# A Comparative Study on Instructional vs Role-based Prompt Engineering in NLP Tasks

Radia Riaz
*Military College of Signals*
radia.riaz986@gmail.com

Mavia Adil
*Military College of Signals*
maviaadilsiddiqui@gmail.com

Amna Alvie
*Military College of Signals*
amnaalviee@gmail.com

***Abstract*—Prompt engineering plays a crucial role in steering language model responses. This study compares two prominent prompt engineering styles—Instructional and Role-based prompts—within the context of NLP tasks. Using a dataset of 50 questions derived from IEEE Std 730-2014, AI-generated responses were evaluated based on six NLP metrics: BLEU, ROUGE-2, ROUGE-L, Cosine Similarity, Readability (Flesch Score), and Sentiment Polarity. In addition, a quantitative user study with 51 participants from diverse academic and professional backgrounds was conducted, where participants compared AI-generated responses framed by both prompt types. The evaluation focused on six qualitative metrics, including wording, sentence flow, clarity, tone, and overall preference. T**

***Index Terms*—Prompt Engineering, NLP Evaluation, Role-based Prompts, Instructional Prompts, BLEU, ROUGE, Cosine Similarity, Readability**

## I. Introduction

Large Language Models (LLMs), such as GPT-4, have demonstrated remarkable proficiency in generating coherent and contextually relevant responses, solidifying their role in a broad spectrum of natural language processing (NLP) applications. A key factor in harnessing their full potential lies in prompt engineering—the art of crafting inputs that effectively steer the model's outputs.

Among the many prompt engineering strategies, two widely adopted approaches stand out: **instructional prompting**, which provides the model with direct and explicit instructions, and **role-based prompting**, which assigns the model a contextual persona to simulate expert behavior or tone. While both have shown success in practice, existing studies often focus on anecdotal results or specific tasks, lacking a comprehensive, empirical comparison across a range of quality dimensions.

This research aims to bridge that gap by conducting a systematic evaluation of instructional and role-based prompts using a consistent LLM backend (GPT series) across diverse tasks. Unlike previous studies that either rely on task-specific outcomes or subjective judgments, our work incorporates a **multi-faceted evaluation framework**—including lexical overlap (BLEU, ROUGE), semantic similarity (cosine distance), sentiment neutrality, and readability (Flesch Reading Ease)—to quantify the performance of each prompting style.

Additionally, we manually curated **baseline answers** from authoritative IEEE sources to ensure reliable comparisons against model outputs. Our findings not only reveal nuanced differences between the two prompting paradigms but also offer practical guidance on selecting the most effective style based on the desired outcome. This research thus contributes to the growing field of prompt engineering by offering deeper insights and more generalizable conclusions than prior work.

## II. Literature Review

Prompt engineering has emerged as a critical technique for optimizing the performance of Large Language Models (LLMs) across a wide range of Natural Language Processing (NLP) tasks. As LLMs like GPT-4 continue to demonstrate exceptional versatility in generating human-like responses, the process of crafting effective input prompts has become essential for maximizing the utility of these models. Prompt engineering has garnered considerable attention in research, with various strategies explored to enhance the accuracy, fluency, and relevance of generated responses.

Sahoo et al. (2024) [4] conducted a systematic survey that categorizes prompt engineering techniques based on their applications. Their comprehensive review highlighted the strengths and weaknesses of various prompting strategies, offering a structured taxonomy that facilitates a deeper understanding of how different methods can be optimized for specific NLP tasks. This foundational work has contributed significantly to the classification of prompt engineering techniques and their practical applications.

Chen et al. (2023) [5] explored both foundational and advanced methodologies in prompt engineering, such as self-consistency and chain-of-thought prompting. Their study also delves into the security aspects of prompt engineering, addressing adversarial attacks that exploit vulnerabilities in the models. They presented strategies for mitigating these risks, which are crucial for ensuring the robustness and reliability of LLMs in real-world applications.

Vatsal and Dubey (2024) [6] offered a comprehensive survey of prompt engineering methods tailored for diverse NLP tasks. Analyzing 44 research papers, they reviewed 39 distinct prompting techniques across 29 NLP tasks, providing granular insights into the performance of these strategies on different use. Their work serves as an invaluable resource for understanding how different prompt engineering approaches perform in various task contexts and datasets.

White et al. (2023) [7] introduced a prompt pattern catalog for enhancing prompt engineering in ChatGPT interactions. Their framework documented reusable solutions for common

problems encountered during LLM interactions, providing developers and researchers with valuable tools to refine their prompt engineering processes and improve model outcomes.

While these studies offer a wide-ranging exploration of prompt engineering strategies, they focus primarily on general frameworks and qualitative analyses. A notable gap exists in the empirical, metric-driven comparison of different prompting styles, particularly in domain-specific question answering (QA) tasks. Most existing research lacks a rigorous comparison of Instructional and Role-based prompts, which are among the most widely used prompt styles in LLM applications.

Several studies have assessed the effectiveness of various prompting techniques in LLMs. For example, Liu et al. (2023) [1] explored the role of prompt tuning in improving model accuracy. Reynolds et al. (2021) [2] examined zero-shot and few-shot prompt tuning techniques, while Brown et al. (2020) [3] laid the groundwork for prompt-based interactions in LLMs. However, these works have largely focused on theoretical frameworks or qualitative comparisons rather than systematic evaluations based on linguistic and semantic metrics.

Our study addresses this gap by providing a head-to-head comparison of Instructional and Role-based prompt styles using a fixed dataset of technical questions from IEEE Std 730-2014. We assess the effectiveness of these prompting strategies using six robust NLP evaluation metrics: BLEU, ROUGE-2, ROUGE-L, Cosine Similarity, Flesch Readability Score, and Sentiment Polarity. This quantitative evaluation offers insights into the strengths and weaknesses of each prompt style from a linguistic and semantic perspective.

In addition to the metric-driven evaluation, our research also includes a **Google Forms-based user study**. This user-centered component of the study involved presenting a group of 51 participants with AI-generated responses based on both Instructional and Role-based prompts. Participants were asked to evaluate the responses across six qualitative metrics, including clarity, tone, sentence flow, and overall preference. This survey-based approach provides valuable insights into user preferences and perceptions of the two prompting strategies, complementing the metric-based findings and offering a holistic view of prompt engineering effectiveness.

By combining both quantitative NLP evaluation and qualitative user feedback, our study provides a comprehensive, empirical framework for comparing Instructional and Role-based prompt styles. This approach advances the field by offering actionable insights into how different prompt styles can be leveraged depending on task requirements and user preferences. Moreover, our work sets a precedent for future research that seeks to explore and refine prompt engineering techniques, with the potential to improve the accuracy, engagement, and overall effectiveness of LLMs in various applications.

## III. Methodology

The methodology adopted in this study employs a comprehensive, two-pronged evaluation approach: **Dataset-Based Evaluation** and **User-Based Evaluation**. This dual-strategy ensures that the performance of AI-generated content is measured not only through objective, reproducible metrics but also through subjective human-centered assessments. The overall objective is to examine how different prompt engineering techniques—namely, **Instructional Prompts** and **Role-Based Prompts**—impact the quality, readability, semantic richness, and perceived professionalism of AI-generated responses.

Following is the block diagram that summarizes the details of the methodology and evaluation flow.

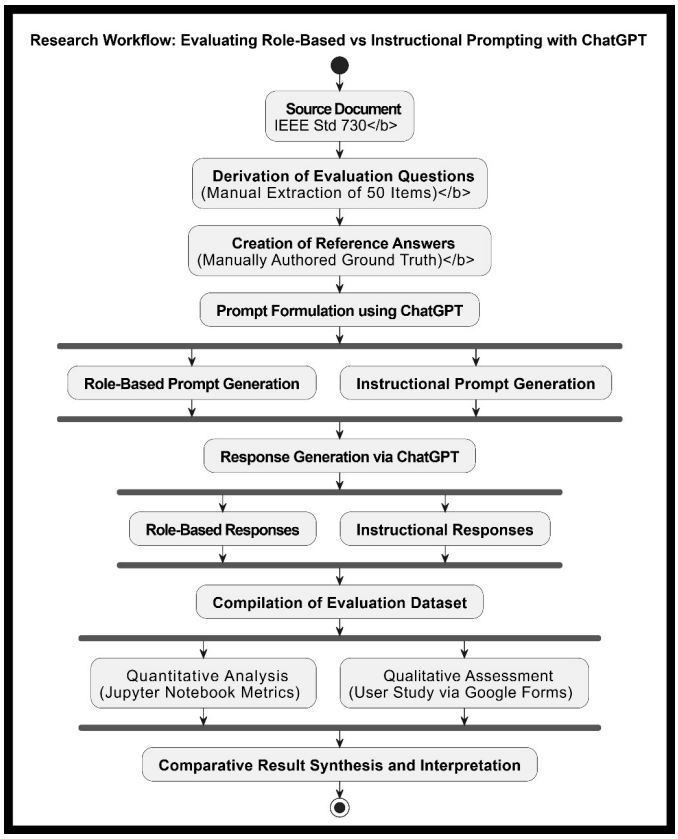


Fig. 1. Preference for Sentence Flow between Instructional and Role-Based Prompts

### A. Dataset-Based Evaluation

The Dataset-Based Evaluation focuses on the quantitative measurement of AI outputs against well-defined reference standards. By using computational linguistic metrics and content similarity measures, this phase aims to systematically evaluate the technical accuracy, language quality, and semantic relevance of AI responses across different prompting strategies.

*1) Dataset Creation:*

*a) Source Document Selection:* To ensure domain specificity and technical rigor, the dataset was constructed using content from the **IEEE Std 730-2014** standard, which pertains to software quality assurance processes. This document was selected for its thorough articulation of best practices and its authoritative standing within the software engineering community. By anchoring the dataset in a formal, structured

document, we ensure that the evaluation focuses on high-quality, non-trivial content that demands both factual accuracy and conceptual understanding from the AI.

*b) Question Formulation:* A total of **50 diverse and representative questions** were manually formulated from the IEEE standard. The question set was carefully curated to cover a wide spectrum of cognitive levels based on Bloom's taxonomy, ensuring a comprehensive assessment of the AI's capabilities. The questions targeted:

- **Factual Recall:** Simple retrieval of information stated explicitly in the document.
- **Conceptual Understanding:** Interpretation of ideas and the ability to explain underlying principles.
- **Application:** Adaptation of concepts to practical or hypothetical scenarios.

For each question, a **baseline answer** was manually prepared. These answers were extracted or paraphrased directly from the IEEE document to serve as gold standards for comparison. Special care was taken to ensure that the baseline answers maintained technical accuracy and adhered closely to the document's intended meaning.

*c) Sample Illustration:*

- *Example Question:* What are the main components of a software quality assurance plan?
- *Baseline Answer:* A comprehensive software quality assurance plan typically outlines the objectives, scope, applicable standards, policies, organizational structure, documentation requirements, review and audit activities, test strategies, and metrics for quality evaluation.

*2) Prompt Design:* To investigate the effect of prompt style on the AI's response quality, two prompt formats were systematically employed for each question:

- **Instructional Prompts:** Plain directives that request information or an explanation without any contextual framing.
- **Role-Based Prompts:** Context-rich prompts that position the AI as a domain expert or professional, thereby simulating a real-world communication scenario.

**Prompt Examples:**

- *Instructional Prompt:* Explain the components of a software quality assurance plan.
- *Role-based Prompt:* As a senior software quality analyst, explain the components of a software quality assurance plan to a new team member.

The rationale for introducing a role-playing dimension is based on existing research suggesting that contextual framing can lead to more detailed, coherent, and audience-appropriate language generation in AI systems.

*3) Response Generation:* For every question:

- One response was generated using the instructional prompt.
- A second response was generated using the corresponding role-based prompt.

Thus, the final dataset comprised **100 AI-generated responses** (50 from each prompting style). To minimize external biases, all responses were generated using the same AI model version, under consistent settings, ensuring a fair basis for comparative analysis.

*4) Evaluation Metrics:* A multi-faceted evaluation framework was adopted, utilizing a combination of lexical, semantic, and readability metrics to holistically assess response quality:

*a) Content Fidelity and Semantic Richness:*

- **BLEU Score:** Measures the degree of n-gram overlap between the AI response and the baseline answer. Though originally developed for machine translation evaluation, BLEU remains a strong indicator of content fidelity in factual responses.

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{1}$$

  Where:
  - $p_n$ = Modified precision for n-grams
  - $w_n$ = Weight for n-gram (usually uniform, e.g., 0.25 for BLEU-4)
  - BP = Brevity Penalty

  **Range:** 0 to 1
  **Good Score:** $> 0.5$
- **ROUGE Scores:**
  - **ROUGE-2:** Captures bigram overlap, providing insights into fluency and co-occurrence patterns.

$$\text{ROUGE-2} = \frac{\text{Number of overlapping bigrams}}{\text{Total bigrams in reference}} \tag{2}$$

    **Range:** 0 to 1
    **Good Score:** $> 0.5$ (A value above 0.5 indicates a strong overlap in phrasing and structure.)

  - **ROUGE-L:** Based on the longest common subsequence, assessing the structural similarity of sentence constructions.

$$\text{ROUGE-L} = \frac{\text{LCS}(X, Y)}{\text{Length of reference } Y} \tag{3}$$

    Where:
    * $\text{LCS}(X, Y)$ = Longest Common Subsequence between the generated text $X$ and reference text $Y$

    **Range:** 0 to 1
    **Good Score:** $> 0.5$ (Higher values indicate closer structural alignment between generated and reference sentences.)
- **Cosine Similarity (TF-IDF Vectorization):** Evaluates the semantic closeness between AI-generated and baseline answers by transforming the text into weighted term vectors and calculating angular similarity.

$$\cos(\theta) = \frac{\vec{A} \cdot \vec{B}}{\|\vec{A}\| \cdot \|\vec{B}\|} \tag{4}$$

Where:

- $\vec{A}, \vec{B}$ are TF-IDF vectors of generated and reference texts

**Range:** -1 to 1
**Good Score:** $> 0.7$

*b) Linguistic Style and Readability:*

- **Flesch Reading Ease Score:** Measures the readability of the generated text, with higher scores indicating simpler, more easily understandable language.

$$\text{FRES} = 206.835 - 1.015\left(\frac{\text{Total Words}}{\text{Total Sentences}}\right) - 84.6\left(\frac{\text{Total Syllables}}{\text{Total Words}}\right) \quad (5)$$

**Range:** 0 to 100
**Good Score:**

- 60–70: Easily understandable
- $> 70$: Very readable
- $< 50$: Complex

- **Sentiment Polarity:** Computed using TextBlob, this metric identifies the emotional tone of the response. Although factual writing should ideally be neutral, measuring sentiment ensures the absence of unintended bias or emotional coloring.
  **Range:** -1 (Negative) to +1 (Positive)
  **Good Score:** $\approx 0$ (Neutral for factual content)

*c) Tools and Libraries Employed:*

- `nltk`: For BLEU computation and basic text preprocessing.
- `rouge-score`: Python implementation for ROUGE metrics.
- `scikit-learn`: Used for TF-IDF vectorization and cosine similarity calculations.
- `textstat`: To compute Flesch reading ease scores.
- `textblob`: For sentiment polarity analysis.

*5) Data Collection and Analysis:* All metric outputs for the 100 generated responses were recorded in structured spreadsheets. Subsequent analysis involved:

- Aggregating mean and median scores across all evaluation dimensions for both prompt types.
- Conducting a paired *t-test* to statistically validate whether observed differences between instructional and role-based prompt responses were significant.
- Visualizing the results through comparative bar graphs, box plots, and trend lines to enable intuitive interpretation.

Wherever necessary, supplementary descriptive statistics such as standard deviation and confidence intervals were also calculated to ensure analytical robustness.

### *B. User-Based Evaluation*

In addition to the dataset-driven evaluation, a user-centered evaluation was conducted to gather qualitative insights into human preferences and perceptions regarding the two prompting styles: instructional and role-based. This component was designed to complement quantitative findings by analyzing subjective feedback from real users on the perceived quality of AI-generated responses.

*1) Participant Selection:*

*a) Demographic Characteristics:* To ensure diverse and balanced feedback, participant selection prioritized inclusivity across academic and professional backgrounds. The details are as follows:

- **Total Participants:** 51 individuals were successfully recruited and completed the evaluation survey.
- **Background:** The participant pool was diverse, consisting of undergraduate and graduate students from various academic disciplines, working professionals—particularly from the technology, business, and creative sectors—and AI hobbyists with differing levels of familiarity with artificial intelligence systems. This diversity helped to capture a broad spectrum of user perspectives.
- **Eligibility Criteria:** No prior expertise in AI, NLP, or prompt engineering was required. This allowed participants to evaluate responses based on their intuitive understanding, without technical bias.

Recruitment was carried out through university mailing lists, online academic and AI-focused communities. This approach helped to reach a wide demographic range from both academic and professional contexts.

*2) Survey Structure:* The survey was distributed using **Google Forms** and was titled: *Evaluation of Prompt Styles for Answer Quality*. Participants were informed of the survey's purpose and given a clear description of the two prompt types being compared:

- **Instructional Prompt:** A direct, task-based instruction.
- **Role-based Prompt:** A prompt given from the perspective of a specified role (e.g., a teacher or subject-matter expert).

Each participant was presented with multiple question blocks. For each block, the following components were shown:

- The original question.
- The correct reference answer (from the IEEE SQA documentation).
- The answer generated using the instructional prompt.
- The answer generated using the role-based prompt.

Participants were explicitly informed which answer corresponded to which prompting style. This means the evaluation was not blind or anonymized; participants were aware of the origin of each response.

For each set, participants answered a series of 6 evaluation questions, choosing which of the two responses they believed performed better across various quality dimensions:

1) **Better wording and sentence structure**
2) **Fluidity and naturalness of language**
3) **Accuracy and factual correctness**

4) **Readability**
5) **Appropriateness of tone for a professional context**
6) **Alignment with the reference answer**
7) **Overall preference**

*3) Data Collection and Analysis:* All survey responses were collected via **Google Forms**. The collected data was analyzed as follows:

- Calculating the percentage of preference for each prompt style across all six evaluation dimensions.
- Identifying patterns and dominant trends favoring one prompt style over the other.
- Visualizing the results using pie charts, stacked bar graphs, and histograms for clarity.

## IV. Results

### A. Dataset-Based Evaluation

This subsection presents a detailed evaluation of the generated responses using a range of automated metrics, comparing the effectiveness of Role-Based and Instructional prompting styles. Across multiple evaluation dimensions, Role-Based prompts consistently outperformed Instructional prompts in producing responses that were closer in quality and structure to human-curated references.

*1) Metric-Based Analysis:*

- **BLEU Score:**
  As illustrated in Figure 2, Role-Based prompts achieved a BLEU score of 0.064, substantially surpassing the 0.020 score obtained by Instructional prompts. This higher n-gram overlap suggests that Role-Based responses were more closely aligned with the wording and phrasing of the reference answers, demonstrating superior surface-level similarity and syntactic fidelity.

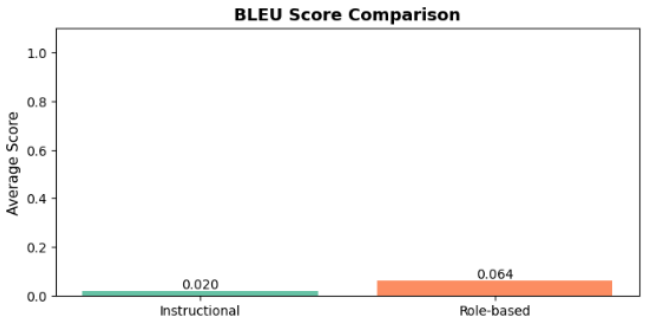


Fig. 2. BLEU Score Comparison between Role-Based and Instructional Prompts

- **ROUGE-2 and ROUGE-L Scores:**
  Consistent with the BLEU findings, Role-Based prompts attained higher ROUGE-2 (0.158 vs. 0.083) and ROUGE-L (0.319 vs. 0.227) scores compared to Instructional prompts, as shown in Figures 3 and 4. These metrics, which capture bigram overlap and longest common subsequence respectively, indicate that Role-Based responses not only retained more content elements but also better preserved the structural organization of the reference answers.

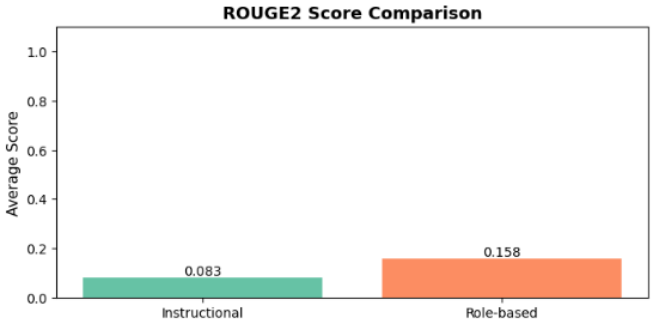


Fig. 3. ROUGE-2 Score Comparison

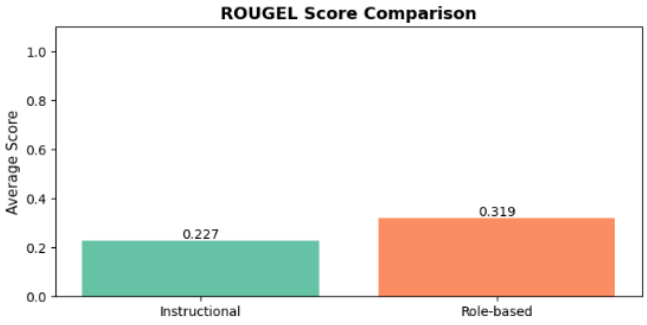


Fig. 4. ROUGE-L Score Comparison

- **Cosine Similarity:**
  A semantic similarity analysis, based on TF-IDF cosine similarity (Figure 5), further favored Role-Based prompts, which achieved a similarity score of 0.421 compared to 0.284 for Instructional prompts. This result suggests that Role-Based responses exhibited not only better surface matching but also deeper semantic alignment with the reference content.
- **Readability (Flesch Reading Ease Score):**
  An inverse trend was observed in terms of readability, as shown in Figure 6. Instructional prompts achieved a significantly higher Flesch Reading Ease score (34.517) compared to Role-Based prompts (3.764). This finding suggests that responses generated through Instructional prompting were considerably easier to read and comprehend, whereas Role-Based responses, likely due to their formal tone and complex structure, demanded greater cognitive effort from readers.
- **Sentiment Polarity:**
  Sentiment analysis, presented in Figure 7, revealed a

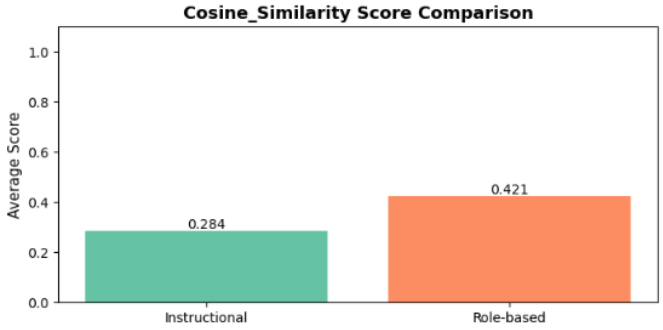


Fig. 5. Cosine Similarity Comparison between Role-Based and Instructional Prompts

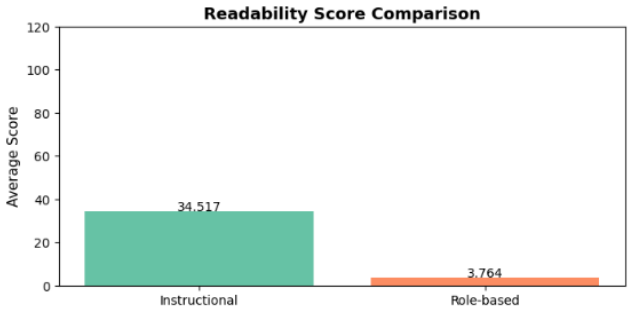


Fig. 6. Readability Score Comparison

slight positive bias in Role-Based responses (0.040) relative to Instructional responses (0.018). Although the difference is modest, it indicates that adopting a role or persona may infuse AI-generated content with a slightly more optimistic or engaging tone.

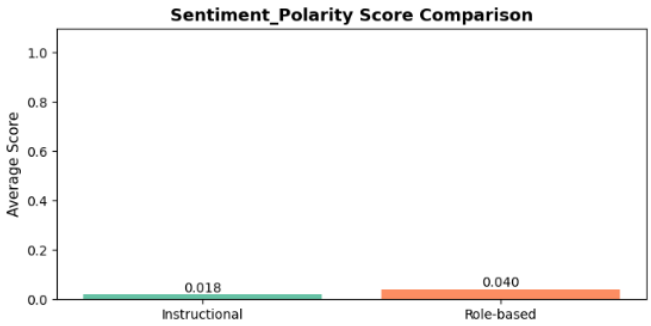


Fig. 7. Sentiment Polarity Score Comparison

*2) Key Observations:* The results consistently demonstrate that Role-Based prompting enhances both lexical overlap and semantic fidelity relative to Instructional prompting. However, this improvement is accompanied by a noticeable reduction in readability, suggesting that Role-Based outputs, while richer and more structurally aligned, require more cognitive effort to process. The slight increase in positive sentiment also points to a higher degree of human-likeness and engagement potential in Role-Based outputs.

Overall, these findings suggest that Role-Based prompting is particularly advantageous for expert-facing applications where depth, precision, and nuanced language are prioritized. In contrast, Instructional prompting, characterized by greater ease of readability, may be better suited for educational or beginner-facing contexts, where accessibility, clarity, and immediate comprehension are of primary importance.

### *B. User-Based Evaluation*

This subsection presents a comprehensive analysis of the User-Based Evaluation results, focusing on participant preferences between Instructional and Role-Based prompt styles across six critical evaluation dimensions. The insights derived from these findings offer a deeper understanding of how different prompt engineering strategies influence the perceived quality, clarity, and professionalism of AI-generated content. Furthermore, the evaluation highlights the strengths and potential trade-offs associated with each approach, enabling a more informed application of prompt engineering techniques in practice.

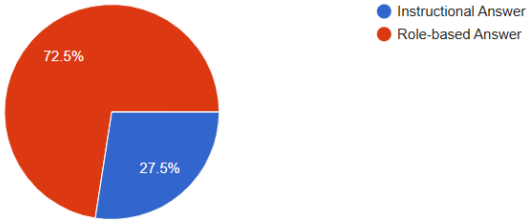


Fig. 8. Overall User Preference between Instructional and Role-Based Prompts

*1) Overall User Preference:* As illustrated in Figure 8, a substantial majority of participants (66.7%) expressed a preference for responses generated through Role-Based prompts, whereas 33.3% favored responses generated via Instructional prompts. This pronounced overall preference highlights the significant advantage offered by Role-Based prompting in producing outputs that users perceive as more engaging, natural, and contextually appropriate. Participants reported that responses generated through Role-Based prompting exhibited a conversational tone, improved relatability, and a more intuitive structure, closely mirroring real-world communication styles. These qualities collectively elevated the overall interaction quality, making Role-Based prompting a particularly effective strategy in enhancing user satisfaction and engagement.

*2) Dimension-wise Comparative Analysis:*

– **Word Choice**

In the word choice dimension (Figure 9), 74.5% of participants preferred Role-Based prompts, compared to only 25.5% who favored Instructional prompts. This suggests that Role-Based prompting

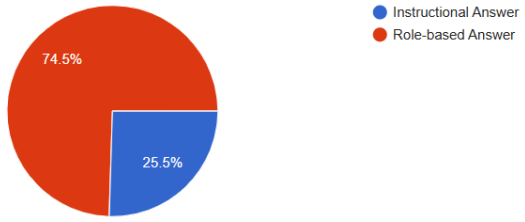


Fig. 9. Preference for Better Wording between Instructional and Role-Based Prompts

allows for more natural and effective wording, enabling the AI to communicate ideas in a clearer, more nuanced manner.

– **Sentence Flow**

As shown in Figure 10, 52.9% of participants

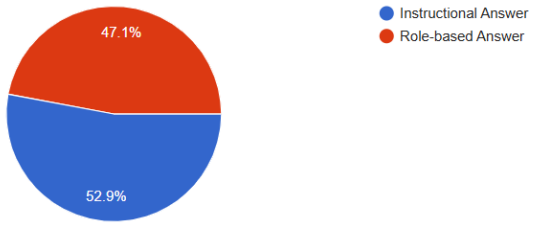


Fig. 10. Preference for Sentence Flow between Instructional and Role-Based Prompts

found Instructional prompts to have better sentence flow, while 47.1% preferred Role-Based prompts. Instructional responses were seen as slightly more logically structured and fluent, suggesting they may better suit tasks requiring step-by-step clarity.

– **Accuracy of Meaning**

In terms of best matching the original intended

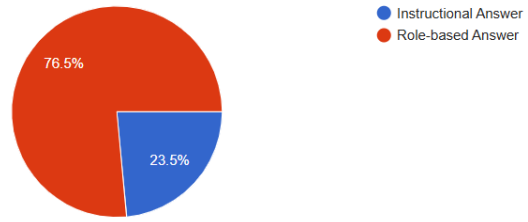


Fig. 11. Best Match for Original Intended Meaning

meaning (Figure 11), 76.5% of participants favored Role-Based prompts, with only 23.5% selecting Instructional ones. This highlights Role-Based prompting's strength in capturing and conveying users' intended messages with higher fidelity.

– **Readability**

In contrast to most other dimensions, 70.6% of participants found Instructional prompts easier to understand, while only 29.4% preferred Role-Based prompts for readability. This suggests that Instructional prompts may be more effective in contexts where clarity and simplicity are paramount.

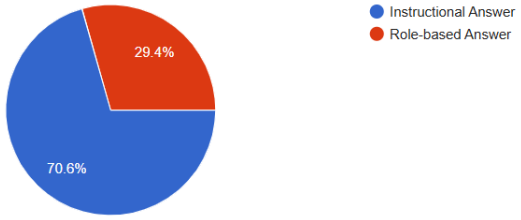


Fig. 12. Preference for Readability between Instructional and Role-Based Prompts

– **Professional Tone**

As depicted in Figure 13, 82.4% of participants

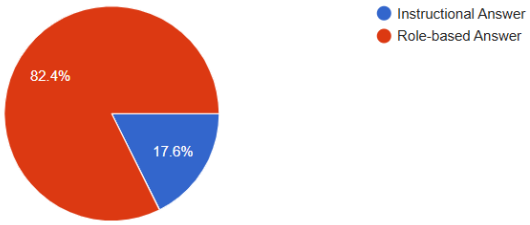


Fig. 13. Preference for Professional Tone between Instructional and Role-Based Prompts

favored Role-Based prompts for maintaining a more professional tone, while only 17.6% preferred Instructional prompts. The impersonation of expert roles appears to help the AI adopt more formal, respectful, and authoritative language.

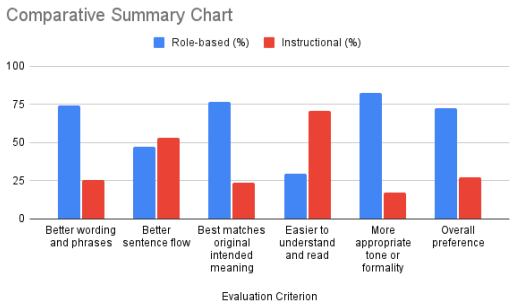


Fig. 14. Participant Preferences Across All Evaluation Dimensions

*3) Comparative Summary Across All Questions:* Figure 14 offers a consolidated summary of participant preferences across all dimensions. Overall, 72.5% of participants preferred Role-Based prompts, whereas 27.5% preferred Instructional prompts. These results show that Role-Based prompting generally provides more effective communication, especially in terms of tone, meaning accuracy, and word choice. However, Instructional prompts were notably favored in the readability and sentence flow dimensions.

*4) Key Observations:* The data from this user-based evaluation clearly indicate that Role-Based prompting delivers distinct advantages in most dimensions, especially in conveying professional tone, accurate meaning, and better word choices. These strengths make Role-Based prompts particularly useful in contexts requiring authoritative and nuanced communication.

Nevertheless, Instructional prompts demonstrated superiority in sentence flow and ease of understanding, suggesting they are better suited for contexts that prioritize clarity and instructional coherence.

In summary, Role-Based prompting is ideal when the goal is to produce professional, precise, and engaging content. However, for simpler, more instructional outputs focused on ease of reading and logical progression, Instructional prompting remains a reliable alternative. The choice between the two should be guided by the communication objectives at hand.

## V. Discussion

The comparative analysis between Role-Based and Instructional prompting styles reveals several key insights into their respective strengths and limitations. Role-Based prompting consistently outperformed Instructional prompting across most automated evaluation metrics, including BLEU, ROUGE-2, ROUGE-L, and cosine similarity. These results suggest that Role-Based prompts yield responses with greater lexical overlap and semantic similarity to human-curated references. Notably, the BLEU score for Role-Based prompting (0.064) was more than triple that of Instructional prompting (0.020), indicating improved n-gram alignment and surface-level resemblance to reference answers.

ROUGE-2 and ROUGE-L scores further supported this trend, with Role-Based prompting achieving values of 0.158 and 0.319, respectively, compared to 0.083 and 0.227 for Instructional prompting. These metrics highlight the enhanced content retention and structural alignment of Role-Based outputs. Cosine similarity results reinforced this semantic advantage, with Role-Based responses scoring 0.421 versus 0.284 for Instructional, reflecting stronger conceptual coherence and topical relevance.

However, this improvement in content fidelity comes at the expense of readability. Instructional prompts achieved a markedly higher Flesch Reading Ease score (34.517) than Role-Based prompts (3.764), indicating that while Role-Based responses are richer and more structurally faithful, they are also more complex and cognitively demanding to process. This trade-off is crucial when considering use-case scenarios, particularly in educational or novice-facing applications where simplicity and ease of comprehension are paramount.

Sentiment analysis added another dimension to the comparison, showing a slight positive polarity in Role-Based responses (0.040) compared to Instructional (0.018). Although modest, this suggests that Role-Based prompting may naturally adopt a more optimistic or engaging tone, potentially enhancing user experience in interactive or conversational AI settings.

User-based evaluations echoed the findings from automated metrics. A significant majority (66.7%) preferred Role-Based prompts overall, citing more natural, conversational tone and greater contextual relevance. In specific dimensions such as word choice, meaning accuracy, and professional tone, Role-Based prompting was favored by 74.5%, 76.5%, and 82.4% of users, respectively. These results affirm the ability of Role-Based strategies to deliver more refined, articulate, and authoritative responses. Nonetheless, Instructional prompting held an advantage in readability (70.6%) and sentence flow (52.9%), emphasizing its strength in producing clear, logically ordered responses. This suggests its continued relevance for scenarios demanding high accessibility and instructional clarity.

In summary, Role-Based prompting is highly effective for generating sophisticated, professional, and semantically rich content, making it ideal for expert-facing applications. Instructional prompting, on the other hand, is better suited for contexts prioritizing clarity, readability, and structured guidance. The choice between these prompting styles should therefore be guided by the intended audience and communicative goals of the application.

## VI. Future Work

While the current study provides valuable insights into the comparative effectiveness of Instructional and Role-Based prompts, several avenues remain open for further exploration. The findings suggest promising directions for enhancing the design and application of prompt engineering techniques. Below are a few key areas for future research and potential improvements.

### *A. Exploring Additional Prompt Styles*

Although this study focused on Instructional and Role-Based prompting styles, there exists a broad spectrum of other prompting strategies that could be explored in future work. For instance, experimenting with hybrid approaches that combine elements of both Instructional and Role-Based prompting might yield unique benefits, leveraging the strengths of each style. Investigating additional styles, such as Question-Based or Contextual prompts, could also reveal novel insights into how different approaches affect user engagement, clarity, and overall satisfaction.

### *B. Longitudinal Studies on User Interaction*

Future research could include longitudinal studies that track user preferences and behavior over time. By examining how user feedback evolves with extended interactions, researchers could gain a deeper understanding of the long-term effectiveness and adaptability of various

prompting styles. This type of research could shed light on how user expectations change in more complex or sustained conversational contexts, and how prompting strategies can be adapted for improved retention and performance.

### C. Expanding the Evaluation Dimensions

In the present study, six evaluation dimensions were examined: clarity, fluidity, accuracy, ease of understanding, professional tone, and overall preference. Future studies could expand this evaluation to include additional dimensions such as user trust, emotional engagement, or perceived creativity. Evaluating prompt styles based on a broader range of metrics would provide a more holistic view of their impact on user experience, helping to refine prompt engineering strategies for various domains.

## VII. Conclusion

This study conducted a comprehensive comparison between Role-Based and Instructional prompting strategies in the context of AI-generated content evaluation. Through both dataset-based and user-based assessments, Role-Based prompting consistently demonstrated superior performance across key dimensions such as lexical similarity, semantic alignment, clarity, fluidity, and professional tone. Despite a lower readability score, Role-Based outputs were more aligned with human-like communication in structure and tone, making them particularly well-suited for expert-facing and professional use cases.

User-based evaluation further validated these findings, with the majority of participants preferring Role-Based responses for their clarity, natural flow, and formal expression. While Instructional prompts maintained competitiveness in factual accuracy and ease of reading, their overall appeal was comparatively lower.

These results highlight the importance of aligning prompting strategies with application-specific goals. Role-Based prompting is ideal for scenarios that require contextual depth and user engagement, whereas Instructional prompting remains effective for tasks prioritizing simplicity and directness. Future research can explore hybrid prompting techniques that combine the strengths of both styles for optimized response generation.

## References


[1] P. Liu et al., "Pre-train Prompt Tune: Towards a unified paradigm for language model tuning," *arXiv preprint arXiv:2107.13586*, 2021.

[2] L. Reynolds and K. McDonell, "Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm," *arXiv preprint arXiv:2102.07350*, 2021.

[3] T. Brown et al., "Language Models are Few-Shot Learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877-1901, 2020.

[4] S. Sahoo et al., "A Survey on Prompt Engineering for Large Language Models: Taxonomy and Recent Advances," *arXiv preprint arXiv:2402.07927*, 2024.

[5] C. Chen et al., "A Comprehensive Survey of Prompt Engineering Methods and Security: Techniques, Applications, and Challenges," *arXiv preprint arXiv:2310.14735*, 2023.

[6] V. Shah and P. Dubey, "Prompt Engineering for NLP Tasks: A Comprehensive Survey," *arXiv preprint arXiv:2407.12994*, 2024.

[7] B. White et al., "Prompt Patterns: A Catalog to Enhance Prompt Engineering with ChatGPT," *arXiv preprint arXiv:2302.11382*, 2023.