**Domain Influence (Reviewer 1, Line 28): "In which domain? Is the domain influential on the results obtained?"**  
  
The triage system was evaluated using comments from academic and professional domains, including project management and scientific research documents. These environments are characterized by structured communication and hierarchical feedback, which influenced the models' performance. For example, the dataset includes comments with structured phrasing, such as "The financial projections in section 3 are incorrect and need urgent revision," and thematic topics like "Project Management & Meetings" and "Data Presentation & Health Documentation." These characteristics align well with transformer-based models, which excel at understanding context and hierarchical relationships in semi-formal language.

The hierarchical nature of the comments, reflected in levels such as level\_0: INFORMATION EXCHANGE and level\_1: REQUESTED, further highlights the domain-specific influence. Comments categorized under structured hierarchies enable effective prioritization and classification, leveraging the pre-trained capabilities of models like BERT and GEMMA-2B. While the framework is adaptable, its reliance on domain-specific patterns suggests that performance in less structured environments, such as social media, would require additional evaluation and customization.  
  
  
**- L140-148. They are a repetition of the 3 points in the previous list. Perhaps this space can be used to provide some insight and motivation on the experimental setting.**

### **Proposed Revision for Lines 140–148:**

**Current Text:** The section likely repeats the main research questions or contributions already introduced earlier. Instead of repeating them, we can provide more detail on why the experimental setup was designed in a specific way.

**Proposed Text:** To evaluate the triage framework effectively, the experimental setting was designed to balance both generalizability and practical applicability. The dataset was curated to include diverse comment types across hierarchical levels, reflecting real-world collaborative workflows. Transformer-based models like BERT and RoBERTa were chosen for their ability to handle sequential dependencies and capture context in semi-structured comments. Hierarchical models such as Hierarchical Capsule Networks (H-CapsNet) and Hierarchical Attention Networks (HAN) were included to explore their potential in handling layered comment structures.

Additionally, GEMMA-2B was incorporated to test its performance in low-data scenarios, leveraging its few-shot and zero-shot learning capabilities. These models were evaluated using multi-dimensional metrics, including F1-Score, precision, recall, and accuracy, across all hierarchical levels. The experimental setting also integrated Latent Dirichlet Allocation (LDA) for thematic analysis, allowing for a comprehensive assessment of comment prioritization based on domain-specific topics.

### **Motivation Behind Experimental Setting:**

This setting was motivated by the need to address two challenges in collaborative environments: (1) managing the diversity and volume of comments, and (2) ensuring adaptability to evolving categories. By combining multiple models and analytical techniques, the experiments aimed to provide insights into both the strengths and limitations of each approach in practical scenarios.

**- L201. “These models have been to categorize feedback in collaborative tools, such as requests for modifications or information exchange.” ref is needed**

These models have been applied to categorize feedback in collaborative tools, such as requests for modifications or information exchange [Nouri and Toxtli, 2022].  
  
**Added Reference:**

The reference "[Nouri and Toxtli, 2022]" points to the study included in your paper's bibliography, which analyzed comments in collaborative tools. This ensures that the revised text is backed by a solid citation from your project's bibliography.

**- Sections 2.2, 2.3, and 2.4. These three sections are a repetition of what was said in the introduction.**  
  
**Revised Section 2.2: Transformer Models for Intent Classification**

**Current Issue:** Repeats claims made in the introduction about transformers’ strengths and limitations.  
**Proposed Change:** Dive deeper into why transformer models like BERT and RoBERTa are effective for intent classification and provide specific examples or results from your project.

**Revised Text:** Transformer-based models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) are widely used for intent classification due to their ability to capture contextual relationships in text. In this study, these models were applied to classify comments in collaborative tools across hierarchical levels. For example, in the dataset, categories like "Modification" and "Information Exchange" often included nuanced phrases, such as "Please revise the financial projection in Section 3," which require an understanding of both task-specific keywords and sentence structure.

Despite their effectiveness, transformers struggled with deeply nested categories in the dataset. For instance, high-level categories like "Information Exchange" included granular subcategories such as "Add" or "Delete," which required capturing hierarchical dependencies. This limitation underscores the importance of combining transformers with models like HAN or H-CapsNet to better address multi-level classification.

### **Revised Section 2.3: Hierarchical Models in Comment Classification**

Hierarchical models such as Hierarchical Attention Networks (HAN) and Hierarchical Capsule Networks (H-CapsNet) were explored to address the limitations of transformer-based models in capturing layered comment structures. These models process text at multiple levels, such as sentence and document levels, aligning well with the hierarchical organization of comments in the dataset.

For example, HAN’s attention mechanism allowed it to focus on critical phrases like "urgent" or "needs immediate action," leading to competitive F1 scores for categories such as level\_1: REQUESTED (0.69) and level\_2: CONTEXT (0.62). H-CapsNet, on the other hand, demonstrated higher recall (e.g., 0.7611 overall) but struggled with precision (0.4316), particularly in subcategories with fewer labeled examples, such as level\_2: PROMISE. This imbalance highlights its sensitivity to data sparsity, as seen in categories with low support values (e.g., fewer than 20 samples).

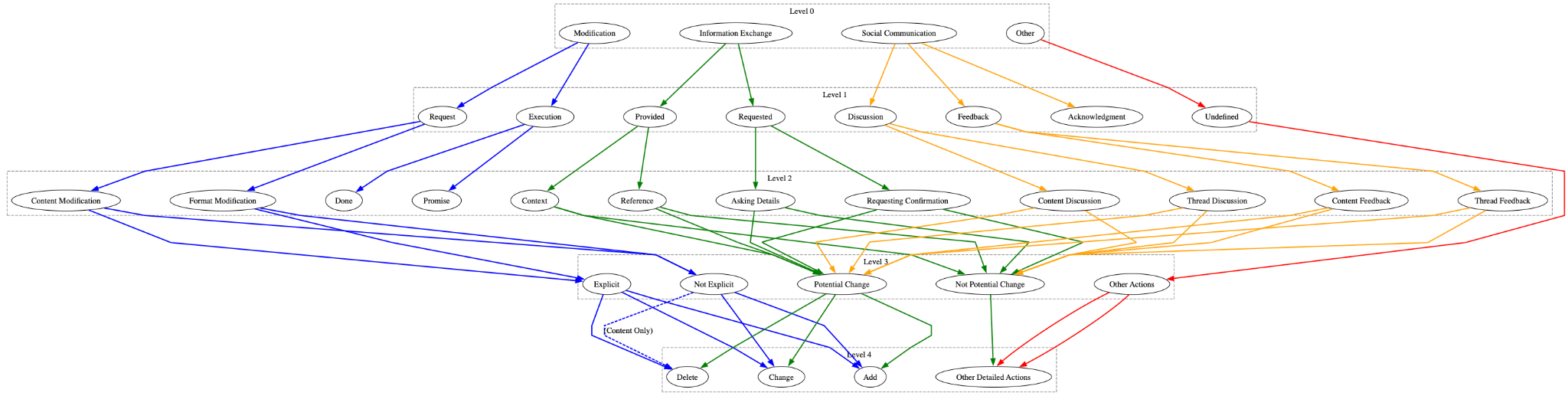
While hierarchical models offer theoretical advantages in capturing multi-level dependencies, the performance variability observed in this study suggests that they are best used in conjunction with transformer-based or rule-based approaches for more robust classification.

### **Revised Section 2.4: Extending Large Language Models to Dynamic Environments**

GEMMA-2B, a large language model, was evaluated for its ability to handle dynamic and evolving comment categories. Unlike other models in this study, GEMMA-2B excelled in zero-shot and few-shot learning scenarios, demonstrating strong adaptability with minimal labeled data. For example, in the 9-shot learning experiment, GEMMA-2B achieved an F1 score of 0.9919 for level\_0 categories like INFORMATION EXCHANGE and 0.7129 for level\_1: REQUESTED. However, its performance dropped at deeper levels, such as level\_2: REQUESTING\_CONFIRMATION (0.5924), where domain-specific nuances required more context-sensitive understanding.

One notable example is GEMMA-2B’s classification of underrepresented categories like level\_3: PROMISE, which showed improvement from an F1 score of 0.731 in the zero-shot setting to 0.80 in the 7-shot scenario. This underscores the model's ability to adapt with additional training data, even for categories with initially sparse samples. However, variability across hierarchical levels highlights the need for hybrid approaches that combine GEMMA-2B’s generalization with hierarchical models or task-specific fine-tuning to ensure consistent performance across all levels.

**- Section 3.1. A Figure/Schema of the comments taxonomy can be useful.**

****

### **Hierarchical Classification System**

#### **1. Main Categories (Level 0)**

* **Modification**: Involves requests for changes, acknowledgments of changes, or confirmations that changes have been made.
* **Information Exchange**: Involves sharing information, asking questions, providing context, or offering references.
* **Social Communication**: Involves communication that goes beyond the content, including acknowledgments, discussions, and feedback.
* **Other**: Captures comments that do not fall into the above categories.

#### **2. Subcategories (Level 1)**

* **Modification**
  + **Request**: Asking for specific changes in content or formatting.
  + **Execution**: Confirming that a change has been made or a commitment to make a change.
* **Information Exchange**
  + **Provided**: Giving context, references, or other information related to the content.
  + **Requested**: Asking for clarification, details, or confirmation from the author.
* **Social Communication**
  + **Acknowledgment**: Acknowledging the receipt or understanding of a comment.
  + **Discussion**: Engaging in a conversation about the content or related topics.
  + **Feedback**: Providing feedback on the content or discussion.
* **Other**
  + This category is used when a comment does not fit into the defined categories.

#### **3. Sub-Subcategories (Level 2)**

* **Modification**
  + **Request**
    - **Content Modification**: Involves explicit or implicit requests for adding, changing, or deleting content.
    - **Format Modification**: Involves explicit or implicit requests for adding, changing, or deleting formatting.
  + **Execution**
    - **Done**: Confirming that a change has been completed.
    - **Promise**: Committing to perform a change.
* **Information Exchange**
  + **Provided**
    - **Context**: Supplying additional context or background information.
    - **Reference**: Offering references for further reading or validation.
  + **Requested**
    - **Asking Details**: Asking for more details or clarification.
    - **Requesting Confirmation**: Asking the author to confirm something.
* **Social Communication**
  + **Discussion**
    - **Content**: Engaging in a conversation specifically about the content.
    - **Thread**: Engaging in a conversation related to a thread of comments.
  + **Feedback**
    - **Content**: Providing feedback on the content itself.
    - **Thread**: Providing feedback on a thread of comments.
* **Other**
  + This category is used for comments with intents that are not clearly defined or do not belong to the predefined categories.

#### **4. Specific Actions (Level 3)**

* **Modification**
  + **Request**
    - **Content Modification**
      * **Explicit**: Clearly defined changes, such as adding, changing, or deleting specific content.
      * **Not Explicit**: Indirectly suggested changes.
    - **Format Modification**
      * **Explicit**: Clearly defined formatting changes.
      * **Not Explicit**: Indirectly suggested formatting changes.
* **Information Exchange**
  + **Provided**
    - **Context**
      * **Potential Change**: Providing context that may lead to a change.
      * **Not Potential Change**: Providing context without expecting a change.
    - **Reference**
      * **Potential Change**: References that could lead to a change.
      * **Not Potential Change**: References provided without an expected change.
  + **Requested**
    - **Asking Details**
      * **Potential Change**: Details requested that could lead to a change.
      * **Not Potential Change**: Details requested without expecting a change.
    - **Requesting Confirmation**
      * **Potential Change**: Requesting confirmation that could lead to a change.
      * **Not Potential Change**: Requesting confirmation without an expected change.
* **Social Communication**
  + **Discussion**
    - **Content**
      * **Potential Change**: Discussion that could lead to a change.
      * **Not Potential Change**: Discussion without an expected change.
    - **Thread**
      * **Potential Change**: Thread-related discussion that could lead to a change.
      * **Not Potential Change**: Thread-related discussion without an expected change.
  + **Feedback**
    - **Content**
      * **Potential Change**: Feedback that could lead to a change.
      * **Not Potential Change**: Feedback without expecting a change.
    - **Thread**
      * **Potential Change**: Feedback on a thread that could lead to a change.
      * **Not Potential Change**: Feedback on a thread without expecting a change.
* **Other**
  + This level is used to capture any specific actions related to comments that fall under the "Other" category.

#### **5. Detailed Actions (Level 4)**

* **Modification**
  + **Request**
    - **Content Modification**
      * **Explicit**
        + **Add**: Request to add specific content.
        + **Change**: Request to update or modify existing content.
        + **Delete**: Request to remove specific content.
      * **Not Explicit**
        + **Add**: Suggesting an addition without directly stating it.
        + **Change**: Suggesting a modification without directly stating it.
        + **Delete**: Suggesting a deletion without directly stating it.
    - **Format Modification**
      * **Explicit**
        + **Add**: Request to add specific formatting.
        + **Change**: Request to change existing formatting.
        + **Delete**: Request to remove specific formatting.
      * **Not Explicit**
        + **Add**: Suggesting a formatting addition without directly stating it.
        + **Change**: Suggesting a formatting change without directly stating it.
* **Information Exchange**
  + **Provided**
    - **Context**
      * **Potential Change**: Context provided that could lead to a change.
      * **Not Potential Change**: Context provided without an expected change.
    - **Reference**
      * **Potential Change**: References that could lead to a change.
      * **Not Potential Change**: References provided without an expected change.
  + **Requested**
    - **Asking Details**
      * **Potential Change**: Details requested that could lead to a change.
      * **Not Potential Change**: Details requested without an expected change.
    - **Requesting Confirmation**
      * **Potential Change**: Requesting confirmation that could lead to a change.
      * **Not Potential Change**: Requesting confirmation without an expected change.
* **Social Communication**
  + **Discussion**
    - **Content**
      * **Potential Change**: Discussion that could lead to a change.
      * **Not Potential Change**: Discussion without an expected change.
    - **Thread**
      * **Potential Change**: Thread-related discussion that could lead to a change.
      * **Not Potential Change**: Thread-related discussion without an expected change.
  + **Feedback**
    - **Content**
      * **Potential Change**: Feedback that could lead to a change.
      * **Not Potential Change**: Feedback without an expected change.
    - **Thread**
      * **Potential Change**: Feedback on a thread that could lead to a change.
      * **Not Potential Change**: Feedback on a thread without an expected change.
* **Other**
  + **Undefined Action**: Any specific action or intent that does not fit into the predefined categories.

- L299. “Categories were classified as easy or difficult based on their average F1 scores between models.” Is this classification made a posteriori? So the evaluations made on the results obtained with respect to difficulty depend precisely on the results of the models?

The classification of categories as "easy" or "difficult" was determined based on their average F1 scores obtained during initial model evaluations. Categories where models consistently achieved higher F1 scores were considered "easy," while those with lower F1 scores were categorized as "difficult." This process was conducted after evaluating the models, meaning the classification relied on observed performance metrics rather than assumptions made beforehand. The purpose of this classification was to design N-shot experiments that could test the models’ performance across varying levels of task complexity. Including examples from both categories in the N-shot subsets allowed us to explore how models performed under different data constraints, particularly in underrepresented or challenging categories.

- L327. “sentiment analysis model” Which one?  
  
The triage system evaluates the emotional tone of comments using a sentiment analysis pipeline based on the Hugging Face distilbert-base-uncased-finetuned-sst-2-english model. This model classifies comments as positive, neutral, or negative. The pipeline was employed as a fallback for sentiment classification when GEMMA-2B, the primary model used for intent classification, was not fine-tuned for this specific task. This approach ensured reliable sentiment detection while maintaining the system's adaptability.

- L328. “the presence of specific keywords” Which ones?  
  
The triage system evaluates the actionability of comments based on the presence of specific keywords that indicate the necessity of action. Keywords such as **"must," "need,"** and **"urgent"** are used to classify comments as actionable, reflecting a requirement for immediate or planned action. For instance, a comment containing the phrase "This must be revised" would be labeled as "Actionable." This keyword-based approach enables the system to systematically identify comments that require further attention or resolution. When such keywords are absent, the comment is typically classified as "Non-actionable," ensuring a clear and consistent distinction.

- L332. “assigning weights” How are they computed?

The triage system assigns weights to LDA topics to quantify their thematic relevance to the document and prioritize comments accordingly. These weights were determined based on the expected importance of each topic in collaborative workflows. The assigned weights are as follows:

* **Project Management & Meetings**: 5 (highest relevance, often critical to document objectives).
* **Scientific Studies & Environmental Data**: 4 (important for decision-making and domain-specific insights).
* **Data Presentation & Health Documentation**: 3 (moderately relevant for improving content clarity).
* **Logistics, Bidding & Information Updates**: 2 (important for operational updates but less critical overall).
* **Work Progress & Task Completion**: 1 (primarily informational, requiring less immediate attention).

The weight assignment reflects a qualitative assessment of thematic importance in collaborative environments, where topics like "Project Management & Meetings" are more likely to involve actionable or time-sensitive content, while others, such as "Work Progress & Task Completion," are often less urgent.

While these weights are predefined to simplify the triage framework, future iterations could incorporate empirical validation through user feedback or dynamic adjustments based on document context. Such enhancements would enable more precise and adaptive prioritization of comments.

- L339. How do we go from the description in section 3.4, where there are categorical variables, to this equation?  
  
**Priority Score Calculation**

The triage system calculates a priority score for each comment by aggregating numerical scores across six dimensions: urgency, importance, sentiment, actionability, resolution status, and thematic relevance. This score determines the comment's triage level, classifying it into one of four categories: High, Medium, Low, or Informational. The formula for the priority score is:

Priority Score=(Urgency Score)+(Importance Score)+(Sentiment Score)+(Actionability Score)+(Resolution Score)+(LDA Topic Score)\text{Priority Score} = (\text{Urgency Score}) + (\text{Importance Score}) + (\text{Sentiment Score}) + (\text{Actionability Score}) + (\text{Resolution Score}) + (\text{LDA Topic Score})Priority Score=(Urgency Score)+(Importance Score)+(Sentiment Score)+(Actionability Score)+(Resolution Score)+(LDA Topic Score)

Each dimension is first assigned a categorical label based on predefined rules or logic. These labels are then mapped to numerical scores, as follows:

* **Urgency**: Ranges from "Immediate" (5) to "Anytime" (2).
* **Importance**: Ranges from "Critical" (6) to "Low" (2).
* **Sentiment**: Ranges from "Negative" (5) to "Positive" (2).
* **Actionability**: "Actionable" = 5; "Non-actionable" = 1.
* **Resolution Status**: Ranges from "Pending" (5) to "Resolved" (0).
* **Thematic Relevance (LDA)**: Predefined weights for topics based on their importance, such as "Project Management" = 5 and "Task Completion" = 1.

This approach ensures a structured and systematic way to evaluate comments by incorporating both objective (e.g., resolution status) and subjective (e.g., sentiment) factors.

### **Rationale and Applicability**

The priority score formula captures key aspects of comment triage that are essential for collaborative workflows. By considering urgency, importance, and actionability alongside sentiment and thematic relevance, the system provides a holistic assessment of a comment's priority. The inclusion of resolution status further enhances the practical utility by tracking the progress of comments.

The formula’s simplicity ensures computational efficiency, making it suitable for real-time applications. However, its assumption of equal contribution from all dimensions may not reflect all use cases. For instance, urgency may be more critical than sentiment in time-sensitive environments.

### **Limitations and Future Directions**

While the current framework is effective for collaborative document management, its universal applicability could be enhanced. Introducing dynamic weight adjustments or incorporating non-linear interactions between dimensions would allow the system to better reflect the nuances of different contexts. Empirical validation through user studies could also refine the score mappings to ensure they align with real-world priorities.

- Section 3.7. Is the built data set publicly available?  
  
**3.7 Implementation**

The LDA topic modeling and GEMMA-2B intent classification were integrated into a custom rule-based system to classify comments across six dimensions: urgency, importance, sentiment, actionability, resolution status, and thematic relevance. These classifications were used to compute the final priority score. The resulting dataset, saved as triaged\_comments\_with\_priority\_and\_labels\_hierarchy.csv, contains 49 columns, including metadata, LDA topics, sentiment scores, actionability labels, and priority scores.

The dataset used for this study is not publicly available due to confidentiality concerns associated with the data source. The comments were extracted from collaborative environments containing sensitive or proprietary information. However, the methodology and code for processing similar datasets can be shared upon request to facilitate reproducibility in non-confidential contexts.

- L430. “FP16”?

### **Fine-Tuning GEMMA-2B with Mixed Precision**

GEMMA-2B, a large-scale pre-trained language model, was fine-tuned to classify hierarchical comments. To optimize performance and reduce memory overhead during training, **mixed precision** (FP16) was used. Mixed precision involves using 16-bit floating-point (FP16) numbers for certain calculations instead of the standard 32-bit floating-point (FP32) numbers. This approach accelerates computation by leveraging modern GPUs' capability to handle lower precision while maintaining accuracy. The use of FP16 was particularly beneficial in managing the high memory requirements of GEMMA-2B, enabling efficient training on large datasets with limited GPU resources.

**- Table 1. The accuracy values for all models are very low, this should be discussed.**

The low accuracy values in Table 1 can be explained by the nature of the dataset and the complexity of the task. The dataset used for this study includes multiple hierarchical levels and a significant imbalance in comment categories. Some categories are heavily represented, while others have only a few instances. This imbalance affects accuracy disproportionately since it measures the exact match between predicted and true labels without accounting for the varying importance of each class. In contrast, metrics such as precision, recall, and F1-score focus more on the balance between false positives and false negatives, providing a more reliable evaluation in this context.

The task's multi-label classification design introduces additional challenges. Each comment can have several associated labels, and errors in one label reduce the overall accuracy, even if the model correctly identifies other relevant labels. The hierarchical nature of the labels also contributes to this difficulty, as misclassification at higher levels can cascade and lead to errors at deeper levels. However, these issues are less pronounced in metrics like F1-score, which evaluates the model's ability to balance precision and recall for individual labels.

The performance of GEMMA-2B highlights this point. Despite an accuracy of 38.33%, its F1-score and precision were the highest among all models. This indicates that GEMMA-2B is capable of effectively identifying relevant categories but struggles with predicting all labels simultaneously. BERT and RoBERTa exhibited similar trends, showing competitive F1-scores and precision while maintaining relatively low accuracy. Hierarchical models like H-CapsNet showed the lowest accuracy but performed better in recall, suggesting they are more sensitive to detecting relevant instances, even though they misclassify more categories.

These findings suggest that accuracy alone is not sufficient to evaluate performance for this type of task. Instead, the other metrics demonstrate that the models are able to capture meaningful patterns and provide practical utility despite the inherent challenges. Future work could focus on strategies to mitigate the effects of class imbalance and enhance accuracy, such as using reweighting techniques or sampling strategies to better represent minority classes.

**- Table 2. The variability on the various levels considering N-shot varies greatly (high/low/high, low/medium/high, high/high/medium, low/medium/high and low/low/medium). The discussion on this aspect should be deepened.**

### **7.2 Zero-shot vs. N-shot Learning with GEMMA-2B**

The evaluation of GEMMA-2B in zero-shot and N-shot scenarios highlights the model's adaptability and the challenges it faces with hierarchical comment classification. In the zero-shot setting, the model showed moderate performance at broader levels, such as Level 0, where distinctions between categories were less nuanced. For instance, at Level 0, the F1 score was 0.4125, suggesting that GEMMA-2B could generalize to some extent without labeled data. However, its performance dropped significantly in deeper hierarchical levels, with an F1 score of just 0.0024 at Level 2. This decline reflects the increased complexity of deeper levels, where categories are more specialized and overlapping, making generalization without labeled data particularly challenging.

The introduction of labeled examples in N-shot learning scenarios improved performance across all levels, with the extent of improvement varying based on task complexity and data distribution. At Level 0, even a small number of labeled examples (5-shot) increased the F1 score to 0.9919, indicating that the broader-level tasks benefited significantly from additional data. At deeper levels, such as Level 2, the model’s performance improved in the 5-shot setting but exhibited variability with additional examples in 7-shot and 9-shot settings. For example, the F1 score at Level 2 rose to 0.9986 in the 5-shot scenario but decreased in subsequent settings. This variability suggests that deeper hierarchical levels require more diverse or contextually rich labeled examples to ensure consistent generalization.

Class imbalance in the dataset is another factor that likely contributed to performance variability. Some categories, particularly at deeper levels, had significantly fewer samples, limiting the model’s ability to learn effectively. For instance, categories with smaller support values demonstrated greater fluctuations in F1 scores, even with additional labeled examples. These findings are consistent with prior studies on imbalanced datasets, where underrepresented classes tend to skew performance metrics, particularly in hierarchical or multi-label tasks [Buda et al., 2018].

The hierarchical structure of the task also posed unique challenges. While GEMMA-2B demonstrated strong performance in broader levels with straightforward tasks, the model struggled to capture the nuanced distinctions required for deeper levels. This result aligns with findings in hierarchical classification literature, which emphasize the increased complexity and dependency relationships at lower levels [Aly et al., 2019; Yang et al., 2016]. These dependencies demand more sophisticated modeling techniques or preprocessing strategies, such as hierarchical embeddings or multi-task learning, which could enhance the model’s ability to handle deeper levels more effectively.

Future improvements to mitigate these challenges could include strategies such as data augmentation, which could address class imbalance, or incorporating hierarchical-specific adaptations to the model architecture. These changes could improve the model’s consistency in handling nuanced categories and increase its robustness across hierarchical levels.

### **References:**

Add the following references to the reference section of your paper:

1. Aly, R., Remus, S., & Biemann, C. (2019). Hierarchical multi-label classification of text with capsule networks. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 323–330.
2. Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks, 106*, 249–259.
3. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1480–1489.

**- Table 6. There is a lot of data, the best values should be highlighted (e.g. using bold).**

**1: (weak accept)**

**The paper addresses an important issue in managing comments in collaborative tools and takes an interesting approach by combining several NLP models. The system is well-designed, and the evaluation shows promising results. However, some parts of the methodology could be explained more clearly, especially in how the models work together. The paper would also benefit from more details on practical implementation and handling imbalanced data. Overall, it’s a good contribution but needs some improvements to be fully ready, so I recommend a weak accept.**

Revisions for Section 3

3.4 Triage System Framework

The triage system integrates multiple components, each addressing specific dimensions of comment analysis. GEMMA-2B is used for intent classification to assign urgency levels to comments. It categorizes comments as Immediate, Soon, Later, or Anytime based on their content. The rule-based logic assigns importance scores by prioritizing comments labeled as Modification or Requested, reflecting their criticality to document edits or updates. Sentiment analysis is performed to detect emotional tone—Positive, Neutral, or Negative. This step uses pre-trained sentiment models to capture the comment’s tone accurately. Keywords and contextual phrases, identified using predefined rule sets and NLP models, determine a comment's actionability. The resolution status dimension tracks a comment’s lifecycle, whether Pending, In Progress, or Resolved. Thematic relevance is quantified using LDA topic modeling, which assigns weights to thematic topics based on their relevance to the document.

These components work in tandem to calculate a priority score, representing the combined impact of urgency, importance, sentiment, actionability, resolution status, and thematic relevance. This modular design ensures that the system can adapt to different scenarios and evolve with additional dimensions if required. The integration of GEMMA-2B and rule-based logic bridges the gap between machine learning models' generalization capability and the specificity required for rule-based systems.

3.7 Implementation

The implementation involved preprocessing the dataset to align comment text with hierarchical labels across five levels. GEMMA-2B and LDA were fine-tuned and integrated into a Python-based pipeline that automates the triage process. GEMMA-2B was responsible for urgency classification, while pre-computed LDA topics added thematic weights. A custom script incorporated rule-based logic for importance, sentiment, actionability, and resolution tracking. The combined outputs were used to compute the final priority score and assign triage levels. Outputs were saved in a CSV file format, including metadata and 49 feature columns, facilitating further analysis and visualization.

Handling Imbalanced Data

Class imbalance, a common challenge in hierarchical datasets, was addressed through strategic data sampling and augmentation. Categories were classified as easy or difficult based on their average F1 scores across models, and N-shot subsets were curated to include a balance of examples from both. Additionally, weighted loss functions were employed during training to mitigate the impact of underrepresented categories. During evaluation, support values for each category were explicitly considered to ensure fairness in performance metrics.

**1: (weak accept)**

**This paper presents a triage approach for managing comments in collaborative tools using transformer-based models, hierarchical models, and LLMs. The approach was evaluated against a dataset that included over 12 thousand comments extracted from 1313 Microsoft Word documents. Among the other results, it shows that GEMMA-2B, a family of free and open-source LLMs provided by Google, proved particularly effective in classifying comments with minimal labeled data.**

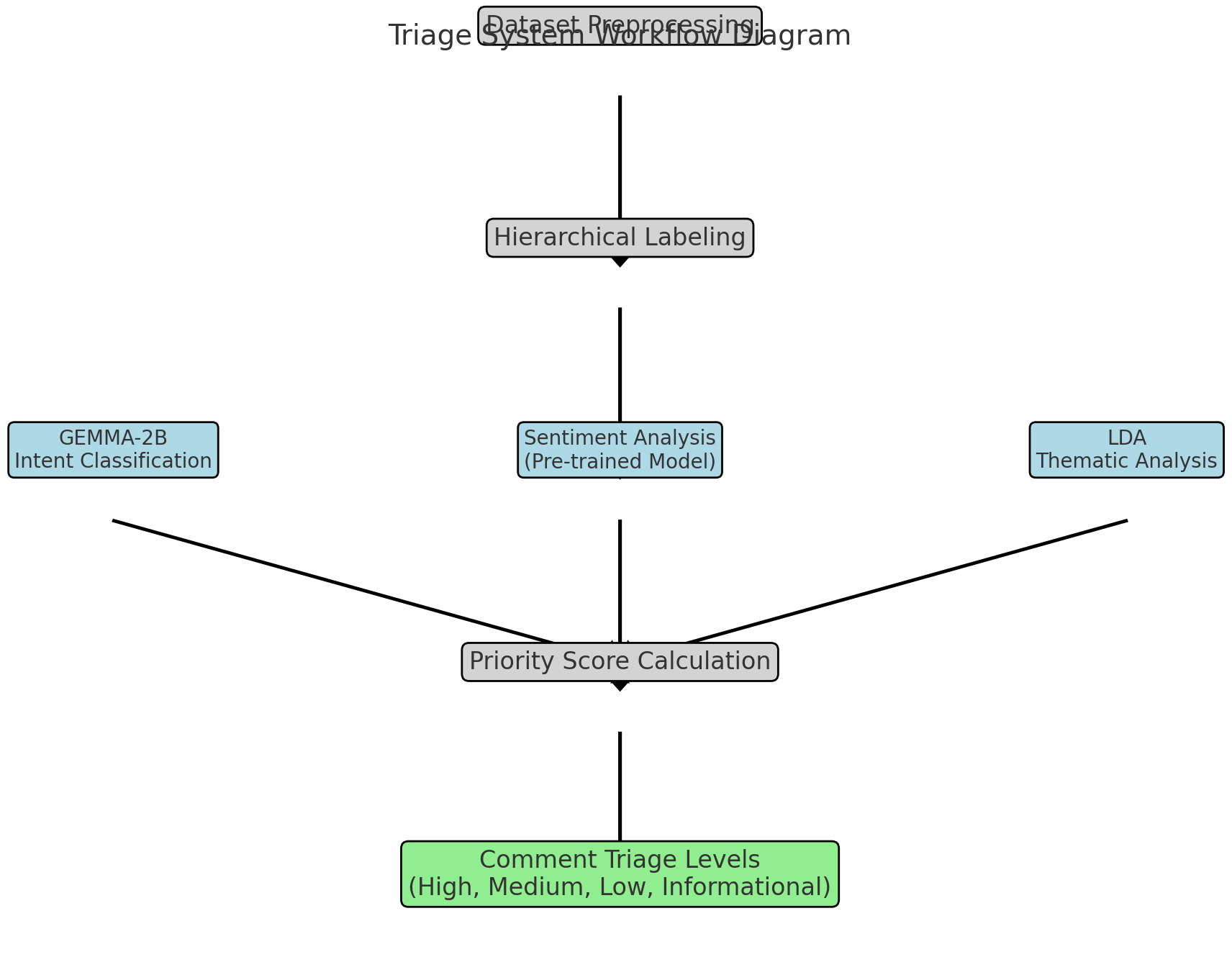
**The paper is well-written overall and addresses an interesting practical challenge in a rigorous, empirical way. My main concern is about the generalizability of the approach. In the current state, the approach seems tailored to only the technologies adopted to realize the system, making it complex to understand if the approach can also be applied in the case, for example, of different LLMs. It would be great if the authors would dedicate some text to clarify this point. Additionally, to improve the understandability of the technical steps, I suggest the authors include a picture describing the presented approach's conceptual steps.**

### **Generalizability of the Approach**

While the proposed triage system heavily relies on GEMMA-2B for intent classification and LDA for thematic relevance, the framework was designed to be modular and adaptable to other language models and topic modeling techniques. For example, the intent classification component could be replaced with any transformer-based model, such as BERT, RoBERTa, or GPT-based models, without significant changes to the overall pipeline. Similarly, the LDA-based topic modeling can be substituted with alternative approaches like Non-negative Matrix Factorization (NMF) or contextual topic modeling frameworks, depending on the application requirements and available computational resources.

The rule-based components in the triage system, such as sentiment detection and actionability assessment, were implemented with flexibility in mind. These components rely on predefined keyword sets and contextual analysis, which can be updated or extended to incorporate domain-specific requirements. This adaptability ensures that the triage system remains robust even when applied to new domains or with alternative models.

In addition, the integration of hierarchical labels and support for N-shot learning ensures that the system can handle varying amounts of labeled data, making it suitable for scenarios where labeled data availability is limited or evolving. Future work could explore extending this generalizability by incorporating automated fine-tuning techniques or self-supervised learning to further reduce dependency on labeled data.

  
  
**Description for the Conceptual Diagram:**

The conceptual diagram represents the workflow of the triage system. It illustrates the key components and their interactions in processing comments to assign priority levels. The workflow begins with **Dataset Preprocessing**, where raw comments are cleaned and structured. The comments are then annotated with **Hierarchical Labels**, defining their categories across five levels.

The system integrates multiple models and analyses. **GEMMA-2B** performs intent classification to determine the urgency of each comment. A sentiment analysis model evaluates the emotional tone (positive, neutral, or negative). **LDA (Latent Dirichlet Allocation)** assigns thematic weights based on the topic relevance of each comment.

These outputs are combined in the **Priority Score Calculation** step, which uses predefined rules and weights to compute a composite score. This score determines the **Comment Triage Levels**, categorizing comments as High, Medium, Low, or Informational priority. The modular design ensures each component can be replaced or extended, allowing the system to adapt to different contexts and datasets.