Master's Thesis

Estimating Text Concreteness in Online Discussions

HU Mingxi

Supervisor Professor HASEGAWA Shinobu

Japan Advanced Institute of Science and Technology
Division of Advanced Science and Technology
(Information Science)

March 2025

Abstract

With the increasing prevalence of online forums as platforms for discussion, evaluating the quality of user-generated content has become a significant challenge. One crucial aspect of discussion quality is concreteness, which influences readability, engagement, and effective communication. Concreteness refers to the extent to which a text includes specific, detailed, and vivid information, as opposed to abstract and generalized statements. However, prior research has primarily focused on sentence- or word-level concreteness, with limited exploration of how entire comments in online discussions exhibit concreteness. To address this gap, this study explores the estimation of abstraction-concreteness (AC) scores in online discussion texts, aiming to improve the detection of high-quality contributions. While online forums facilitate knowledge exchange, the quality of content varies significantly, making it essential to understand how concreteness impacts discussion effectiveness. This understanding is crucial for enhancing content filtering and recommendation systems. Prior research has examined various linguistic factors affecting text quality, yet a systematic approach to estimating AC in user-generated content remains underdeveloped. This study seeks to bridge this gap by constructing predictive models for AC scores and investigating five supporting dimensions—Actionability, Clarity, Orientation, Relevance, and Specificity—as key factors influencing AC.

To achieve these goals, a dataset was constructed by collecting Reddit comments from the ExplainLikeImFive subreddit, a platform where users explain complex topics in simple terms. The dataset includes human-annotated AC scores and supporting dimensions, obtained through Amazon Mechanical Turk (MTurk). A rigorous filtering process was applied to improve inter-rater consistency, ensuring that the annotations reflect a shared understanding of the abstract-concrete spectrum. The study employed a combination of human annotations, dictionary-based estimates, and machine learning models to analyze and predict AC scores. The research methodology involved three primary modeling approaches: (1) rule-based weighting using Pearson correlations, (2) linear regression with five supporting dimensions as predictors, and (3) feature extraction models based on TF-IDF regression and GPT-4 few-shot prompting. Additionally, a baseline estimate was derived using word concreteness scores from psycholinguistic dictionaries to provide a reference for model performance. The dataset underwent pre-processing, including filtering to remove inconsistencies and ensuring that only high-agreement annotations were retained for model training.

Experimental results indicate that models leveraging structured feature inputs outperform text-based approaches. The linear regression model demonstrated the highest

accuracy, achieving the lowest Mean Absolute Error (MAE = 0.34) and Root Mean Squared Error (RMSE = 0.41), while also being the only model with a positive R-squared value ($R^2 = 0.24$). This suggests that integrating structured linguistic features significantly enhances predictive performance. In contrast, text-based models, including TF-IDF regression and GPT-4 predictions, exhibited significantly higher errors, with GPT-4 generating particularly inconsistent scores. The findings confirm that concreteness perception extends beyond word-level features and is better captured through structured linguistic dimensions rather than raw textual analysis alone.

Further analysis of the supporting dimensions revealed that Clarity and Specificity exhibited stronger correlations with AC, suggesting that well-defined and detailed content is perceived as more concrete. Actionability, while relevant, showed weaker correlations with AC, indicating that while providing actionable advice contributes to text quality, it does not necessarily align with higher concreteness perceptions. The study also assessed the feasibility of automatically extracting these dimensions from raw text. TF-IDF regression consistently outperformed GPT-4 across all dimensions, particularly in Clarity and Specificity. However, GPT-4 showed substantial inconsistencies, especially in Relevance and Orientation, indicating that while large language models capture general linguistic patterns, they struggle with nuanced text quality assessments without domain-specific fine-tuning.

The implications of this research extend beyond theoretical linguistics, offering practical applications for improving online discourse. By integrating AC estimation models into discussion platforms, moderators can enhance content ranking mechanisms, prioritizing well-structured and informative comments. Additionally, automated concreteness scoring could benefit educational applications, helping instructors assess the clarity and specificity of student responses in discussion-based learning environments. Content creators and writers can also leverage AC scoring to refine their writing, ensuring that their work is engaging and easy to understand. Furthermore, AC scoring could be integrated into automated writing assistants and feedback systems, improving real-time text suggestions and readability evaluations.

One of the broader applications of AC estimation lies in misinformation detection and fact-checking. More concrete statements often contain verifiable information, whereas highly abstract statements may be more prone to misinterpretation or fabrication. By assessing concreteness levels in social media discourse, AC scoring could serve as an additional layer of verification for content credibility assessments. Similarly, in corporate and legal communication, ensuring high concreteness can aid in drafting clearer policies and legal documents, reducing ambiguities that may lead to misinterpretation. Another

key application is in search engine optimization and content recommendation, where highlighting concrete and informative content can improve user engagement and knowledge retention.

Future research should refine the definitions of AC and its supporting dimensions to improve annotation consistency and model interpretability. Expanding the dataset to include diverse text genres, such as scientific literature, journalistic writing, and instructional materials, would enhance model generalization. Additionally, analyzing cross-linguistic differences in AC perception could provide valuable insights into how concreteness varies across cultures and languages. Investigating how contextual features, such as the discourse structure of comments, affect AC ratings could also improve model robustness. Further, deep learning techniques, including fine-tuned transformer models like BERT, could be explored to improve prediction accuracy by leveraging contextual embeddings. Hybrid models that integrate linguistic features with neural networks may provide a balanced approach, combining interpretability with predictive power.

Another promising avenue for future work is refining automated annotation techniques to reduce reliance on human labeling. Leveraging active learning strategies, where models select the most uncertain samples for human review, could improve annotation efficiency while maintaining high-quality data. Additionally, evaluating how AC interacts with sentiment, engagement metrics, and user trustworthiness could further enhance our understanding of how concreteness contributes to effective communication in online discussions. Another direction worth exploring is how AC levels correlate with engagement metrics such as comment popularity, upvotes, and response rates, which could provide additional insights into the impact of concreteness on online interactions.

Finally, the study highlights the need for a broader discussion on how concreteness influences communication effectiveness across different domains. Future research could examine its role in persuasive writing, policymaking, and legal discourse, where clarity and specificity are crucial for effective information dissemination. By further refining computational methods for AC estimation, this research contributes to the broader field of natural language processing, fostering more structured and meaningful interactions in digital communication environments. Additionally, practical implementations of AC scoring in education, journalism, and content moderation could be explored to create user-friendly tools that assist in generating more effective and engaging textual communication. Such advancements would reinforce the importance of concreteness in knowledge dissemination and digital interaction, paving the way for further research on optimizing communication strategies through computational analysis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research Background

Online Forum is a network-based interactive platform that allows users to engage in discussions and share information around specific topics. It is a user-generated content space, primarily composed of topic-based threaded discussions, where each thread contains user posts centered around a particular topic, arranged in chronological or logical order [16]. With the development of the internet, online forums have made it possible to address social issues through collective intelligence, removing limitations of time and place [14]. An analysis showed that during 2020, as COVID-19 spread, the number of posts on large asynchronous online forums like Reddit [59] related to "depression, anxiety, and medication" significantly increased, and posts on topics related to "social relationships and friendships" continued to grow [1].

Reaching consensus on solutions to social problems through online discussions is promising [15]. Since forums are conversational social network spaces, the quality of user contributions varies greatly [18]. Navigating this knowledge base to find useful information can be challenging and time-consuming. Some key online discussion forums have already used collaborative intelligence to highlight noteworthy posts. For instance, most forums allow users to rate posts on a five-point scale (1 being the lowest, 5 the highest), and some forums offer more granular rating systems [2]. These ratings help filter online forum content based on the value of posts, enabling users to access knowledge more easily. However, a substantial part of the conversation may have occurred in threaded discussions before users identify valuable posts. At this point, the comments within posts significantly influence the perceived value of the posts. These factors collectively affect the visibility of knowledge within online discussion forums.

In recent years, the evaluation of text quality and popularity has been studied across various domains [2]. Some research has focused on text comprehensibility, defined as "the ease of understanding," which is a crucial factor in document usability. The gap between the text and the reader (e.g., measured by school grade levels) determines whether a text appears to be read. For skilled and educated readers, texts on complex topics such as science, philosophy, or legal issues may be easily understood; however, these texts pose considerable cognitive burdens for a significant portion of readers.

While comprehensibility depends on various factors, such as syntactic difficulty measured by surface text features (e.g., sentence or word length) or document coherence, we focus on concreteness, a key aspect of content comprehensibility. Concreteness refers to the extent to which text includes detailed and vivid information. It contrasts abstraction, which generalizes or omits detailed information to summarize broader concepts. Concreteness often enhances the reader's ability to understand and connect with the text, as it appeals directly to sensory experiences and memory [10]. In contrast, while useful for summarizing complex ideas, abstraction can sometimes make the text less engaging or harder to comprehend [5][20]. Thus, emphasizing concreteness is essential for creating effective and impactful communication. Studying how to predict the level of concreteness in text is highly significant, as it provides insights into enhancing text comprehensibility and ensuring effective communication across different contexts.

Concreteness not only affects comprehensibility but also directly influences users' interest and attitudes toward the text [21]. Readers may find texts filled with excessive generalizations and abstractions to be dull, confusing, or vague. Concrete content tends to impact readers more than abstract, generalized content because it engages their sensory experience and memory. By reading, these memories almost allow them to "feel, see, hear, touch, smell, and taste" the content [4]. A good writing style should capture readers' attention and stimulate their senses using many concrete words.

Additionally, interest and attitudes toward information appear to be directly linked to memory, as human memory is believed to be emotionally driven [5]. Consider the following situation: when teaching employees workplace safety, simply stating, "Accidents can happen if you’re careless," will be less effective than sharing specific examples, such as "Wearing loose clothing near machinery can result in it getting caught, leading to injuries." Humans learn by generalizing from specific cases and applying the knowledge to analogous situations, but abstract warnings alone often lead to poor understanding and retention.

This study seeks to address two key issues: First, while previous research has extensively focused on sentence-level or word-level concreteness using dictionaries or word embeddings [52][54] , there is a lack of datasets that directly reflect human perceptions of the concreteness of entire forum comments. Second, the influence of various dimensions, such as clarity, specificity, and actionability, on the perceived concreteness of comments remains underexplored. By addressing these gaps, this study aims to construct a dataset and develop a model to evaluate comment-level concreteness, which is mentioned as Abstraction-Concreteness (AC) score, and its influencing factors.

1.2 Research Objective

The objective of this research is to construct a model to estimate the concreteness (AC score) of comment texts in online discussions. The following research questions are proposed to achieve this objective:

(1) How can we identify and investigate potential factors influencing the AC score of comment texts?

(2) How can data on comment texts and their AC score be effectively collected and annotated?

(3) How can models be built and evaluated to predict AC score based on the collected data?

This research makes three key contributions: 1. We propose a novel research problem: assessing the concreteness of comment texts to support quality detection in online discussions. 2. We collect data on concreteness scores and analyze the potential factors influencing them. 3. We develop and evaluate models to estimate concreteness based on these factors, providing insights into model performance and limitations.

1.3 Thesis Structure

The structure of this thesis is as follows:

Chapter 1: Introduction - Provides the research background, research objective, and an overview of the thesis structure.

Chapter 2: Related Work - Discusses previous research related to this research.

Chapter 3: Proposed Model - Describes the methodology used in this research, including data preparation, model development, and the evaluation protocol.

Chapter 4: Experimentation and Evaluation - Presents the experimental setup, evaluation metrics, and results of the proposed approach.

Chapter 5: Conclusion - Summarizes the findings and future work.

Chapter 2

Related Work

2.1 Online Discussions

Online discussions are interactive exchanges within forums or platforms where users post and respond to messages on specific topics in text [22][23]. These discussions are typically structured into threads, with each thread consisting of:

Initial Post: A question, opinion, or topic introduced by a user to start the discussion.

Replies: Comments to the initial post or to other replies, forming a conversational hierarchy.

Metadata: Information such as timestamps, user identifiers, and engagement metrics (e.g., likes, upvotes).

Figure 2.1: A part of an online discussion on Reddit.com[24]

2.2 Text Quality Assessment

The study of text quality has evolved significantly since 1944, starting with Robert Gunning's consultancy work, which defined text quality as factors that make writing fluent and easy to read [27]. Early efforts (1944–1970s) focused on readability metrics like the Gunning Fog Index and Flesch-Kincaid, linking text complexity to reader comprehension through word and sentence length analysis [27][30][31]. While these metrics provided simple measures for evaluating text readability, they largely focused on surface features.

The Gunning Fog Index is a readability metric developed by Robert Gunning to assess how easy a text is to read [27]. It estimates the education level needed to understand the text on a first reading.

$$\text{Fog Index} = 0.4 \times \left(\frac{\text{Total Words}}{\text{Sentences}} + \text{Percentage of Complex Words} \times 100 \right) \quad (1)$$

Where:

Complex Words: Words with three or more syllables, excluding proper nouns, compound words, or simple verb forms.

The Fog Index considers two key factors: the average sentence length and the percentage of complex words, which are words with three or more syllables excluding proper nouns, compound words, and simple verb forms. Texts with a Fog Index between 8 and 10 are considered easy to read and suitable for a general audience, such as newspapers. At the same time, scores above 16 indicate very complex texts, suitable for advanced readers [30]. Although widely used in journalism, business, and technical writing, the Fog Index has limitations, as it focuses only on surface features like word length and sentence structure, ignoring deeper aspects like context and logic.

From the 1970s to the 1990s, linguists such as Halliday, Hasan, Mann, and Thompson expanded this foundation by exploring cohesion and rhetorical structures [25]. Their work introduced concepts like cohesion devices and Rhetorical Structure Theory (RST), which analyzed how logical connections between text segments improve coherence and guide readers' interpretations [32] [33]. These frameworks significantly influenced fields like computational linguistics and education, laying the groundwork for modern discourse analysis.

Since 1995, research on entity coherence and discourse analysis has emphasized the

importance of maintaining a logical flow of topics through consistent use of entities [35]. Centering Theory [34] formalized this concept, proposing that coherent texts guide readers by smoothly transitioning between key entities. These frameworks significantly improved understanding of text coherence, influencing fields like automated scoring, summarization, and natural language processing (NLP) tasks.

2000s-Present: Advances in machine learning enabled the integration of lexical, syntactic, and discourse features into readability models [36]. Modern approaches, leveraging large corpora and deep learning like BERT, shifted focus from surface metrics to more sophisticated analyses of text complexity, enabling precise evaluations applicable to education, journalism, and personalized learning systems [37]. In the following, we will explore several recent studies on text quality to illustrate the latest advancements in this field.

2.2.1. A Unified Framework for Predicting Text Quality

Pilter's study focuses on developing a readability assessment model that combines lexical, syntactic, and discourse features. This model advances beyond traditional metrics like the Flesch-Kincaid Index or Gunning Fog Index, which rely primarily on surface-level features [25]. By incorporating deeper linguistic structures, such as discourse relations and syntactic complexity, this study provides a comprehensive understanding of how different textual properties interact to influence perceived readability.

Traditional readability metrics have been widely used but are limited in their ability to predict human judgments of text quality accurately. These metrics generally focus on surface features such as sentence length, word syllables, and overall text length, but they fail to capture syntactic and semantic relationships between sentences. Earlier works have attempted to address this by employing more sophisticated features. For instance, Si [38] and Schwarm [39] integrated language models to predict readability based on vocabulary likelihood and syntactic structures, providing more robust predictions for texts aimed at specific grade levels. Additionally, studies by Barzilay [40] explored entity coherence, emphasizing the role of consistent topic development across sentences in maintaining readability. These approaches, however, often treat features independently, overlooking the interplay between lexical, syntactic, and discourse-level factors.

The contribution of Pilter's study lies in its novel integration of discourse features using the Penn Discourse Treebank (PDTB), which annotates explicit and implicit discourse relations, including expansion, contingency, and temporal relationships. The authors empirically demonstrate that discourse relations, alongside vocabulary and syntactic features, are one of the most predictive factors for readability. For instance, the likelihood

of specific discourse relations and the number of verb phrases per sentence strongly correlate with human judgments of text quality. This finding highlights the need to move beyond simple metrics and incorporate a multi-dimensional approach to readability assessment.

Pilter's study evaluates its model on Wall Street Journal articles, focusing on readability rankings provided by college-educated readers. Regression analysis and pairwise ranking experiments show that combining features like lexical likelihood, discourse relations, and syntactic complexity achieves superior performance compared to models using only surface features. Notably, the combination of entity coherence and discourse relations produces the best results, achieving high predictive accuracy.

Table 2.1: SVM prediction accuracy in Pilter's work[25]

Features	Accuracy
None (Majority Class)	50.21%
ALL	88.88%
log_l_discourse_rels	77.77%
number_discourse_rels	74.07%
N-O transition	70.78%
O-N transition	69.95%
Avg_VPs_sen	69.54%
log_l_NEWS	66.25%
number_of_words	65.84%
Grid only	79.42%
Discourse only	77.36%
Syntax only	74.07%
Vocab only	66.66%
Length only	65.84%
Cohesion only	64.60%
no cohesion	89.30%
no vocab	88.88%
no length	88.47%
no discourse	88.06%
no grid	84.36%
no syntax	82.71%

This work underscores the importance of integrating lexical, syntactic, and discourse features to capture the complexity of human readability judgments. Its emphasis on discourse relations provides a strong foundation for advancing readability research and aligns with recent computational linguistics trends focusing on multi-dimensional text analysis.

2.2.2. A Neural Local Coherence Model for Text Quality Assessment

Mesgar's study introduces a neural local coherence model designed to assess text quality, emphasizing its application to readability assessment and essay scoring [41]. Unlike traditional entity-based or lexical coherence models that rely on explicit tools like coreference resolution systems, this approach leverages distributional semantic representations and neural networks to capture sentence-to-sentence transitions based on semantic information. The model represents coherence using semantic patterns extracted via a convolutional neural network (CNN) [42] and employs a recurrent neural network (RNN) [43] with Long Short-Term Memory (LSTM) cells to encode word contexts within sentences. By focusing on the two most semantically similar RNN states from adjacent sentences, the model captures salient sentence-level semantic relations and encodes their transitions as coherence vectors.

The study positions its model against previous coherence approaches [44], highlighting limitations such as the dependency on external tools in entity-based models and the lack of context consideration in lexical approaches. It also contrasts its CNN-based pattern extraction with graph-mining methods previously employed for coherence assessment [43]. The proposed model surpasses these limitations by capturing distant word relations and contextual nuances, resulting in a more robust coherence representation.

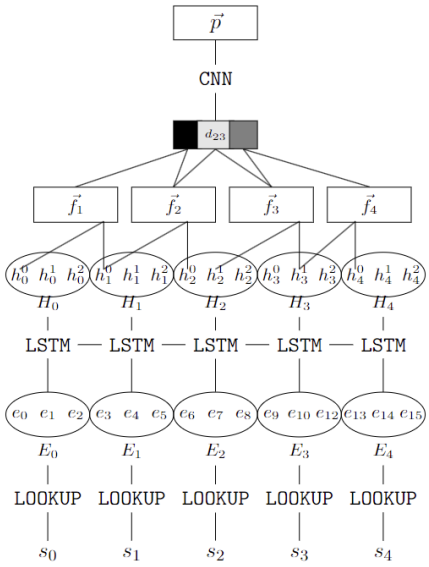

Figure 2.2: Coherence model in Mesgar's work[41]

The model was tested for evaluation on readability assessment and essay scoring tasks. In readability assessment, the model achieved state-of-the-art performance, significantly outperforming previous graph-based coherence models and readability systems. It excelled at ranking text pairs based on their readability, utilizing coherence patterns effectively. In essay scoring, the model's coherence vectors, when combined with linguistic features from existing scoring systems, improved performance, demonstrating the utility of its coherence representations in diverse domains.

The paper acknowledges that while its model significantly enhances coherence assessment, it exclusively focuses on local coherence. This raises potential limitations in evaluating global coherence or overarching text structure, which might be relevant in broader contexts of text quality evaluation. Additionally, the model's reliance on pre-trained embeddings and computationally intensive neural networks may pose scalability challenges in resource-constrained environments.

This work contributes to the broader discourse on text quality assessment by providing a neural coherence framework that integrates semantic pattern recognition with sentence-level context, bridging gaps left by earlier methods. It lays a foundation for future research into coherence patterns and their implications for diverse linguistic tasks.

2.3 Text Concreteness Features

Text concreteness features in this study are derived from the concept of word concreteness—the degree to which a word refers to an entity that can be perceived through our senses, such as something we can see, hear, touch, taste, or smell [10]. Word concreteness has been widely studied in psycholinguistics, where words are rated on how directly they evoke sensory experiences [45]. This research extends the idea of concreteness beyond individual words to entire text segments, such as sentences and paragraphs, capturing how concrete, vivid, and sensory-rich the language is throughout a larger unit of discourse.

2.3.1. Word Concreteness

The study by Brysbaert, Warriner, and Kuperman introduces a comprehensive dataset of concreteness ratings for over 40,000 English word lemmas and nearly 3,000 two-word expressions, collected through a large-scale crowdsourcing effort [10]. Concreteness, defined as the extent to which a word’s meaning refers to perceptible entities, has been

widely studied in psycholinguistics and cognitive science due to its impact on memory, language processing, and comprehension.

The dataset was created using a curated list of 60,099 English words and 2,940 two-word expressions, drawn from multiple sources, including the SUBTLEX-US corpus [46] [47], English Lexicon Project, and the British Lexicon Project [48]. Using Amazon Mechanical Turk (AMT), over 4,000 participants rated words on a 5-point scale, ranging from abstract (e.g., "justice") to concrete (e.g., "apple"). Ratings were based on participants' sensory and experiential understanding of the words. To ensure data quality, the authors implemented rigorous controls, such as the inclusion of calibrator words and checks for participant reliability. After excluding inconsistent responses, the final dataset contained ratings for 37,058 lemmas and 2,896 expressions, all known by at least 85% of raters.

The concreteness ratings strongly correlated with existing norms in the Medical Research Council Psycholinguistic Database (MRC) [49] ($r = 0.92$), validating the reliability of the new dataset. However, the study highlighted a modality bias: participants primarily relied on visual and tactile senses when rating concreteness, with less consideration for auditory or gustatory experiences. Additionally, the dataset revealed a bimodal distribution of concreteness ratings, suggesting that concreteness and abstractness may represent distinct categories rather than a single continuum.

Gregori's study builds on Brysbaert et al.'s work [10] by exploring the contextual nature of word concreteness in both English and Italian, moving beyond static, word-level ratings [50]. As part of the CONCRETEXT task at EVALITA 2020 [51], the authors introduced a dataset of 1,096 sentences (550 in Italian, 534 in English) from WikiHow instructions, where target nouns and verbs were annotated with concreteness scores on a 7-point Likert scale by over 300 native speakers per language. The task challenged participants to predict the concreteness of words in context, using systems that integrated distributional models, BERT as transformer-based embeddings, and behavioral norms. The ANDI system achieved the best performance, highlighting the value of combining contextual and lexical features [52]. While the dataset's small size and reliance on human annotations limit generalizability, the study advances the field by emphasizing how context affects concreteness perception and providing a framework for future NLP tasks, including semantic representation and lexical disambiguation.

The results show that word concreteness is significantly affected by context, with polysemous words exhibiting notable shifts depending on sentence meaning [53]. Contextual variability also revealed subtle cross-linguistic differences between English and Italian due to linguistic and cultural factors. While the dataset is smaller in scope, it

provides high inter-rater reliability (Cronbach's alpha > 0.9) and adds depth to the study of contextual concreteness.

2.3.2. Concreteness for Document Comprehensibility

Tanaka's study estimates the concreteness of terms and documents to evaluate and improve document comprehensibility [54]. It addresses the limitations of traditional readability metrics, such as Flesch Reading Ease or Dale-Chall Formula, which primarily rely on syntactic features like sentence length or word syllables, by introducing a concreteness-driven approach to comprehensibility. The research highlights the critical role of concreteness, defined as the ease of perceiving or visualizing concepts, in determining document quality and user satisfaction.

Building on psycholinguistic theories of concreteness [55][56], this study extends prior efforts by automating term- and document-level concreteness estimation using machine learning. It demonstrates the practical value of incorporating concreteness into readability models, offering a novel perspective for applications in information retrieval and personalized search. Focusing on concreteness complements existing readability measures, bridging the gap between user comprehension and content accessibility.

The study introduces methods for estimating term- and document-level concreteness to assess document comprehensibility. At the term level, an SVM regression model with 21 features—such as visual representativeness, sensory verb co-occurrence, ontology depth, and sentiment levels—is used to capture perceptual and imagistic properties. Training data is sourced from the MRC, which provides human-annotated concreteness ratings for over 3,400 nouns. For document-level estimation, two approaches are proposed: averaging the concreteness scores of all terms in a document and identifying the most concrete paragraph. These methods aim to link concreteness with overall text readability.

2.4 Models for Automatic Scoring of Text

Automatic scoring systems have been developed to evaluate diverse types of user-generated content, such as essays and discussion posts. In Wang's study, a reinforcement learning framework optimized scoring accuracy by incorporating rating schemas [8]. This approach demonstrated the benefits of advanced machine learning techniques in addressing subjective evaluation tasks.

In the context of online discussions, Wanas' study proposed a classification system to rate posts as high, medium, or low value based on relevance, originality, forum-specific

traits, and surface features [9]. This work highlighted the importance of combining content-specific and structural features to improve the discoverability of valuable posts. These findings align with the objectives of this research, which aims to construct models that score comments based on their concreteness, an aspect critical for understanding contribution quality in discussions.

Joo's study proposed a model to measure discussion validity using participants' discussion capabilities, highlighting the importance of evaluating contributions based on specific text features [6]. Similarly, in software-related forums, an algorithm leveraging surface, lexical, syntactic, forum-specific, and similarity features achieved high accuracy in assessing post quality. These works demonstrate that feature-based models can effectively evaluate text quality by capturing diverse aspects of user contributions.

Deokgun's study introduced CommentIQ, a system that combines analytic scores and interactive visualizations to assist moderators in identifying high-quality comments [7]. This research underscores the value of integrating scoring systems to enhance online interactions and highlights the potential for similar methodologies to be applied to assess other dimensions of text quality, such as concreteness.

2.5 Research Positioning

Building on the insights from these related works, this study aims to address a critical gap in evaluating online discussion texts by focusing on the concreteness of comment contributions. Unlike prior studies emphasizing overall text quality, this research isolates concreteness as a measurable and impactful dimension, leveraging feature-based models and advanced scoring techniques. By integrating social interaction, surface, and content features, this study seeks to develop robust models that enhance the understanding and visibility of high-quality contributions in online discussions.

Chapter 3

Proposed Model

3.1 Method Overview

This study follows a structured workflow to assess and predict text concreteness. First, raw comment text data is collected from online discussion forums and removing irrelevant entries. Next, key linguistic and contextual features are identified and defined. The dataset is then annotated through a crowdsource platform, where workers annotate data on text concreteness and other features. Following this, the data undergoes analysis and parsing to check for consistency and inter-annotator agreement. Models are constructed and evaluated to predict concreteness. Finally, evaluation methodologies are established to assess model performance and determine the most influential factors in concreteness prediction.

This study will proceed as Figure 3.1:

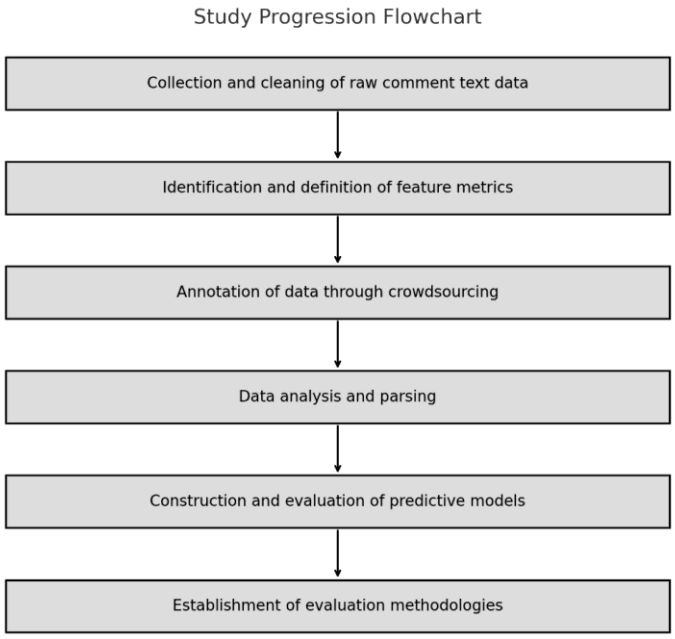

Figure 3.1: Study progression flowchart

3.2 Data Collection

1. Reddit

Reddit [59] is one of the largest online platforms for community discussions, hosting millions of active users who engage in various topics through posts and comments [57]. Its structure, organized into topic-specific subreddits, fosters highly interactive and hierarchical discussions. These threaded discussions, where replies are linked to specific parent posts, enable researchers to analyze conversational dynamics and user interactions comprehensively. The platform's diversity ensures access to comments with varying levels of concreteness, from abstract musings to detailed, example-driven explanations. For this research, Reddit's dynamic and global user base provides authentic, real-world data crucial for studying how comments reflect and influence text concreteness.

2. Subreddit: ExplainLikeImFive

ExplainLikeImFive (ELI5) is a popular educational subreddit where users ask and answer questions in a simple and accessible manner, often as if explaining to a five-year-old [58]. The subreddit encourages detailed yet comprehensible responses, making it a valuable source of high-quality, concrete content. Comments in ELI5 are often upvoted based on their ability to simplify complex topics. This unique environment makes ELI5 an ideal dataset for studying comment concreteness and assessing the quality of explanations in online discussions.

3. Reddit API

The Reddit API [60] provides programmatic access to Reddit's vast data resources, enabling efficient collection of both textual content and metadata. This section provides detailed information on the API configuration, the structure of the scraping script, and the measures taken to ensure compliance with Reddit's guidelines.

Credentials were generated through Reddit's developer portal to access the Reddit API. These credentials included:

1) Client ID and Client Secret: Unique identifiers to authenticate API requests.
2) User Agent: A descriptive string (e.g., "Bot/1.0") identifying the application making requests.
3) OAuth Tokens: These are used for secure, authenticated API interactions. The script employed the `PRAW` library in Python [71] to simplify token management and ensure authenticated access.

Posts were retrieved from the ELI5 subreddit using Reddit's `hot` algorithm, which ranks posts based on a combination of factors such as upvotes, age, and user engagement. This algorithm prioritizes posts that are both popular and recent, ensuring the dataset reflects current and actively discussed topics.

The Reddit API enforces a rate limit of 60 requests per minute. To comply with this restriction, the script incorporated a delay of 1000 milliseconds (1 second) after processing each post. This ensured uninterrupted data collection without violating API restrictions.

As an initial filtering step, 50 posts were selected, each with at least 10 comments to ensure substantive discussions. For each post, all nested comments were fully expanded and collected resulting in a total of 3,073 comments. The collected data was initially stored in JSON format for flexibility and later converted to CSV for analysis.

Table 3.1: Raw data format

Field	Explanation	Example
Post_ID	A unique identifier for each post in Reddit	1i2wv5u
Post_Body	The main content or body of the post	ELI5 is it true that the way burned fat actually leaves your body is when you exhale co2?
Comment_ID	A unique identifier for each comment within a post	m7ixv20
Comment_Text	The content of the comment	All of your skin has tiny pores that help you lose water, you just don't notice it because it's in over 2square meters of skin. It's about 12.5ml/h if water in the same space of a king size bed You notice your own sweat only when you massively increase your water elimination in order to keep the body from overheating (aka sweating)
Parent_ID	The identifier of the parent object to which the comment is responding	t1_m7iwqp1
Depth	Indicates the hierarchical level of the comment within the thread	0

4. Raw Data Format

The collected data was structured in CSV format for compatibility with analysis tools. Key fields as Table 3.1.

The ‘Depth’ field helps visualize and analyze the structure of discussions. Comments at deeper levels are typically part of sub-conversations, such as:

Depth = 0: A comment directly responding to the post.

Depth = 1: A comment responding to another comment.

Depth = 2: A nested reply within the thread.

5. Data Cleaning and Sampling

Data cleaning is a crucial step to ensure the quality and reliability of the dataset. The following processes were implemented for this study:

1) Unified Encoding: All text data was converted to UTF-8 encoding to handle non-ASCII characters and ensure compatibility with analytical tools.
2) Empty Content Removal: Comments where `Comment_Text` was null or consisted only of whitespace were removed, as they do not provide meaningful information.
3) Random Sampling: Using Python’s `random` library, 120 comments were randomly selected from the cleaned dataset.
4) Manual Filtering: During manual review, 3 comments that contained only URLs or were unreadable were removed. After this process, 100 comments were randomly selected to form the final dataset for analysis.

3.3 Feature Definitions

In this study, the core focus is on the AC score, which measures how abstract or concrete a comment is. Since concreteness is a subjective perception that varies between individuals, this study employs a Likert scale [73] to systematically quantify human judgments.

A Likert scale is a widely used psychometric tool for measuring people's attitudes or perceptions by presenting them with a statement and allowing them to express their level of agreement on a predefined scale [72]. In the case of AC, annotators were presented with the statement:

"This text is concrete."

They were then asked to rate their agreement with the statement using a 5-point Likert scale, where:

5 = Strongly Agree (The text is highly concrete.)

4 = Agree (The text is fairly concrete.)

3 = Neutral (The text is neither particularly concrete nor abstract.)

2 = Disagree (The text is fairly abstract.)

1 = Strongly Disagree (The text is highly abstract.)

This section explains the core features and the auxiliary dimensions that influence AC. The same Likert-scale approach is applied to the auxiliary dimensions, including Clarity, Specificity, Relevance, Actionability, and Orientation, ensuring consistency in measuring different aspects of text concreteness.

3.3.1. Core Feature: AC

The AC score reflects people's intuitive sense of whether a text feels abstract (e.g., generalized ideas, conceptual language) or concrete (e.g., specific examples, tangible information) as a whole. Building on Brysbaert's research on word concreteness, which measures the degree to which words are associated with perceivable entities [10], this study extends the concept to entire text segments. Specifically, it explores how people perceive the abstraction or concreteness of an entire passage by linking the textual content to sensory and tangible experiences in the real world.

Table 3.2: Examples for AC

High Level	because thousands of vehicles already set times there. let's build an almost identical circuit somewhere. why would a car manufacturer go and test their car there? what's the purpose? there is no previous data to compare. when you test your car at Nurburgring, you can easily say "our car is faster than X but slower than Y"
Low Level	Unless you're building an ultra-high performance track car, an oval pipe isn't going to make any noticeable difference. Many manufacturers squish exhaust pipes into all kinds of irregular shapes to tuck them up tightly against the vehicle and snake them around other components. One little oval section isn't going to have a tangible effect on performance.

3.3.2. Dimensions Influencing AC

1) Clarity

Clarity measures how understandable and unambiguous a comment is. Clarity reflects the degree to which the language used in a text is understandable, free from ambiguity, and easy to interpret. In studies evaluating human-generated texts, clarity has been

strongly associated with comprehensibility and readability, directly impacting a reader's ability to process information efficiently without confusion [61]. Text with both high clarity and concreteness may be like below:

Table 3.3: Examples for Clarity

High Level	My pop kept the glow plug from his 1997 GMC Sierra and moved it to his 2015 F-150, which didn't come with one. Still works.
Low Level	Lots and lots and lots of worthless footage and even more patience.

2) Specificity

Specificity measures the richness and depth of detail in a text. Highly specific comments provide precise information, such as numerical data, concrete examples, or descriptive elements. In contrast, vague or superficial statements tend to feel abstract. Specificity enhances text concreteness by grounding the content in vivid, relatable details, making it easier for readers to process and understand the message.

Specificity and Concreteness are often discussed together in linguistic studies, but they are not synonymous [62]. A text can be highly specific while remaining abstract, containing numerous details but lacking sensory or tangible elements.

> *"Quantum entanglement occurs when two or more particles become interconnected such that their quantum states are instantaneously correlated, regardless of the distance separating them. This phenomenon, mathematically described by Bell's Theorem, challenges classical notions of locality and is a fundamental principle underlying quantum computing."*

This passage contains detailed, specific information about quantum entanglement (high specificity), but it lacks sensory or physical references, making it abstract and difficult for a general audience to visualize (low concreteness).

Despite their differences, specificity and concreteness are often correlated. Highly specific descriptions incorporating concrete details tend to be more comprehensible and engaging.

Table 3.4: Examples for Specificity

High Level	When any gas - including air or refrigerant - is squeezed, it's temperature increases, it becomes warmer. When any gas - including air or refrigerant - which had been squeezed is allowed to expand, it's temperature decreases, it becomes cooler.

	Inside of a refrigerator... * refrigerant gas is squeezed, making it hotter * the hot refrigerant moves through the inside of a tube cooled by outside air * the less hot refrigerant passes through an expansion valve * the now cold refrigerant passes through another tube, this one inside of the refrigerator. * the cold refrigerant makes this inside tube cold, which makes the air inside of the refrigerant cold too. The part which squeezes the refrigerant is called a compressor. The tubes are usually shaped into a spiral to take less space. The refrigerant usually changes between a gas and a liquid and back, because that allows the refrigerator to use less tubing total.
Low Level	Everywhere that anything happens. Life happens by exponential cellular division. The way you're phrasing the question is weird because it's like you're asking when ducks whip out their calculators but math is a language for expressing real-life changes and trying out hypotheses without doing an actual experiment. Asking 'when does it happen in nature' doesn't make sense any more than asking what language blackberries would speak if they could.

3) Relevance

Relevance assesses how well a comment aligns with the topic or question posed in the post. Unlike other forms of text, comments in online discussions are inherently context-dependent, meaning they derive meaning and purpose from the original post to which they respond. Based on observations, comments that build directly upon the topic of the post tend to be more concrete, as they are grounded in a defined subject and often provide relevant details or explanations. Thus, Relevance and Concreteness are interrelated—a highly relevant comment is more likely to provide specific and tangible details, reinforcing its concreteness. Conversely, off-topic or tangential remarks often lack direct references to the discussion, making them feel more abstract and less structured.

Table 3.5: Examples for Relevance

High Level	There is NO disease-modifying treatment to slow the spread of alpha-synuclein proteins in Parkinson's (and related Lewy Dementia, which is in my genes) among tens of million patients worldwide who have died since the disease was named in 1817. Drug developers have been mystified. That will change someday considering we're finally starting to see breakthroughs in Alzheimer's anti amyloid-protein treatments. In the meantime, Parkinson's symptom

	progression can be slowed down to some extent with dopamine drugs and some high-intensity exercises (e.g. boxing, "forced cycling").
Low Level	Krebs Cycle is the chemical basis for most life.

4) Actionability

Actionability evaluates whether a comment provides clear, actionable advice or steps that the reader can follow. Texts with high actionability typically include practical guidance or instructions that readers can implement immediately. In contrast, abstract or theoretical texts may describe ideas without offering concrete steps for application.

Research in sentiment analysis has suggested that Actionability—or the practicality of a text—is often associated with the text's attitude, such as approval or disapproval [63]. Texts with higher actionability tend to convey stronger, more explicit attitudes, which enhances the accuracy of sentiment recognition. For example, a text offering step-by-step advice on solving a problem is more likely to be aligned with a specific sentiment (e.g., positive, supportive) than a generalized or abstract statement.

Table 3.6: Examples for Actionability

High Level	I broke my humerus recently, and the doctor told me if I hadn't been wearing so many layers, the bone would have been more free to move and could have cut my brachial artery and I'd have been dead before the ambulance got to me.
Low Level	My pop kept the glow plug from his 1997 GMC Sierra and moved it to his 2015 F-150, which didn't come with one. Still works.

5) Orientation

Orientation considers whether a text leads toward a clear outcome or conclusion. Orientation can have multiple meanings. In this study, orientation specifically refers to the degree to which a comment is directed toward a particular conclusion or result. A strongly oriented comment provides a clear sense of direction, guiding the reader toward a defined takeaway. In contrast, a weakly oriented comment may feel open-ended, ambiguous, or lacking a clear resolution.

Studies on online discussions have not yet extensively examined orientation in comment texts, making it an underexplored dimension. Based on observations, comments with strong orientation explicitly connect their points to a tangible or conceptual goal, whereas abstract ones may lack direction or resolution. Orientation plays a key role in enhancing text concreteness, as comments that lead to clear conclusions often provide

structured and goal-driven information.

Table 3.7: Examples for Orientation

High Level	Not necessarily broken bones, but lots of tissue trauma from a nasty fall or crash can cause proteins and electrolytes from the muscles and skin to leak out into the blood stream. This can cause rhabdo, electrolyte imbalances that can cause arrhythmias, and so on.
Low Level	One, there's like no market (allegedly). Two, it's hard to make salty/bitter/savory flavors that are all the same strength and will last the exact same amount of time.

3.4 Annotating Data

Psycholinguistic studies have established numerous lexical dictionaries for English words, capturing human perceptions of word concreteness[55][10]. These dictionaries serve as critical references for concreteness research. Recent studies have explored using deep learning methods to expand these dictionaries and to predict sentence concreteness by averaging word-level concreteness scores[50][51][52][53]. However, datasets directly reflecting human perceptions of entire comments' concreteness are absent. To address this gap, this study aims to construct a dataset that encapsulates human perspectives on comment-level concreteness, providing a standard for evaluating the impact of influencing dimensions.

3.4.1. Amazon MTurk

1. Overview

Amazon Mechanical Turk (MTurk) is a crowdsourcing platform developed by Amazon [64]. It allows requesters to publish tasks, referred to as Human Intelligence Tasks (HITs), which workers can complete for compensation. This platform enables scalable data collection and annotation by leveraging a diverse workforce.

MTurk's name originates from an 18th-century hoax machine, "The Turk," which was presented as an automated chess-playing device but was actually operated by a hidden human chess master. Similarly, MTurk emphasizes human involvement in completing tasks challenging for machines to perform autonomously.

In psycholinguistic research, MTurk has been widely used as a reliable tool for collecting data [10]. Studies have shown that data collected through MTurk has high validity and is comparable in quality to data collected from traditional lab-based experiments. This makes MTurk a valuable platform for tasks requiring subjective judgment, such as rating the concreteness of textual content, which is central to this study.

2. Workers

As of 2019, over 250,000 individuals had completed at least one task on MTurk, with approximately 85,000 active workers [64]. MTurk workers come from diverse geographic locations, offering various perspectives. However, to ensure that most annotators were native English speakers, this study restricted participation to workers from English-speaking regions. This filtering step helped maintain annotation quality and linguistic consistency across responses.

MTurk workers are typically younger and more internet-savvy than the general population, making them particularly suitable for digital annotation tasks. Studies indicate that most workers on the platform are non-specialist part-time participants. Given that this study aims to capture public perception of text concreteness and related dimensions, such a workforce composition aligns well with the research objectives.

While MTurk provides a cost-effective and scalable means of data collection, selection biases should be considered when designing questionnaires, as demographic factors such as income and education may influence responses. To mitigate potential biases, task instructions were designed to be clear and accessible to a general audience, ensuring that responses reflected an intuitive and representative evaluation of text features. MTurk also provides a qualification metric called approval rate, allowing requesters to filter workers based on past performance. The approval rate represents the percentage of HITs a worker has submitted that have been accepted by requesters over a given period (typically one month). In this study, the approval rate was set at a minimum of 90%, meaning that only workers with a record of at least 90% approved HITs were eligible to participate.

3. HITs and Assignments

A Human Intelligence Task (HIT) represents a single task or unit of work published on MTurk. For example, labeling ten images with categories such as "dog" or "cat" would constitute a single HIT. Assignments refer to individual responses to a task; multiple workers can complete the same HIT to ensure reliability. For this study, each Reddit comment was rated by three workers to mitigate individual biases and enhance annotation

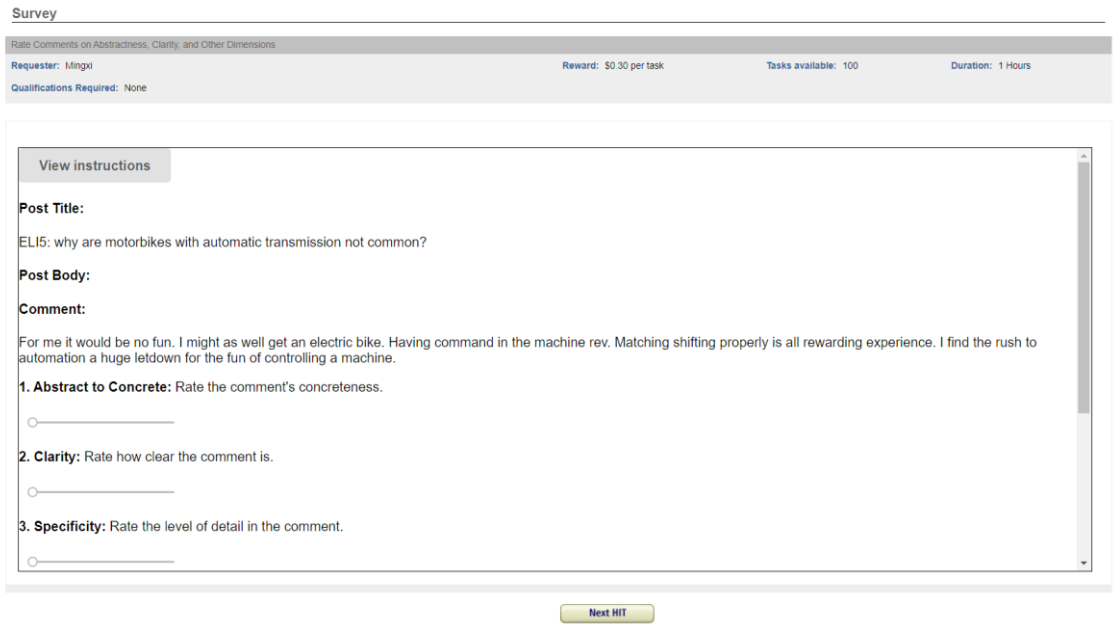

Figure 3.2: A HIT sample in this survey

consistency.

4. Compensation

Workers were compensated at rates above the platform average to attract reliable participants and ensure task quality. Compensation rates were calculated based on task complexity and estimated time requirements, with bonuses provided for exceptional performance.

3.4.2. Survey Design

This study utilized MTurk's survey templates with custom designs tailored for the annotation task. Key components included:

1) Title: "Rate English Texts on Abstractness and Other Dimensions (~1 min)". This title briefly described the task, indicating that participants would rate English texts across several dimensions. It also highlighted the short time requirement for each task, allowing workers to gauge the effort and reward ratio.
2) Description: "In this task, you will rate comments based on their abstractness, clarity, specificity, relevance, actionability, and orientation. Each comment requires evaluating six dimensions using a provided scale. Strong proficiency in English is required." This description provided more detail about the survey, helping workers understand the nature of the task before choosing to participate.
3) Keywords: "text, comments, rating, English, language analysis, abstractness, concreteness." These keywords were selected to attract workers interested in linguistic and text analysis tasks.
4) Reward per Response: $0.2, set as fair compensation for a task requiring less than one minute to complete.
5) Number of Respondents: 3. Each comment was annotated by three workers to ensure reliability and mitigate bias.
6) Time Allotted per Worker: 2 hours. The time allocated for each HIT was determined based on the complexity of the annotation task.
7) Survey Expiration: 14 days. Deadlines were set to ensure timely data collection.
8) Auto-approve and Pay Workers in: 7 days. Payments were automated to process after a short review period.

As a result of this study's data collection process, a total of 300 annotation entries were gathered, ensuring a diverse set of evaluations for analysis.

In MTurk, tasks are designed using an HTML editor that supports customization

through HTML, CSS, and JavaScript. Crowd HTML Elements can be used for simpler layout creation. This layout is shared across all tasks in the project, allowing for consistency and scalability.

1) Key features of the layout design include:
2) HTML and JavaScript Integration: These elements allow for creating of interactive and dynamic task interfaces tailored to specific research needs.
3) Variable Definition: Variables (e.g., ${variable_name}) can be embedded within the layout, enabling data from a CSV input file to dynamically populate the task interface. This ensures that each HIT is unique while adhering to the overall task format.
4) Customization for Task Clarity: Layouts were designed to be intuitive and user-friendly, minimizing the cognitive load on workers and ensuring a clear understanding of annotation requirements.

A survey questionnaire in this study is composed of the following sections:

1) Introductions Button: A clickable element that opens a window providing workers with detailed task instructions, including:
   Summary: A brief overview of the task.
   Detailed Introductions: Comprehensive guidelines for completing the task.
   Examples: Positive and negative examples to clarify the expectations.
2) English Text Information: Displays dynamic content, including:
   ${Post_Title}: The title of the Reddit post.
   ${Post_Body}: The body of the Reddit post.
   ${Comment_Text}: The text of the comment to be rated.
3) Rating Sliders: Six sliders corresponding to the dimensions of abstractness, clarity, specificity, relevance, actionability, and orientation. Each slider allows workers to select a value between {1, 2, 3, 4, 5}.
4) Submit Button: The submit button is enabled only after all six sliders have been assigned values. This ensures the completeness of responses before submission.

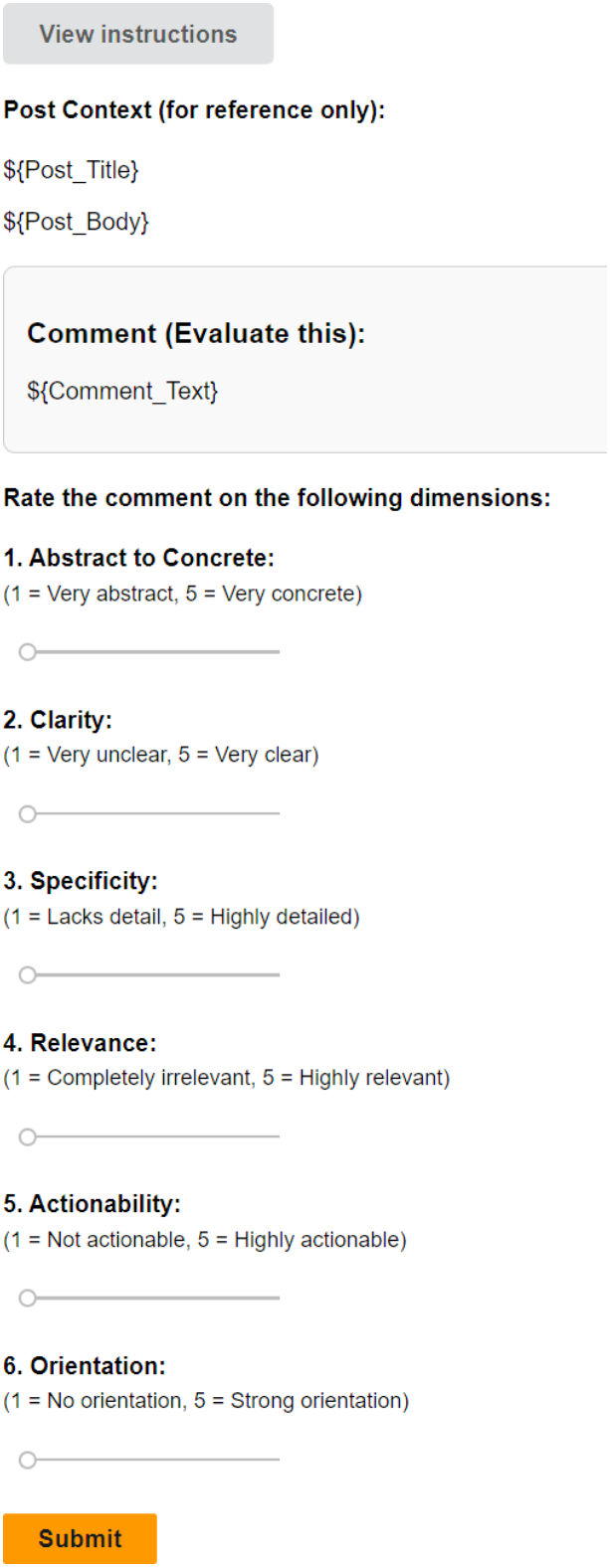

Figure 3.3: Survey layout

3.5 Data Analysis

After receiving all submissions from the crowdsourcing task, the annotated data was downloaded from MTurk in CSV format. The file included the following fields as Table 3.8:

Table 3.8: Raw data structure

HITId
WorkerId
Input.Post_Title
Input.Post_Body
Input.Comment_ID
Input.Comment_Text
Answer.Abstract_Concrete
Answer.Actionability
Answer.Clarity
Answer.Orientation
Answer.Relevance
Answer.Specificity

This structured dataset was used for subsequent statistical analysis, consistency checks, and dataset refinement for model training and evaluation.

3.5.1. Basic Statistical Analysis

This section aims to understand the overall trends in the data for the six dimensions of ratings (AC, Clarity, Specificity, Relevance, Actionability, and Orientation). By analyzing summary statistics and visualizing score distributions, we aim to: 1. Detect potential concentration of scores at high or low values. 2. Identify skewness, kurtosis, or outliers in the distributions. 3.Assess whether the dataset meets the assumptions for subsequent modeling and analysis.

The dataset comprises 300 individual rating instances, generated by 100 unique samples. A Comment_ID uniquely identifies each sample. Three randomly assigned workers rated each sample. Workers were not necessarily involved in rating multiple samples, meaning the assignments were randomized across the dataset. In total, 268 unique workers participated in the rating process. The dataset includes 31 distinct posts,

and a total of 100 comments were evaluated. Since three workers rated each comment, the total number of ratings in the dataset is 300, reflecting the structured evaluation process.

Table 3.9: Summary statistics

	Worker	Post	Comment	Rating
Count	268	31	100	300

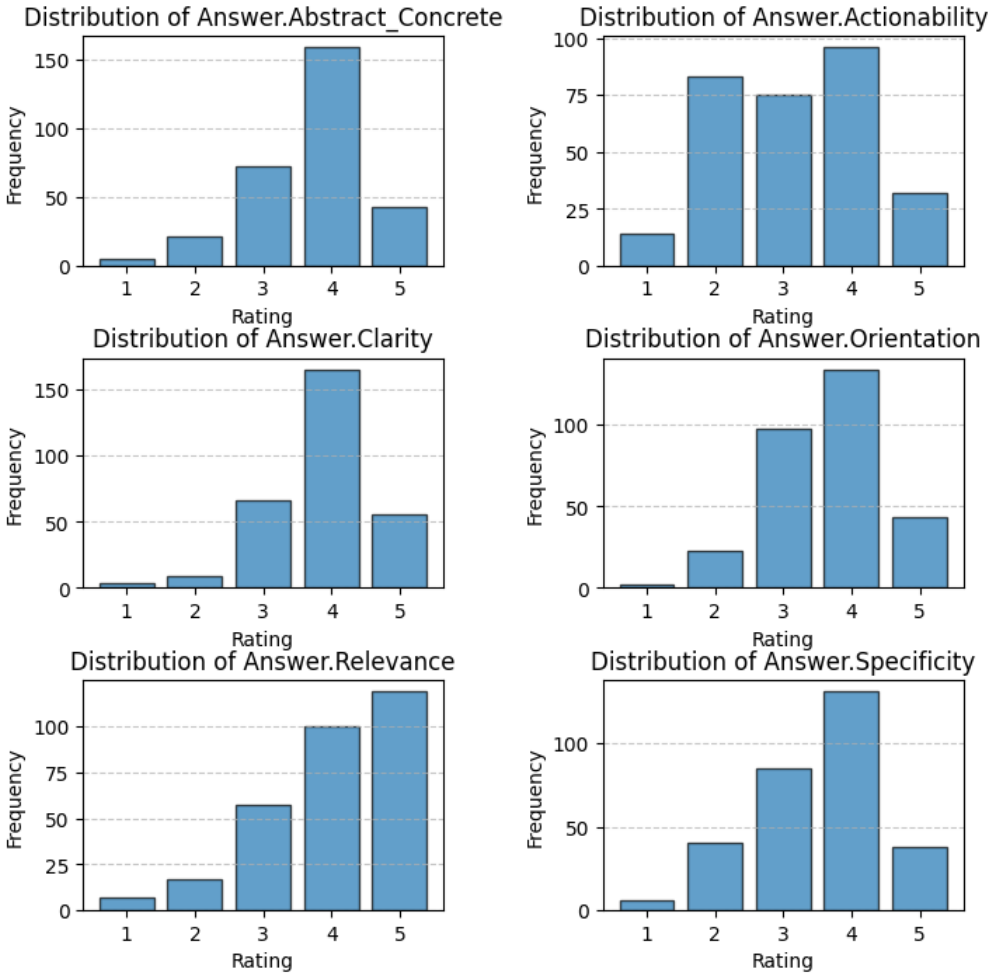

Figure 3.4: Distribution of ratings

Generally, most ratings are concentrated between 3 and 4, with relatively fewer 1 and 2. Higher ratings (4 and 5) appear more frequently in certain dimensions, suggesting that

the nature of the comments may inherently lead to favorable evaluations in those aspects. Fewer low ratings (1 and 2) may indicate that either the dataset contains predominantly well-formed comments or that annotators tend to avoid lower ratings due to uncertainty in evaluation criteria.

These trends can be explained by the comment texts originating from a scientific Q&A setting (Subreddit ELI5), where information is generally structured, relevant, and clear. However, the rating scale may not be well-calibrated for scientific discussion contexts, making it difficult for annotators to use lower ratings consistently.

For each dimension:

1) Mean: The average score across all comments.
2) Standard Deviation: The spread of scores around the mean.
3) Minimum Value: The lowest score observed.
4) Maximum Value: The highest score observed.

Table 3.10: Mean, std of ratings

	count	mean	std	min	25%	50%	75%	max
Answer.Abstract_Concrete	300	3.71	0.86	1.0	3.0	4.0	4.0	5.0
Answer.Actionability	300	3.16	1.09	1.0	2.0	3.0	4.0	5.0
Answer.Clarity	300	3.87	0.79	1.0	3.0	4.0	4.0	5.0
Answer.Orientation	300	3.64	0.84	1.0	3.0	4.0	4.0	5.0
Answer.Relevance	300	4.02	1.01	1.0	3.0	4.0	5.0	5.0
Answer.Specificity	300	3.52	0.94	1.0	3.0	4.0	4.0	5.0

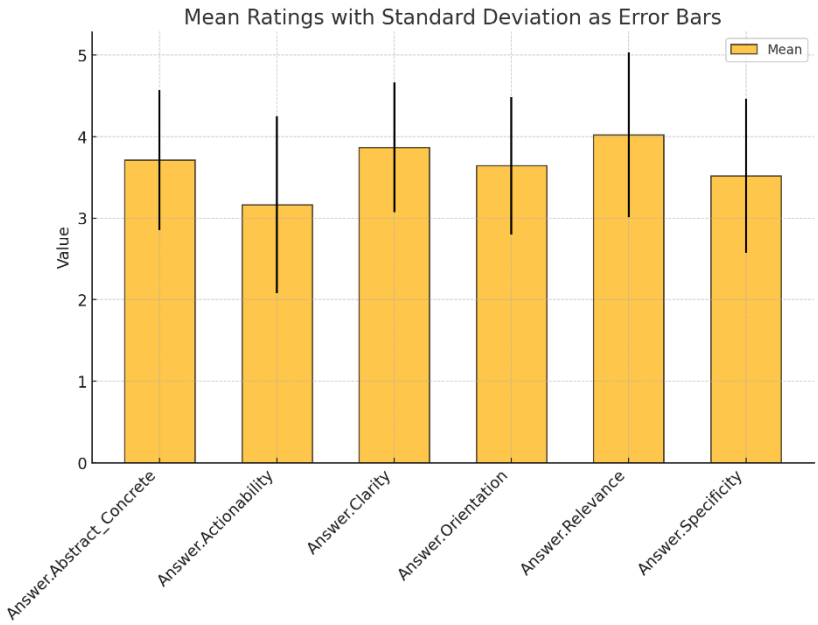

Figure 3.5: Mean, Standard Deviation of Ratings

High mean scores in Clarity and Relevance indicate that comments are generally rated as clear and relevant. A lower mean in Actionability may suggest that many comments are not perceived as actionable.

Actionability and Specificity show greater score variability, implying subjectivity in evaluation. This may suggest that raters interpret these dimensions differently, potentially requiring clearer rating guidelines. Clarity and AC exhibit lower variability, indicating that raters tend to agree more on these dimensions. This suggests a more consistent interpretation of these criteria.

3.5.2. Consistency Analysis

Consistency analysis is a critical step in evaluating the reliability of human-annotated data. In this study, we use the Intraclass Correlation Coefficient (ICC) [65] to measure the agreement among raters across different evaluation dimensions. ICC is particularly useful when multiple raters assess the same set of items, as it quantifies both the degree of absolute agreement and the consistency of ratings.

Unlike other reliability measures (e.g., Cohen's Kappa, Fleiss' Kappa [11]), ICC is

well-suited for continuous ratings (such as our 1-5 scale) and accounts for both inter-rater reliability and within-item variability.

There are multiple types of ICC, each addressing different study designs and reliability concerns. Given that this study is conducted using a crowdsourcing approach, where each comment is rated by a different set of raters selected randomly from a large pool of potential annotators, it is appropriate to use ICC(1, k) for consistency analysis. The nature of crowdsourcing means that raters are not fixed across samples, making random rater selection a key factor in determining reliability. Additionally, the average of three raters is used as the basis for evaluation to reduce data variability. Thus, ICC(1, k) is chosen as the measure of consistency.

The formula for ICC(1, k) is defined as follows:

$$ICC(1,k) = \frac{MS_B - MS_W}{MS_B + \frac{MS_W}{k}} \quad (2)$$

Where:
MS_B= Mean Square Between Groups (Between Comments)
MS_W= Mean Square Within Groups (Between Raters)
k = Number of Raters per comment (3 in our case)

$$MS_B = \frac{\sum_{i=1}^{n} k\,(\bar{X}_i - \bar{X})^2}{n-1} \quad (3)$$

Where:
n = Number of comments
k= Number of raters per comment (3 in our case)
$\bar{X}_i$= Mean rating for comment i
$\bar{X}$= Grand mean across all ratings

$$MS_W = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k}\left(X_{ij} - \bar{X}_i\right)^2}{n(k-1)} \quad (4)$$

Where:
X_{ij}= Score given by rater j for comment i

Interpretation of ICC Values:

0.75 – 1.00: Excellent consistency

0.60 – 0.74: Good consistency

0.40 – 0.59: Moderate consistency

0.00 – 0.39: Poor consistency

Table 3.11: ICC(1, k) of raw data

Dimension	ICC	F	df1	df2	pval	CI95%
AC	0.1150	1.130	99	200	0.234	[-0.23, 0.38]
Actionability	-0.0693	0.935	99	200	0.642	[-0.49, 0.25]
Clarity	-0.1928	0.838	99	200	0.837	[-0.66, 0.16]
Orientation	0.0352	1.036	99	200	0.411	[-0.34, 0.32]
Relevance	0.0462	1.048	99	200	0.385	[-0.33, 0.33]
Specificity	0.1771	1.215	99	200	0.125	[-0.15, 0.42]

Overall, the ICC(1, k) values for all dimensions are quite low, and all p-values are greater than 0.05, indicating weak rating consistency among raters. Additionally, most dimensions have wide confidence intervals that include negative values, further suggesting instability in the ratings.

A possible explanation for these results is that the rating criteria may not have been clearly defined, and raters were not provided with standardized training before assigning their scores. Since this study is an initial experiment, individual preferences and rating randomness were considered, which is reflected in the collected data. The differences in personal interpretations of the rating criteria might have led to substantial variability in scores, reducing overall agreement.

To ensure that the model can accurately predict the collective judgment of a broader audience rather than the subjective preferences of individual raters, the next step is to filter the data to reduce extreme discrepancies among raters and improve the reliability of the consistency analysis. By applying data filtering methods, such as setting thresholds for inter-rater disagreement, it is possible to eliminate outliers and increase the overall stability of the dataset. This process will help refine the dataset so that the remaining ratings better reflect the shared perspectives of the majority, improving the foundation for future predictive modeling.

3.5.3. Data Filtering

To ensure the reliability of the dataset used for modeling, we apply a filtering process to remove ratings with low inter-rater agreement. This step is essential to mitigate the impact of highly inconsistent ratings, which may introduce noise and reduce the robustness of the analysis. By retaining only high-consistency ratings, we aim to improve data quality and enhance the validity of subsequent analyses.

First, it was ensured that each comment received exactly three ratings. Since some comments might have missing ratings due to data collection inconsistencies, only samples with three complete ratings were retained for further analysis, ensuring a stable dataset.

Next, the three ratings for each comment were expanded into separate columns, allowing for a more direct comparison of differences among raters. This restructuring made it easier to assess variations in scoring for the same comment.

A score difference metric was calculated to quantify the level of disagreement among raters. This metric measures the total absolute differences between the three ratings, indicating rating inconsistency. A higher score difference signifies greater disagreement among raters, indicating lower reliability in the given ratings.

The filtering criterion for selecting consistent ratings can be formally expressed as follows:

1. Computing Rating Discrepancy (Δ)

For each comment (per Comment_ID), three independent ratings were collected. Let (S_1, S_2, S_3) be the three ratings given to a comment. To quantify the level of disagreement among raters, Δ was computed based on the absolute differences between all pairs of ratings. The total rating disagreement is computed as:

$$\Delta = |S_1 - S_2| + |S_2 - S_3| + |S_1 - S_3| \quad (5)$$

2. Determining the Optimal Discrepancy Threshold Δ*

To retain enough samples while improving rating consistency, a dynamic filtering approach was adopted. The threshold Δ* for filtering was determined based on the standard deviation (σ) of Δ values in the dataset. Specifically, a set of candidate thresholds was defined as follows:

$$\Delta^* = k \cdot \sigma_\Delta, \quad k \in [1.0, 3.0], \quad \text{step size} = 0.1 \quad (6)$$

where k is a scaling factor controlling the strictness of the filtering process.

For each candidate Δ*, comments with $\Delta > \Delta^*$ were removed, and the resulting dataset was evaluated in terms of:

a) Retention rate (R): The proportion of remaining samples after filtering, ensuring $R \geq 50\%$.
b) ICC(1, k) value: The reliability of the filtered dataset, aiming for the highest possible ICC.

The optimal threshold Δ* was selected based on the highest ICC(1, k) value while maintaining the retention rate above 50%.

3. Filtering Comments Based on Δ*

After selecting the optimal Δ* for each rating dimension, comments exceeding this threshold were excluded, leading to a refined dataset with improved rating consistency.

Table 3.12: ICC(1, k) of filtered data

Dimension	Original Count	Filtered Count	ICC After Filtering	F-value	df1	df2	p-value	CI95%
AC	100	66	0.6474	2.8365	65	132	2.03e-07	[-0.23, 0.38]
Actionability	100	78	0.3725	1.5936	77	156	0.0074	[0.09, 0.58]
Clarity	100	70	0.4268	1.7447	69	140	0.0029	[0.15, 0.63]
Orientation	100	59	0.6477	2.8384	58	118	8.26e-07	[0.46, 0.78]
Relevance	100	56	0.7658	4.2707	55	112	3.89e-11	[0.64, 0.85]
Specificity	100	59	0.7860	4.6720	58	118	6.83e-13	[0.67, 0.87]

The filtering process significantly improved the reliability of the dataset by removing extreme discrepancies among raters while maintaining a sufficient sample size. The application of an optimized Δ threshold resulted in a substantial increase in ICC(1, k) values across all rating dimensions, confirming an enhancement in rating consistency.

Before filtering, most dimensions exhibited low or even negative ICC(1, k) values, indicating weak agreement among raters. Wide confidence intervals and high p-values suggested that the observed rating inconsistencies were largely due to random variations rather than systematic patterns. The lack of clearly defined rating criteria and the absence of rater training may have contributed to this inconsistency, allowing individual preferences and subjective interpretations to influence the scores.

After filtering, ICC(1,k) values increased considerably, with Relevance, Specificity, and Orientation reaching values above 0.64, suggesting substantial agreement among raters. Confidence intervals narrowed significantly, and all p-values dropped below 0.05, confirming that the remaining ratings exhibit statistically significant consistency. The highest improvements were observed in Relevance (ICC(1,k) = 0.7658, CI [0.64, 0.85]) and Specificity (ICC(1,k) = 0.7860, CI [0.67, 0.87]), both of which now demonstrate strong inter-rater agreement.

The retention rate remained above 50% across all dimensions, ensuring the dataset retained a representative portion of the original samples. This suggests that the filtering process effectively removed the most inconsistent ratings while preserving sufficient data for further analysis. The increased F-values observed across all dimensions indicate improved variance consistency in ratings, further supporting the effectiveness of the filtering approach.

These results confirm that filtering successfully reduced rating variability and enhanced dataset reliability. The dimensions with the highest post-filtering ICC(1,k) values—Relevance, Specificity, and Orientation—are now suitable for predictive modeling, as they reflect a more consistent shared understanding among raters.

3.5.4. Dataset Construction

To facilitate further analysis, six separate datasets were created, one for each rating dimension. Within each dataset, the three ratings assigned to each comment were averaged, producing a single representative score per comment.

$$\text{Score} = \frac{S_1 + S_2 + S_3}{3} \tag{7}$$

Where:

(S_1, S_2, S_3) are the three ratings given to a comment (per Comment_ID)

This transformation was applied independently to each dimension. The filtered dataset is stored as six separate CSV files, one for each feature

Each file follows the same structure:

Table 3.13: Refined data structure

Label	Interpretation
Post_Title	The title of the original post related to the comment.
Post_Body	The main content of the post (may be empty for some samples).
Comment_ID	A unique identifier for each comment.
Comment_Text	The text of the comment being rated.
Score	The final averaged rating after filtering and group-wise averaging.

3.6 Estimating Model Construction

In this section, three different approaches to model construction for predicting the AC Score are presented: a rule-based model, and a linear regression model. These models aim to predict AC based on the five feature dimensions: Clarity, Specificity, Relevance, Actionability, and Orientation.

As a baseline, a model estimates AC scores by averaging the word concreteness scores from Brysbaert's word concreteness dictionary [10].

3.6.1. Rule-Based Model

A Rule-Based Model is constructed to predict AC scores based on the five dimensions: Clarity, Specificity, Relevance, Actionability, and Orientation. This approach leverages Pearson correlation coefficients[12] to assign weights to each dimension, reflecting their contribution to predicting AC. The model then calculates a weighted sum of the dimension scores to predict the AC value.

The dataset used is a merged version containing these scores for each Comment_ID, where missing values exist for some dimensions. Operations on the dataset:

1. Loading and cleaning

A script loads all datasets and merges them to a whole, ensuring that all relevant columns are available. The merged dataset is filtered to remove rows where the AC score is missing, as it is the target variable for prediction.

2. Computing Pearson correlation weights

To estimate AC, each supporting score is assigned a weight based on its Pearson correlation with the AC column. Correlation coefficients are computed across the dataset.

These values are normalized so that the sum of absolute weights equals 1. Main steps are as follows:

1) Compute Pearson Correlation Coefficients

The Pearson correlation coefficient quantifies the linear relationship between each dimension and the target variable AC. The formula is:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2} \cdot \sqrt{\sum(y_i - \bar{y})^2}} \tag{8}$$

Where:

x_i and y_i are the individual values of the predictor and target variables, respectively.

$\bar{x}$ and $\bar{y}$ represent the mean values of the predictor and target variables.

r is the Pearson correlation coefficient,

ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation).

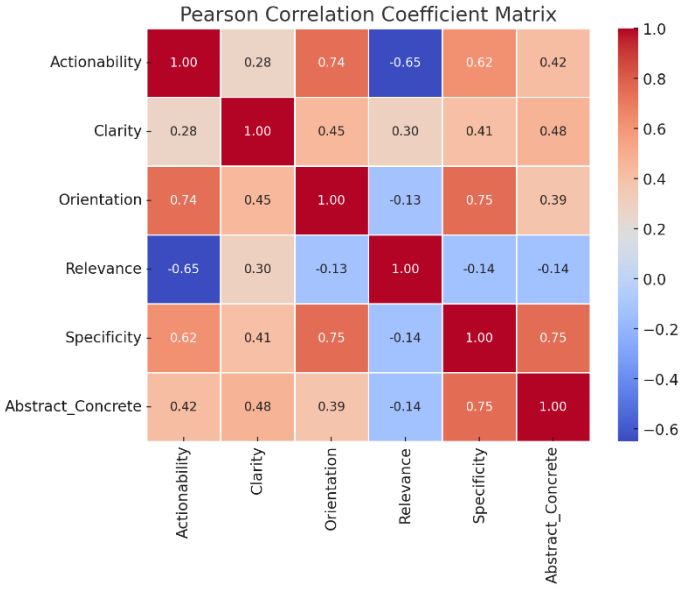

Figure 3.6: Pearson correlation coefficient

2) Normalize Correlation Coefficients to Weights

Convert the absolute values of the Pearson correlation coefficients into normalized weights. This ensures the weights sum to 1:

$$w_i = |r_{i,AC}| / \sum |r_{i,AC}| \quad (9)$$

Where:

w_i: Weight for the i-th dimension.

$r_{i,AC}$: Pearson correlation coefficient for the i-th dimension with AC.

3) Predict AC Score

Using the normalized weights, compute the predicted AC score $AC_{\text{predicted}}$ as a weighted sum of the dimension scores:

$$AC_{\text{predicted}} = \sum_{i=1}^{5} w_i \cdot X_i \quad (10)$$

Where:

X_i: Score for the i-th dimension.

w_i: Weight for the i-th dimension.

3. Handling missing support scores

Since some data according to each Comment_ID may have missing value for one or more supporting scores. A script dynamically adjusts the formula based on available scores per comment. Instead of using fixed weights, the weights are recomputed only for the available scores before applying the weighted sum formula. For each comment where some supporting scores are missing, the prediction formula is modified as follows:

$$AC_{predicted} = \Sigma(w_i * X_i), i \in availablescores \quad (11)$$

where:

X_i is an available supporting score.

w_i is the adjusted weight, recomputed as:

$$w_i = \frac{Pearson\ correlation(X_i, AC)}{\Sigma\left|Pearson\ correlation\left(X_j, AC\right)\right|}, j \in availablescores \quad (12)$$

If all supporting scores are missing, the mean AC score from the dataset is used as a fallback.

3.6.2. Linear Regression Model

Linear Regression is a supervised learning algorithm used for modeling the relationship between a dependent variable (in this case, AC) and one or more independent variables (the five dimensions: Clarity, Specificity, Relevance, Actionability, and Orientation). The relationship is represented as a linear equation [13]:

$$y = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b \quad (13)$$

Where:
y: Target variable (AC score).
x_i: Feature values (dimension scores).
w_i: Coefficients or weights for each feature.
b: Intercept.

The goal of Linear Regression is to minimize the error between the predicted and actual values by optimizing the weights w_i.

The model finds the optimal coefficients w_i and intercept b by minimizing the Mean Squared Error (MSE):

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (14)$$

Where:
y_i: Actual score.
$\hat{y}_i$: Predicted score.

The dataset is processed to handle missing values and evaluated using 5-Fold Cross-Validation to ensure robustness. Data Processing Steps:

1. Handling Missing Values

Rows where AC is missing are removed because it is the target variable. For the supporting scores, missing values are imputed with the column mean.

$$X_{i,j} = \frac{1}{N}\sum_{k=1}^{N} X_{k,j}\,, \quad \text{if } X_{i,j} \text{ is missing} \tag{15}$$

Where:

$X_{i,j}$ is the missing value for the i-th comment and j-th score

N is the total number of available values in the same column.

This ensures that missing values do not result in excessive data loss while maintaining statistical consistency.

2. Applying 5-Fold Cross-Validation

The dataset is divided into five equal folds for training and testing. Each fold serves as the test set once, while the remaining four folds are used for training.

Model Training and Prediction Steps:

1. Training the Linear Regression Model

For each fold, a linear regression model is trained using the five supporting scores as input features. The model learns the relationship between the supporting scores and AC using the equation:

$$AC_{\text{predicted}} = \beta_0 + \beta_1 A + \beta_2 C + \beta_3 O + \beta_4 R + \beta_5 S + \epsilon \tag{16}$$

where:

β_0 is the intercept

β_i are the learned coefficients for each supporting score

ϵ represents residual error

A, C, O, R, S correspond to (Actionability, Clarity, Orientation, Relevance, Specificity).

2. Predicting AC Scores

The trained model predicts AC scores for the test set in each fold. The predicted AC scores for each Comment_ID are stored in a merged dataset for evaluation.

3.6.3. Baseline Score: Word Concreteness Dictionary

To evaluate the effectiveness of the predictive models, we introduce a baseline model that estimates the AC score using Brysbaert's word concreteness dictionary [10]. This serves as a simple heuristic for measuring concreteness based solely on the lexical properties of the comment text, independent of the five dimension scores (Actionability, Clarity, Orientation, Relevance, and Specificity).

The baseline AC score AC_{baseline} is computed as the average concreteness score of words in the comment text using the following formula:

$$AC_{\text{baseline}} = \frac{1}{N}\sum_{i=1}^{N} C(w_i) \tag{17}$$

Where:

N is the number of words in the comment text.

$C(w_i)$ represents the concreteness score of word w_i obtained from a predefined word concreteness dictionary.

Since English naturally contains spaces, split() in Python is sufficient for tokenization. The preprocessing of comment text is designed to ensure that words are properly mapped to their corresponding concreteness scores in the dictionary. To achieve this, several steps are applied:

1. Removing Punctuation

Since the dictionary contains words in their base forms, punctuation marks (e.g., commas, periods, apostrophes) are removed using regular expressions. This prevents words from being mismatched due to attached punctuation.

2. Lowercasing

All words are converted to lowercase to ensure consistency with the dictionary, which stores words in lowercase. This avoids missing matches due to case differences (e.g., "House" vs. "house").

3. Tokenization

The text is split into individual words using whitespace-based tokenization. Since English naturally separates words with spaces, this simple method is sufficient for our purposes.

4. Lemmatization

Words are converted to their base forms (lemmas) using the WordNet [70] lemmatizer.

This step ensures that different grammatical variations of a word (e.g., plural nouns, verb conjugations) are mapped to the same dictionary entry.

For example:

"cats" → "cat"

"running" → "run"

“has” → "have”

5. Handling Missing Words

The dictionary used in this study provides precomputed concreteness scores for 39954 English words, typically ranging from 1 (very abstract) to 5 (very concrete), which meets the scale in this study. If a word is not found in the dictionary, it is assigned to a default value C_{default} based on the average concreteness score of known words:

$$C_{\text{default}} = \frac{1}{M}\sum_{j=1}^{M} C(w_j) \tag{18}$$

Where:

M is the total number of words in the word concreteness dictionary.

$C(w_j)$ represents the concreteness score for the j-th word in the dictionary.

As a result:

$$C_{\text{default}} = 3.0363 \tag{19}$$

This prevents missing words from being arbitrarily excluded while maintaining reasonable estimates for unknown words.

An example of how a baseline score is circulated given as below:

Table 3.14: Examples of AC baseline

Comment Text	Word Concreteness Score	AC Baseline	AC Score annotated
Even if there are no corners, the circle has the minimum boundary and therefore minimum friction against the	even: 2.79 if: 1.19 there: 2.2	2.5905	4.3333

interior surface.	are: 1.96 no: 2.45 corner: 4.61 the: 1.43 circle: 4.44 have: 2.18 minimum: 2.25 boundary: 3.04 and: 1.52 therefore: 1.33 friction: 3.0 against: 1.8 interior: 3.59 surface: 4.26		

3.7 Feature Extracting Model Construction

This section explores various feature extraction models to automatically derive the six rating scores (AC, Actionability, Clarity, Orientation, Relevance, and Specificity) from the comment text. The extracted scores will be compared with human annotations to evaluate their accuracy.

Additionally, we will integrate these models into the predictive framework by combining the extracted five scores (Actionability, Clarity, Orientation, Relevance, and Specificity) with the previously developed prediction models in 3.6 to estimate AC. Finally, we compare:

1. Predicted AC (using five extracted scores as input)
2. Directly extracted AC from feature models
3. Baseline AC (computed from the word concreteness dictionary)

3.7.1. TF-IDF + Regression

In this approach, we employ TF-IDF (Term Frequency-Inverse Document Frequency) [66] with a regression model to predict the six rating scores (Abstract_Concrete, Actionability, Clarity, Orientation, Relevance, and Specificity) from comment text. The approach leverages TF-IDF vectorization to transform raw text into numerical representations and applies machine learning to estimate the target scores.

The dataset is loaded from a merged dataset, which contains Comment_ID, Post_Title, Post_Body, Comment_Text, and the six target scores. A script is programmed to remove rows where data of Comment_Text is missing, as these cannot be processed through TF-IDF.

TF-IDF converts a comment into a numerical feature vector that reflects the importance of each word in the corpus. This process transforms each comment into a high-dimensional numerical representation. The weight of the word w in the comment d is computed as:

$$\text{TF-IDF}(w, d) = \text{TF}(w, d) \times \text{IDF}(w) \tag{20}$$

Where:

TF (Term Frequency): The number of times word w appears in comment d.

$$\text{TF}(w, d) = \frac{\text{count}(w, d)}{\sum_{\text{all words } v} \text{count}(v, d)} \tag{21}$$

IDF (Inverse Document Frequency): Measures how rare word w is across all comments.

$$\text{IDF}(w) = log \frac{N}{\text{DF}(w) + 1} \tag{22}$$

where N is the total number of comments and $\text{DF}(w)$ is the number of comments containing word w.

After transforming comments into TF-IDF vectors, we train a regression model to predict each rating score. The model learns a function:

$$y_i = f\big(\text{TF-IDF}(d_i)\big) \tag{23}$$

where:

y_i is one of the six rating scores for comment d_i.

f is a trained regression function that maps TF-IDF features to a score.

For this study, we use Ridge Regression [67]:

$$\min_{w} \left|\left|Xw - y\right|\right|^2 + \lambda \left|\left|w\right|\right|^2 \quad (24)$$

where:

X is the TF-IDF feature matrix.

w is the weight vector learned by the model.

y is the actual score.

λ is a regularization parameter that prevents overfitting.

The training process in each score prediction task includes:

1) Split the dataset into training and test sets for the current fold.
2) Train the model using TF-IDF features as input and the target score as output.
3) Predict the target score for the test set.
4) Store predictions for future evaluation.

The dataset is randomly split into five folds to apply 5-fold cross validation. The model is trained on 4 folds and then tested on the remaining fold. This process repeats five times, ensuring that each comment is tested once. For each target score, the model generates predictions and stores them in separate files. Predictions are merged in a final dataset after six target scores extracting tasks are finished. The final date included the following columns in Table 3.15:

Table 3.15: Dataset columns

Comment_ID	AC	Actionability	Clarity	Orientation	Relevance	Specificity

3.7.2. GPT-4 Few-Shot Prompt

GPT-4 [68] has powerful natural language understanding capabilities, making it an effective tool for extracting structured ratings from text. In this section, we explore how few-shot prompting [69] can be used to guide GPT-4 in assigning scores for the six predefined dimensions.

Each comment is associated with a 'Comment_ID', and the dataset also contains 'Post_Title' and 'Post_Body' relating the comment to provide context. Prior to applying GPT-4, we ensured that each comment had a corresponding textual input and was formatted appropriately.

Few-shot prompting was used to guide GPT-4 toward producing structured and consistent rating outputs. The prompt provided GPT-4 with:

1) A task definition
2) Definitions of rating scales, ranging from 1 (lowest) to 5 (highest).
3) Example ratings, demonstrating how human annotators assigned scores to various comments.

A generic task defining part of the prompt format is as follows in Table 3.16:

Table 3.16: Few-shot prompt

```
### Task:
You are an expert in evaluating text based on the following dimension:{Dimension}
AC: Measures how concrete (5) or abstract (1) the statement is.
Actionability: Measures whether the comment suggests a clear action (5 = highly actionable).
Clarity: Measures how clear and understandable the comment is (5 = very clear).
Orientation: Measures the forward-looking nature of the comment (5 = highly goal-oriented).
Relevance: Measures how relevant the comment is to the discussion (5 = highly relevant).
Specificity: Measures the level of detail in the comment (5 = very specific).
Below are examples of how comments are rated:
---
Example {$No}:
Topic in the discussion: {$Post_Title} – {$Post_Body}
Comment: {$Comment_Text}
{Dimension} Score: {$Score}
---
Read the given comment and provide a numerical rating (1-5) based on the definitions below:
- 1: Very low in {Dimension}
- 5: Very high in {Dimension}
### Comment:
{Comment_Text}
### Output Format:
Provide only a JSON object:
{
    "{Dimension}": X
```

```
}
```

GPT-4's output is inherently non-deterministic due to its probabilistic nature [74]. The same input can yield slightly different outputs across different runs, which may introduce noise in the rating predictions. To improve the consistency and robustness of the extracted ratings, we implemented a 5-fold cross-validation-like technique. Unlike traditional machine learning models, we do not train GPT-4 but instead use a structured evaluation method to stabilize the predictions by leveraging multiple inference runs.

The key idea is that each comment should be rated multiple times across different subsets of the dataset, reducing variance in its final assigned score. The steps for this process are as follows:

1) Split the dataset into five equal-sized subsets.
2) For each fold, select four subsets to serve as contextual examples and leave one subset out for evaluation.
3) Apply GPT-4 predictions to the left-out subset by providing few-shot examples from the four training subsets.
4) Repeat the process for all five folds, ensuring each comment is rated five times.

Aggregate the final scores by computing the mode of the five predictions. If multiple values share the highest frequency, their mean is taken as the final score.

Chapter 4

Experimentation and Evaluation

4.1 Overview

This chapter presents a comprehensive evaluation of different approaches for estimating the AC score. The data used in this study comes from crowdsourced ratings that have been filtered for consistency using inter-rater reliability metrics in Chapter 3. The evaluation employs a range of assessment methods, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and R^2 to compare model predictions with human annotations.

The study explores AC estimating models based on the five supporting scores (Actionability, Clarity, Orientation, Relevance, and Specificity):

1) Rule-Based Model, which calculates AC as a weighted sum of the five scores, with Pearson correlation coefficients as weights.
2) Linear Regression Model, which learns the optimal mapping between the five scores and AC through supervised learning.

To provide a baseline comparison, AC is also estimated using a word concreteness dictionary, where the AC score is derived by averaging the concreteness scores of words in each comment. The effectiveness of these estimation models is validated by comparing their outputs with human-annotated AC scores.

Beyond models that rely on human- annotated scores, we explore automated approaches for extracting scores directly from raw text. Two methods are tested:

1) A traditional machine learning method based on TF-IDF, which represents comments using word frequency-based features and applies a regression model to predict all six scores.
2) An LLM-based method of GPT-4 Few-Shot Prompting, which directly extracts the six scores from text using structured prompts designed to guide the model.

The extracted scores are first compared against human annotations to evaluate their accuracy. Next, the five predicted supporting scores are used as input to the AC estimating models, allowing us to assess how well a fully automated system can estimate AC without relying on human- annotated data.

Finally, we conduct a comprehensive comparison, evaluating:

1) Human-Annotated Scores (Ground Truth)

2) AC Predictions Based on Human Ratings (Rule-Based & Linear Regression)
3) AC Predictions Based on Automatically Extracted Scores
4) Directly Extracted AC Scores from GPT-4 and TF-IDF Models
5) Baseline AC Scores (Word Concreteness Dictionary)

This evaluation provides a performance comparison between manual annotation, predictive models, and fully automated text-based extraction methods, helping to determine the most effective approach for AC estimation.

4.2 Evaluation Metrics

The following metrics were used to assess the model's performance:

1. Mean Absolute Error (MAE)

MAE measures the average absolute difference between the predicted and human-annotated scores. It provides a simple and interpretable measure of prediction accuracy.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{25}$$

where:

N is the number of samples

y_i is the actual AC score for sample i

$\hat{y}_i$ is the predicted AC score for sample i

A lower MAE indicates better prediction accuracy.

2. Root Mean Squared Error (RMSE)

RMSE also measures the difference between predicted and actual AC scores but penalizes larger errors more than MAE by squaring them before averaging.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{26}$$

where:

N is the number of samples

y_i is the actual AC score for sample i

$\hat{y}_i$ is the predicted AC score for sample i

RMSE is always greater than or equal to MAE, as it squares the differences before averaging. A lower RMSE indicates better prediction accuracy. RMSE penalizes large errors more heavily, making it useful for detecting models that produce high-variance predictions.

3. Coefficient of Determination (R^2)

The Coefficient of Determination R^2, also known as the goodness-of-fit measure, quantifies how well the predicted AC scores explain the variability in the actual AC scores.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \quad (27)$$

where:

N is the number of samples

y_i is the actual AC score for sample i

$\hat{y}_i$ is the predicted AC score for sample i

$\bar{y}$ is the mean of all actual AC scores

$R^2 \in \left(-\infty, 1\right]$ A higher R^2 indicates better predictive performance.

4.3 Experimentation Setup

This section presents the experimental setup for evaluating different approaches to predicting AC scores and extracting the six rating dimensions from text. The experiments are divided into three main parts:

1) Baseline Model: Uses a word concreteness dictionary to estimate AC scores.
2) AC Prediction Models: Predict AC scores from the five supporting dimensions.
3) Feature Extraction Models: Extract AC, Actionability, Clarity, Orientation, Relevance, and Specificity from comment text.

Each experiment is evaluated using MAE, RMSE, and R^2.

4.4 Results and Observations

In this section, we present the results obtained from the experiments and provide observations based on the data collected. The result dataset consists of multiple scoring outputs, each representing different prediction approaches.

4.4.1. AC Estimation

The collected AC scoring datasets include six different sources, each representing a different approach to AC estimation assessment. These datasets are merged and identified by a feature name. Show as:

Table 4.1: Composition of AC scores

Feature	Description	Count	Directly collected from text?
AC_standard	Ground truth. Filtered scores assigned by human annotators.	66	Yes
AC_baseline	Predicted scores generated by a algorithm based on a word dictionary.	66	Yes
AC_rule	Predicted scores by a rule-based model calculating weights of dimension scores.	66	No
AC_linear	Predicted scores generated by a linear regression model built on dimension scores.	66	No
AC_idftf	Predicted scores generated by a regression model built on TF-IDF features.	66	Yes
AC_gpt	Predicted scores generated by GPT-4 with few-shot prompt.	66	Yes

To ensure a direct comparison, we take the intersection of all datasets using the unique identifier 'Comment_ID'. By aligning records based on this identifier, we extract a subset of 66 common data points from each dataset. This allows for direct evaluation and correlation analysis across different scoring methods, ensuring that all models are assessed on the same set of textual data.

We employed three standard evaluation metrics to assess the effectiveness of different AC estimation models: MAE, RMSE, and R^2. The ground truth (AC_standard) was used as the reference, and we compared predictions from five different models:

Table 4.2: MAE, RMSE, and R^2 between models and the standard

Model	MAE	RMSE	R^2
AC_baseline	1.13	1.21	-5.83
AC_rule	0.39	0.50	-0.14
AC_linear	0.34	0.41	0.24
AC_idftf	0.39	0.48	-0.06
AC_gpt	0.82	1.00	-3.64

1. MAE

AC_linear (0.34) achieved the lowest MAE, suggesting it had the most accurate predictions on average. AC_rule (0.38) and AC_idftf (0.39) performed similarly and were only slightly worse than AC_linear. AC_baseline (1.13) had the highest MAE, indicating poor estimation capability. AC_gpt (0.82) also exhibited a relatively high MAE, suggesting that GPT-4 predictions were less aligned with human-annotated AC scores.

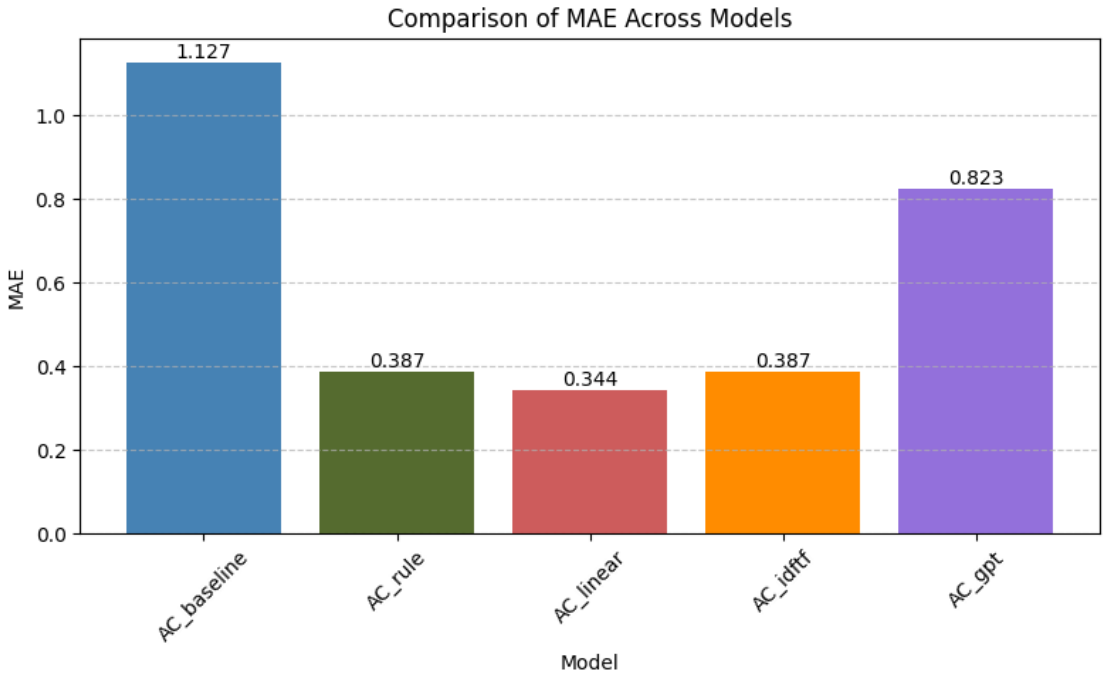

Figure 4.1: MAE between Models and the standard

2. RMSE

AC_linear (0.41) again performed best, followed closely by AC_rule (0.50) and AC_idftf (0.48). AC_baseline (1.22) had the worst RMSE, reinforcing its poor estimation accuracy.AC_gpt (1.00) had a significantly higher RMSE, further confirming that GPT-4's predictions were less reliable.

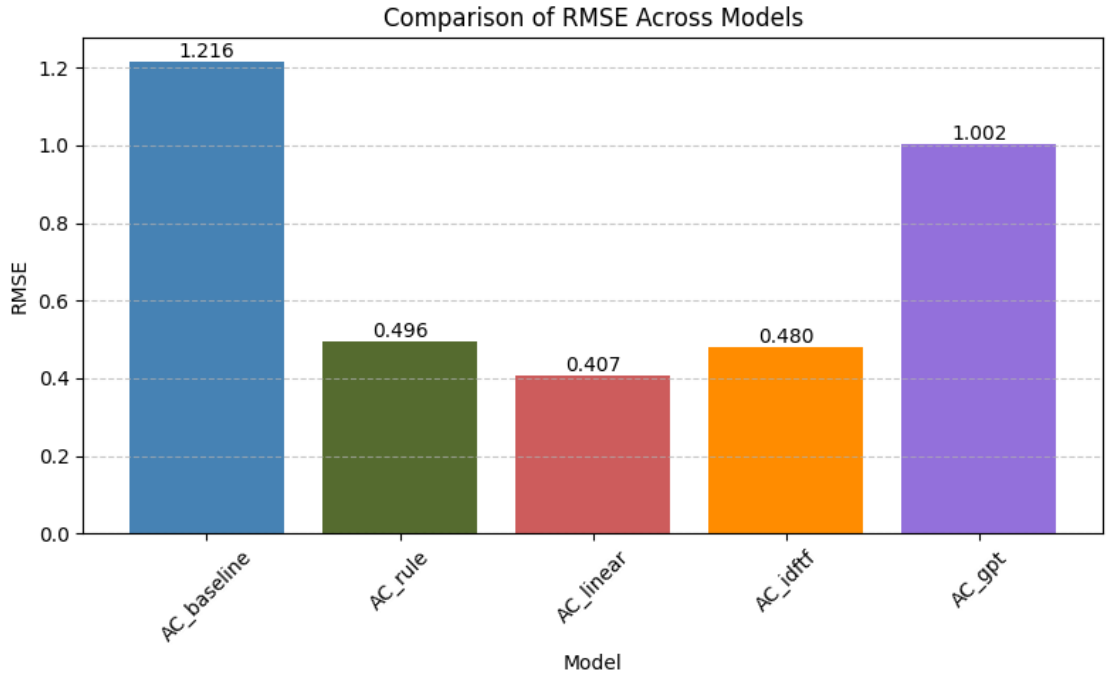

Figure 4.2: RMSE between Models and the standard

3. R^2

AC_linear (0.24) was the only model with a positive R^2, indicating it explained 24% of the variance in AC scores. AC_rule (-0.13) and AC_idftf (-0.06) had slightly negative R^2 values, suggesting that while they performed better than some models, they still did not generalize well. AC_gpt (-3.64) and AC_baseline (-5.83) had highly negative R^2 values, showing they performed significantly worse than a naive mean prediction.

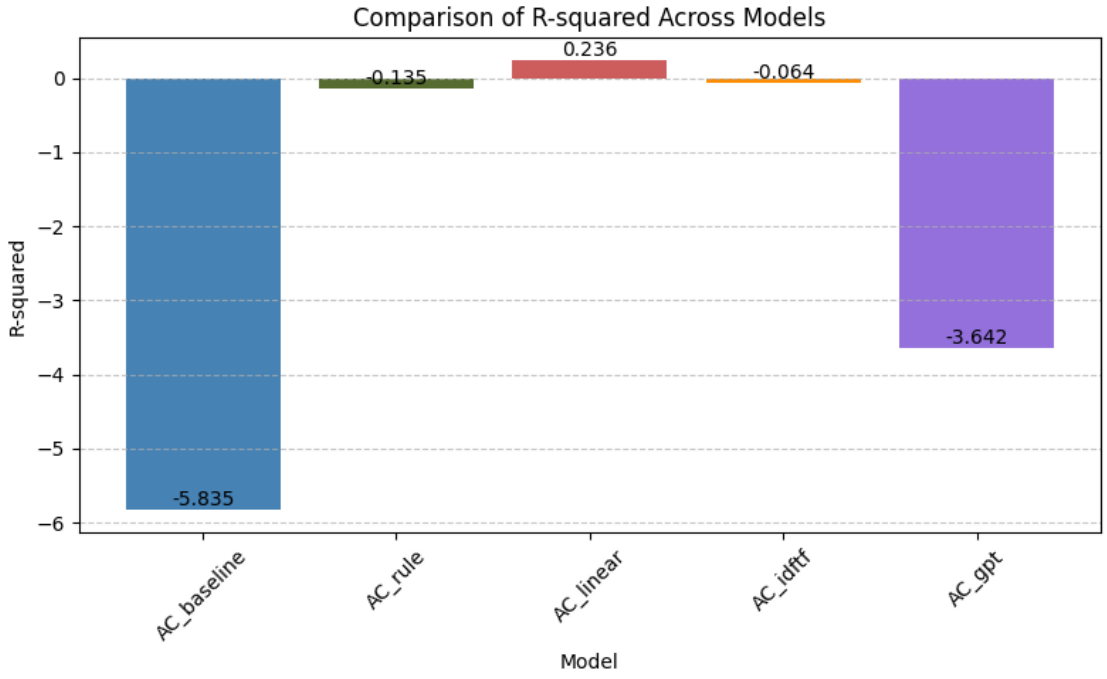

Figure 4.3: R^2 between Models and the standard

Evaluating different AC estimation models highlights the distinction between models based on dimension scores and those extracting features directly from text. The dimension-based models, including AC_rule and AC_linear, rely on structured numerical inputs derived from five supporting dimensions—Clarity, Specificity, Relevance, Actionability, and Orientation. In contrast, the text-based models, such as AC_baseline, AC_idftf, and AC_gpt, attempt to infer AC scores directly from raw textual data, either through dictionary-based methods, regression on TF-IDF features, or GPT-4 predictions.

In terms of accuracy, models utilizing dimension scores consistently outperformed those relying solely on text. The linear regression model AC_linear achieved the lowest MAE (0.34) and RMSE (0.41), making it the most accurate method for predicting AC. The rule-based model AC_rule, which applies weighted calculations to dimension scores, performed slightly worse but still maintained a significantly lower error compared to text-based approaches. These results suggest that the five supporting dimensions proposed in this study play an essential role in accurately predicting AC, as they provide structured and interpretable information that leads to more precise estimations. In contrast, text-based models exhibited significantly higher errors, indicating that extracting features directly from comment text may not accurately reflect human perceptions of AC. The dictionary-based baseline model AC_baseline performed the worst, demonstrating the limitations of relying solely on predefined word concreteness scores. Similarly, GPT-4's

few-shot predictions resulted in high error values, suggesting that directly prompting a language model to assess AC scores does not yield stable results.

Regarding model generalization, the R^2 values further illustrate the performance differences between these approaches. Among all models, only AC_linear achieved a positive R-squared value (0.24), indicating that it explained 24 percent of the variance in AC scores. The rule-based model AC_rule and the TF-IDF-based regression model AC_idftf had slightly negative R^2 values (-0.13 and -0.06), meaning they performed better than random guessing but still had room for improvement. The text-based models AC_baseline (-5.83) and AC_gpt (-3.64) exhibited highly negative R^2 values, confirming that they performed substantially worse than a simple mean prediction. However, the reduced negative impact observed in AC_idftf compared to other text-based methods suggests that regression models can significantly enhance performance when applied to extracted textual features. This indicates that while raw textual representations alone may not reliably predict AC, applying a structured learning process, such as regression, can mitigate prediction errors.

The observed differences in performance align with expectations based on model structure. The dimension-based models benefited from the explicitly defined relationships between AC and supporting dimensions, leading to more stable predictions. On the other hand, text-based models suffered from the inherent variability of language and the difficulty of capturing AC directly from unstructured text. The dictionary-based approach struggled due to its inability to account for contextual variations, while the GPT-4 model exhibited inconsistencies likely due to the randomness in language model outputs. The TF-IDF regression model performed better than the dictionary-based and GPT-4 models, highlighting that transforming text into structured features improves estimation quality before applying a regression framework.

4.4.2. Dimension Features Extraction

This section evaluates the effectiveness of extracting the five supporting dimensions—Clarity, Specificity, Relevance, Actionability, and Orientation—directly from comment text. Two primary approaches were used: a TF-IDF-based regression model and GPT-4 few-shot prompting. The extracted scores were compared against human-annotated dimension ratings to assess the reliability and accuracy of each method.

Table 4.3: Composition of Dimension scores

Feature	Count

Actionability_standard	78
Actionability_tfidf	78
Actionability_gpt	78
Clarity_standard	70
Clarity_tfidf	70
Clarity_gpt	70
Orientation_standard	59
Orientation_tfidf	59
Orientation_gpt	59
Relevance_standard	56
Relevance_tfidf	56
Relevance_gpt	56
Specificity_standard	59
Specificity_tfidf	59
Specificity_gpt	59

The TF-IDF regression model was trained on the original dataset, ensuring its predicted scores correspond directly to the number of available human-annotated standard scores. In contrast, the GPT-4 model always produced scores for all comments, including those without human ratings. All scores are associated with unique identifiers. For the subsequent evaluation, only the intersection of these datasets will be considered, meaning that analysis will be conducted using the subset of data for which human ratings (standard scores) are available. To ensure direct comparability, the extracted ratings from both models were evaluated against the standard scores. The evaluation was performed using MAE, RMSE, and R^2 values.

Table 4.4: MAE, RMSE, and R^2 between models and the standard

Dimension	Model	MAE	RMSE	R2
Actionability	TF-IDF	0.526795566	0.663069243	-0.054769888
Actionability	GPT-4	1.726495726	1.937043302	-8.001577287
Clarity	TF-IDF	0.324782183	0.377069998	0.052980419
Clarity	GPT-4	0.821428571	0.972437618	-5.298519861
Orientation	TF-IDF	0.40090767	0.509657297	-0.029831095
Orientation	GPT-4	1.129943503	1.271148775	-5.406226272
Relevance	TF-IDF	0.518751288	0.619398272	-0.047726821

Relevance	GPT-4	1.529761905	2.011376375	-10.04828254
Specificity	TF-IDF	0.456902154	0.583595824	0.124966474
Specificity	GPT-4	1.146892655	1.277798248	-3.194931934

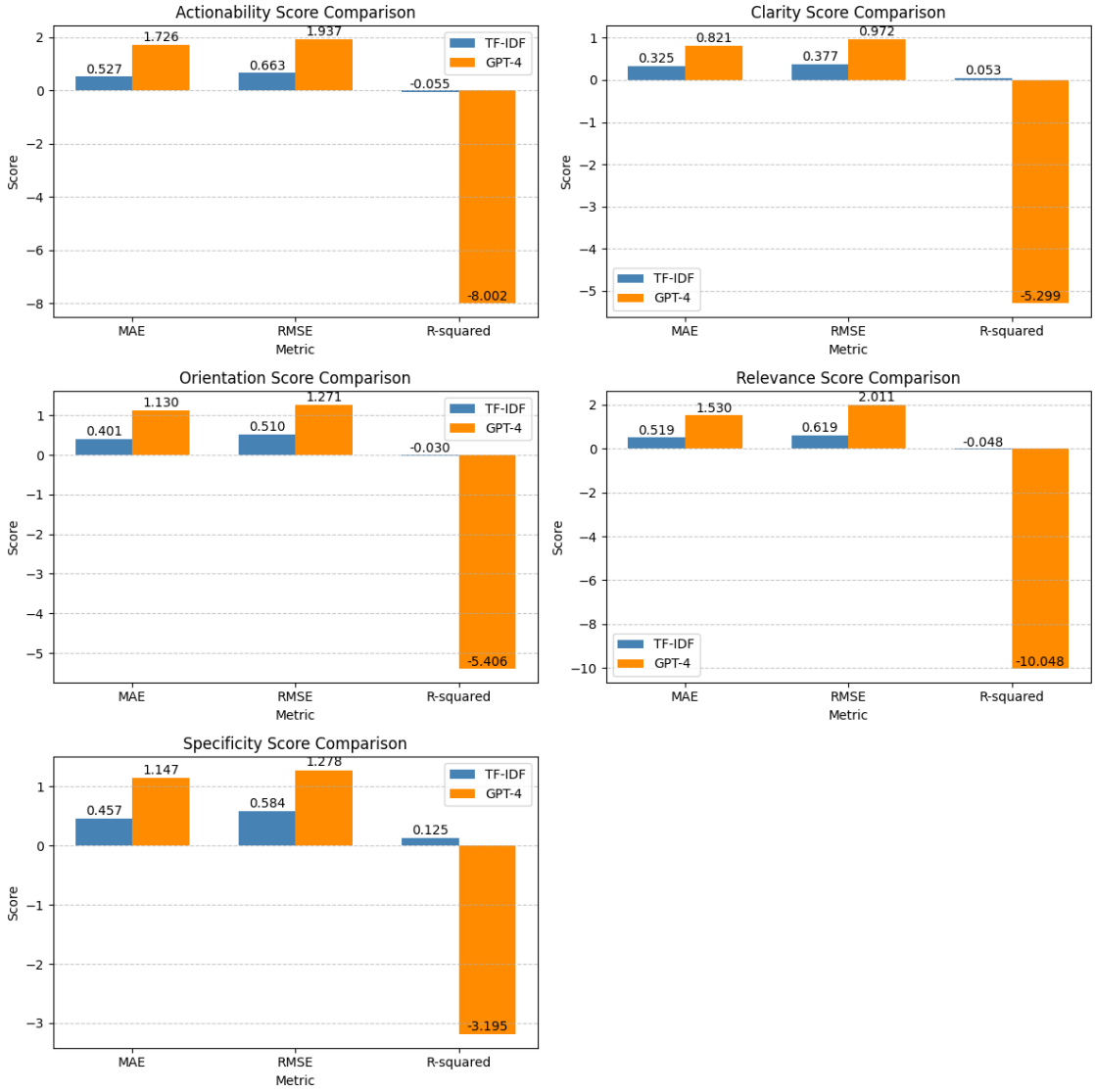

Figure 4.4: MAE, RMSE, and R^2 between models and the standard

The evaluation results of the extracted dimension scores reveal distinct patterns in model performance, highlighting the relative strengths and weaknesses of the TF-IDF Ridge Regression model and the GPT-4 Few-Shot Prompting approach. Across all five dimensions—Actionability, Clarity, Orientation, Relevance, and Specificity—TF-IDF Ridge Regression consistently demonstrated better predictive performance, as reflected by lower MAE and RMSE values. In contrast, GPT-4 exhibited significantly higher errors, suggesting greater prediction variability and a weaker alignment with standard scores.

Among the five dimensions, Clarity and Specificity were where both models performed relatively well. The TF-IDF Ridge model achieved its lowest MAE and RMSE in Clarity, indicating that traditional feature-based regression methods more easily capture textual features contributing to clarity. Specificity also demonstrated higher predictive reliability, as reflected in its comparatively better R2 value. This suggests that the degree of detail in a comment can be effectively quantified using a structured feature-based approach.

On the other hand, Relevance and Orientation posed the greatest challenges for both models, especially for GPT-4. The GPT-4 model exhibited the highest MAE and RMSE in Relevance, along with an extremely low R2 value, indicating that its predictions deviated significantly from human annotations. Orientation also showed poor predictive performance, suggesting that goal-directed aspects of a comment are more difficult to capture using a language model without additional structured guidance.

The overall findings suggest that TF-IDF Ridge Regression provides a more stable and reliable approach for dimension score extraction, as it consistently produces lower error rates across all dimensions. The structured nature of this model, which relies on predefined linguistic features, appears to be advantageous in aligning with human-assigned ratings. Conversely, GPT-4's performance indicates higher variability, likely due to the inherent randomness in its language model predictions. This variability is particularly pronounced in dimensions such as Relevance and Orientation, where subjective context plays a critical role. These findings underscore the limitations of using a large language model for direct numerical estimation without additional fine-tuning or explicit prompt engineering strategies.

Despite its inconsistencies, GPT-4's potential for feature extraction remains promising, particularly if refined through better prompt design and more controlled input structures. The observed weaknesses in Relevance and Orientation suggest that future iterations could benefit from hybrid models that integrate structured feature-based techniques with LLM-generated embeddings. Improving the consistency of LLM-generated scores may also require a larger number of few-shot examples that explicitly demonstrate nuanced distinctions between different levels of clarity, specificity, and relevance. Another promising direction would be the introduction of context-aware models that incorporate post titles and discussion history more effectively to improve relevance-based predictions.

Chapter 5

Conclusion

5.1 Summary

This research aimed to construct a model for estimating the abstraction-concreteness (AC) score of comment texts in online discussions. To achieve this goal, three key research questions were explored: identifying potential factors influencing AC scores, developing effective methods for data collection and annotation, and constructing and evaluating models for AC score prediction.

To address the first research question, this study examined five supporting dimensions—Actionability, Clarity, Orientation, Relevance, and Specificity—as potential factors influencing AC scores. The analysis demonstrated that these dimensions are significant predictors of AC, as models using them as input features consistently outperformed text-based approaches. This finding suggests that abstractness and concreteness are not solely intrinsic to individual words but are shaped by broader linguistic and contextual factors.

For the second research question, a dataset was constructed by collecting comment texts from online discussions and obtaining human-annotated AC scores. To ensure reliability, a filtering process was applied to remove annotations with high disagreement among raters. The dataset also included scores for the five supporting dimensions, providing structured features for AC prediction. Additionally, GPT-4 was employed to generate synthetic scores for comparison, allowing an evaluation of automated feature extraction methods.

The third research question was addressed through the development and evaluation of multiple models for AC score prediction. Dimension-based models, including rule-based weighting and linear regression, were shown to be the most effective, highlighting the predictive power of structured numerical features. In contrast, text-based models, such as those utilizing TF-IDF regression and GPT-4 few-shot prompting, exhibited higher error rates. While TF-IDF regression showed moderate success, the GPT-4 model struggled with consistency, reinforcing the challenges of using large language models for structured predictions without domain-specific fine-tuning.

Overall, this research provides a systematic approach to AC estimation, demonstrating that structured feature-based models outperform direct text-based methods. The findings

confirm that multiple linguistic factors influence AC perception, and leveraging structured annotations significantly improves predictive accuracy. These insights contribute to a deeper understanding of abstractness and concreteness in natural language and offer a foundation for further advancements in text classification and automated linguistic analysis.

5.2 Future Work

Future research should refine the definition of concreteness to improve annotation consistency and model accuracy. Enhancing the survey methodology, such as using pairwise ranking or multi-stage rating systems, could lead to more reliable data collection.

Expanding the dataset to include diverse text genres, such as scientific papers, journalism, and instructional writing, would improve model generalization. Additional linguistic features, such as emotional tone and complexity, could also be explored to better capture factors influencing AC perception.

Deep learning approaches, including fine-tuned transformer models like BERT, may enhance AC prediction. Hybrid models that combine linguistic features with neural networks may improve both interpretability and accuracy.

Practical applications include integrating AC scoring into online platforms to encourage clearer discussions, improving content moderation, and assisting with educational writing feedback. Additionally, AC estimation could help in policy communication by making government and instructional materials more accessible.

Finally, expanding the analysis of AC estimation across different languages and cultures is another important direction. The perception of concreteness may vary depending on linguistic and cultural factors, and future studies could investigate whether AC models trained on English text can be adapted to other languages. Multilingual datasets and cross-linguistic comparisons would provide valuable insights into the universality of AC features and help build more globally applicable models.

Publications

Mingxi Hu, Wen Gu, Koichi Ota, Shinobu Hasegawa. "Estimating Text Concreteness in Online Discussions" in *電子情報通信学会 合意と共創研究会(Consen)* . (2025 in press).

Bibliography

[1] Zhu, Jianfeng, et al. "Investigating COVID-19's impact on mental health: trend and thematic analysis of Reddit Users' discourse." *Journal of medical Internet research* 25 (2023): e46867.

[2] Lampe, Cliff, and Paul Resnick. "Slash (dot) and burn: distributed moderation in a large online conversation space." *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2004.

[3] Wei, Zhongyu, Yang Liu, and Yi Li. "Is this post persuasive? ranking argumentative comments in online forum." *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2016.

[4] Christensen, Jon, et al. "Defining new criteria for selection of cell-based intestinal models using publicly available databases." *BMC genomics* 13 (2012): 1-11.

[5] Cahill, Larry, and James L. McGaugh. "A novel demonstration of enhanced memory associated with emotional arousal." *Consciousness and cognition* 4.4 (1995): 410-421.

[6] Joo, Sungmin, and Hideaki Takeda. "Analysis of Discussion Page in Wikipedia Based on User's Discussion Capability." *2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*. Vol. 1. IEEE, 2012.

[7] Park, Deokgun, et al. "Supporting comment moderators in identifying high quality online news comments." *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2016.

[8] Wang, Yucheng, et al. "Automatic essay scoring incorporating rating schema via reinforcement learning." *Proceedings of the 2018 conference on empirical methods in natural language processing*. 2018.

[9] Wanas, Nayer, et al. "Automatic scoring of online discussion posts." *Proceedings of the 2nd ACM workshop on Information credibility on the web*. 2008.

[10] Brysbaert, Marc, Amy Beth Warriner, and Victor Kuperman. "Concreteness ratings for 40 thousand generally known English word lemmas." *Behavior research methods* 46 (2014): 904-911.

[11] Warrens, Matthijs J. "Cohen's linearly weighted kappa is a weighted average." *Advances in Data Analysis and Classification* 6.1 (2012): 67-79.

[12] Cohen, Israel, et al. "Pearson correlation coefficient." *Noise reduction in speech*

processing (2009): 1-4.

[13] Poole, Michael A., and Patrick N. O'Farrell. "The assumptions of the linear regression model." *Transactions of the Institute of British Geographers* (1971): 145-158.

[14] Zhang T, Schoene A M, Ji S, et al. Natural language processing applied to mental illness detection[J]. 2022.

[15] Wojcieszak, Magdalena. "False consensus goes online: Impact of ideologically homogeneous groups on false consensus." *Public Opinion Quarterly* 72.4 (2008): 781-791.

[16] Dawson, Shane. "Online forum discussion interactions as an indicator of student community." *Australasian Journal of Educational Technology* 22.4 (2006).

[17] Hargreaves, Andy, Lorna Maxine Earl, and James Ryan. "Schooling for change: Reinventing education for early adolescents." *(No Title)* (1996).

[18] Rovai, Alfred P. "Sense of community, perceived cognitive learning, and persistence in asynchronous learning networks." *The Internet and Higher Education* 5.4 (2002): 319-332.

[19] Tinto, Vincent. "Learning communities: Building gateways to student success." The National Teaching and Learning Forum. Vol. 7. No. 4. 1998.

[20] Richardson, John TE. "Imagery, concreteness, and lexical complexity." The Quarterly Journal of Experimental Psychology 27.2 (1975): 211-223.

[21] Paivio, Allan, John C. Yuille, and Stephen A. Madigan. "Concreteness, imagery, and meaningfulness values for 925 nouns." Journal of experimental psychology 76.1p2 (1968): 1.

[22] Yang, Diyi, et al. "Forum thread recommendation for massive open online courses." *Educational Data Mining 2014*. 2014.

[23] Biyani, Prakhar, et al. "I want what I need! Analyzing subjectivity of online forum threads." *Proceedings of the 21st ACM international conference on Information and knowledge management*. 2012.

[24] https://www.reddit.com/r/explainlikeimfive/comments/1ib25kd/eli5_why_is_animation_so_time_consuming/?utm_source=share&utm_medium=web3x&utm_name=web3xcss&utm_term=1&utm_content=share_button

[25] Pitler, Emily, and Ani Nenkova. "Revisiting readability: A unified framework for predicting text quality." *Proceedings of the 2008 conference on empirical methods in natural language processing*. 2008.

[26] Schriver, Karen A. "Evaluating text quality: The continuum from text-focused to reader-focused methods." *IEEE Transactions on professional communication* 32.4 (1989): 238-255.

[27] Gunning, Robert. "The technique of clear writing." (1952).

[28] Spandel, Vicki. *Creating writers: Through 6-trait writing assessment and instruction*. Allyn & Bacon, 2005.

[29] Attali, Yigal, and Jill Burstein. "Automated essay scoring with e-rater® V. 2." The Journal of Technology, Learning and Assessment 4.3 (2006).

[30] Gunning, Robert. "The fog index after twenty years." *Journal of Business Communication* 6.2 (1969): 3-13.

[31] Flesch, Rudolf. "Flesch-Kincaid readability test." *Retrieved October* 26.3 (2007): 2007.

[32] Halliday, Michael Alexander Kirkwood, and Ruqaiya Hasan. *Cohesion in english*. Routledge, 2014.

[33] Mann, William C., and Sandra A. Thompson. "Rhetorical structure theory: Toward a functional theory of text organization." *Text-interdisciplinary Journal for the Study of Discourse* 8.3 (1988): 243-281.

[34] Walker, Marilyn A. *Centering theory in discourse*. Calrendon press, 1998.

[35] Grosz, Barbara, Aravind Joshi, and Scott Weinstein. "Centering: A framework for modeling the local coherence of discourse." *Computational linguistics* (1995).

[36] Tran, Nam Khanh, et al. "Topic cropping: leveraging latent topics for the analysis of small corpora." *Research and Advanced Technology for Digital Libraries: International Conference on Theory and Practice of Digital Libraries, TPDL 2013, Valletta, Malta, September 22-26, 2013. Proceedings 3*. Springer Berlin Heidelberg, 2013.

[37] Athinaios, Konstantinos, et al. *Named entity recognition using a novel linguistic model for greek legal corpora based on BERT model*. Diss. BS Thesis, School of Science, Department of Informatics and Telecommunications, 2020.

[38] Si, Luo, and Jamie Callan. "A statistical model for scientific readability." *Proceedings of the tenth international conference on Information and knowledge management*. 2001.

[39] Schwarm, Sarah E., and Mari Ostendorf. "Reading level assessment using support vector machines and statistical language models." *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL'05)*. 2005.

[40] Barzilay, Regina, and Mirella Lapata. "Modeling local coherence: An entity-based approach." *Computational Linguistics* 34.1 (2008): 1-34.

[41] Mesgar, Mohsen, and Michael Strube. "A neural local coherence model for text quality assessment." *Proceedings of the 2018 conference on empirical methods in*

natural language processing. 2018.

[42] Nguyen, Dat Tien, and Shafiq Joty. "A neural local coherence model." *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2017.

[43] Mesgar, Mohsen, and Michael Strube. "Lexical coherence graph modeling using word embeddings." *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2016.

[44] Barzilay, Regina, and Mirella Lapata. "Modeling local coherence: An entity-based approach." *Computational Linguistics* 34.1 (2008): 1-34.

[45] Fliessbach, Klaus, et al. "The effect of word concreteness on recognition memory." *NeuroImage* 32.3 (2006): 1413-1421.

[46] Brysbaert, Marc, and Boris New. "Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English." *Behavior research methods* 41.4 (2009): 977-990.

[47] Balota, David A., et al. "The English lexicon project." *Behavior research methods* 39 (2007): 445-459.

[48] Davies, Mark. "The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights." *International journal of corpus linguistics* 14.2 (2009): 159-190.

[49] Coltheart, Max. "The MRC psycholinguistic database." *The Quarterly Journal of Experimental Psychology Section A* 33.4 (1981): 497-505.

[50] Gregori, Lorenzo, et al. "CONCRETEXT@ EVALITA2020: The concreteness in context task." *CEUR WORKSHOP PROCEEDINGS*. Vol. 2765. CEUR, 2020.

[51] Basile, Valerio, et al. "Evalita 2020: Overview of the 7th evaluation campaign of natural language processing and speech tools for italian." *CEUR WORKSHOP PROCEEDINGS*. Vol. 2765. CEUR-ws, 2020.

[52] Rotaru, Armand Stefan. "ANDI@ CONCRETEXT: Predicting concreteness in context for English and Italian using distributional models and behavioural norms." *EVALITA Evaluation of NLP and Speech Tools for Italian-December 17th, 2020* (2020): 319.

[53] Montefinese, Maria, et al. "CONcreTEXT norms: Concreteness ratings for Italian and English words in context." *Plos one* 18.10 (2023): e0293031.

[54] Tanaka, Shinya, et al. "Estimating content concreteness for finding comprehensible

documents." *Proceedings of the sixth ACM international conference on Web search and data mining*. 2013.

[55] Paivio, Allan, John C. Yuille, and Stephen A. Madigan. "Concreteness, imagery, and meaningfulness values for 925 nouns." *Journal of experimental psychology* 76.1p2 (1968): 1.

[56] Richardson, John TE. "Imagery, concreteness, and lexical complexity." *The Quarterly Journal of Experimental Psychology* 27.2 (1975): 211-223.

[57] Proferes, Nicholas, et al. "Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics." *Social Media+ Society* 7.2 (2021): 20563051211019004.

[58] https://www.reddit.com/r/explainlikeimfive/

[59] https://www.reddit.com/

[60] https://www.reddit.com/dev/api/

[61] Roy, Sherre, Colin Beer, and Celeste Lawson. "The importance of clarity in written assessment instructions." *Journal of Further and Higher Education* 44.2 (2020): 143-155.

[62] L'Abate, Luciano. *Concreteness and specificity in clinical psychology: Evaluations and interventions*. Springer, 2015.

[63] Simm, William, et al. "Classification of short text comments by sentiment and actionability for voiceyourview." *2010 IEEE Second International Conference on Social Computing*. IEEE, 2010.

[64] https://www.mturk.com/

[65] Koo, Terry K., and Mae Y. Li. "A guideline of selecting and reporting intraclass correlation coefficients for reliability research." *Journal of chiropractic medicine* 15.2 (2016): 155-163.

[66] Ramos, Juan. "Using tf-idf to determine word relevance in document queries." *Proceedings of the first instructional conference on machine learning*. Vol. 242. No. 1. 2003.

[67] Marquardt, Donald W., and Ronald D. Snee. "Ridge regression in practice." *The American Statistician* 29.1 (1975): 3-20.

[68] Achiam, Josh, et al. "Gpt-4 technical report." *arXiv preprint arXiv:2303.08774* (2023).

[69] Reynolds, Laria, and Kyle McDonell. "Prompt programming for large language models: Beyond the few-shot paradigm." *Extended abstracts of the 2021 CHI conference on human factors in computing systems*. 2021.

[70] Miller, George A. "WordNet: a lexical database for English." *Communications of the*

ACM 38.11 (1995): 39-41.

[71] https://praw.readthedocs.io/en/stable/

[72] Joshi, Ankur, et al. "Likert scale: Explored and explained." *British journal of applied science & technology* 7.4 (2015): 396-403.

[73] Wu, Margaret, and Ray Adams. *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne: Educational Measurement Solutions, 2007.

[74] Ouyang, Shuyin, et al. "An empirical study of the non-determinism of chatgpt in code generation." *ACM Transactions on Software Engineering and Methodology* (2024).