label Classification:* Recognizing that institutional guidelines often contain multiple evaluation criteria in the same sentence, the prompt permits the assignment of multiple topics to a single directive.
- *Gap Identification:* The prompt includes a strict instruction to assign a null label ("-") if the input text does not semantically align with any of the 13 data-driven topics.
- *Chain-of-Thought Justification:* The model is required to output a "Reasoning" column, forcing it to articulate the logical link between the guideline text and the assigned topic.

The complete prompt template is presented below:

```
You are an expert academic assistant tasked with classifying peer review guidelines and questions. I will provide you with input sentences containing specific highlighted (bolded) text. Your job is to analyze these sentences and assign them to the most relevant topic(s) from the list below.

**Instructions:**
1. **Analyze the Input:** Read the text thoroughly. Pay strict attention to the **bolded** words, as they indicate the specific focus of the inquiry.
2. **Assign Topics:** Classify the text into one or more (large context overlap) of the provided topics. If the text clearly fits multiple categories based on the definitions, list all of them separated by a line break (`<br>`).
3. **Provide Reasoning:** Briefly explain why you chose the topic(s).
4. **Format:** Output a 3-column Markdown table with the headers: `guideline_text`, `assigned_topics`, and `reasoning`.
5. **Constraint:** Do not alter the input text in the first column; preserve it exactly as given, including the markdown bolding.
6. **Exception:** Assign "-" if none of the topics can be applied.

**List of Topics and Definitions:**
{topic_description_pairs}

**Few-Shot Examples:**

**Input:**
What is the **problem** that the paper **aims to address**?

**Output:**
| guideline_text | assigned_topics | reasoning |
| :--- | :--- | :--- |
| What is the **problem** that the paper **aims to address**? | Inquiry About Motivation and Contribution | The phrase "aims to address" and "problem" relates to the purpose and rationale of the research, which fits the scope of Motivation and Contribution. |

**Input:**
```

```
Are there any **errors**, **unstated assumptions**, or **missing details**?

**Output:**
| guideline_text | assigned_topics | reasoning |
| :--- | :--- | :--- |
| Are there any **errors**, **unstated assumptions**, or **missing details**? | Inquiry About Theoretical Foundations<br>Inquiry About Validity and Reproducibility | The mention of "**unstated assumptions**" directly maps to **Inquiry About Theoretical Foundations**. The inquiry regarding "**errors**" and "**missing details**" pertains to correctness and replication, falling under **Inquiry About Validity and Reproducibility**. |

**Current Task:**

Input: {guideline_chunk_image}
```