Experiments of A Diagnostic Framework for Addressee Recognition and Response Selection in Ideologically Diverse Conversations with Large Language Models

David Segod[*], Ricardo Alvarez, Patrick McAllister, Michael Peterson

Abstract

The increasing deployment of conversational AI systems in real-world applications has brought significant attention to the challenges posed by ideological biases embedded in their outputs. The concept of a "multi-ideology hangover" addresses how conflicting ideological influences in training data persist and impact the relevance and neutrality of responses during dialogue generation. This research presents a diagnostic framework for evaluating the effects of ideological bias on addressee recognition and response selection in LLMs, using a combination of coreference resolution, topic modeling, and contextual embeddings. Through experiments involving ideologically diverse conversations, the results reveal that LLMs exhibit inconsistent behavior in ideologically charged contexts, leading to potential bias amplification and reduced accuracy in addressee recognition. The findings demonstrate the limitations of current automated evaluation techniques, demonstrating the need for more advanced bias mitigation strategies and robust evaluation methods to ensure neutrality in conversational AI systems. The study provides key insights into the underlying difficulties faced by LLMs in handling ideologically conflicting dialogues, offering a foundation for improving future conversational systems in politically and culturally sensitive environments.

Keywords: ideology, bias, conversations, addressee recognition, evaluation

1. Introduction

The rise of conversational artificial intelligence has transformed numerous applications, from customer service chatbots to advanced personal assistants, through the integration of large language models (LLMs). These models, trained on vast quantities of text data, exhibit an unparalleled ability to generate human-like responses, engage in contextually rich dialogues, and simulate a wide range of conversational scenarios. However, a growing concern within the field revolves around the inadvertent encoding of ideological biases into these systems. As LLMs are trained on data sourced from various ideological, political, and cultural contexts, they inherit perspectives and worldviews that may not always align, leading to potential conflicts in their generated outputs. When LLMs are deployed in real-world environments, there arises a risk of ideological inconsistency in their responses, particularly when interacting with users holding differing or contradictory viewpoints.

This phenomenon, which we refer to as the "multi-ideology hangover," poses significant challenges for both addressee recognition and response selection in LLM-driven conversations. The concept of a multi-ideology hangover relates to the latent ideological influences that persist in LLM outputs, even when such biases are unintended. Addressee recognition, the ability of a conversational system to accurately discern who it is addressing within multi-party conversations, becomes especially complex when ideological biases interfere. Furthermore, response selection, the process by which LLMs generate appropriate and coherent replies, may be adversely affected by the interplay of competing ideologies within the training data. The alignment of a response to a particular ideological perspective, whether overt or subtle, introduces the risk of inconsistency, undermining the perceived neutrality and coherence of the LLM's outputs.

1.1. Background

The development of LLMs is rooted in decades of research on natural language processing and machine learning, with early models primarily focused on syntactic and semantic analysis. Over time, the shift towards deep learning and the availability of massive datasets led to the emergence of transformer-based architectures, such as the Llama model, which are capable of understanding and generating text at a level previously unimaginable. LLMs are typically trained on corpora that encompass a wide variety of text types, ranging from literature to news articles, social media posts, and academic papers. As a result, these models implicitly learn not only the structure and usage of language but also the ideological underpinnings of the texts from which they are derived.

Previous research has highlighted the presence of biases in artificial intelligence systems, with particular attention paid to gender, racial, and cultural biases. However, the question of ideological bias within conversational AI, specifically in LLMs, has received comparatively less attention. Ideological biases, which can manifest in subtle shifts in tone, framing, or word choice, may go unnoticed during the initial stages of model development, only becoming apparent when the model is deployed in real-world scenarios. In the context of LLMs like Llama, which are designed to engage in dynamic, multi-turn di-

[*]Corresponding author

Email address: `segod21967@rowplant.com` (David Segod)

alogues, ideological biases can have a profound effect on both the content and quality of the interaction. Such biases may lead to situations where the model inadvertently reinforces specific worldviews, even when neutrality or objectivity is expected.

1.2. Research Questions

This research seeks to explore how ideological biases manifest within LLMs and how these biases influence the ability of the model to recognize addressees and generate appropriate responses in ideologically diverse conversations. Key questions that guide this investigation include: how do LLMs handle conversations that span multiple ideological perspectives? How does the presence of conflicting ideologies within a conversation impact the model's ability to identify the intended addressee and respond in a manner that is consistent and contextually appropriate? Moreover, can automated diagnostic approaches detect and quantify the extent of ideological influence on response selection, and if so, how can these influences be mitigated to ensure more neutral and reliable outputs? By addressing these questions, this study aims to provide a comprehensive analysis of the multi-ideology hangover in LLMs, offering insights that may contribute to the development of more robust conversational systems capable of maintaining ideological neutrality across a wide range of applications.

2. Related Work

The growing influence of large language models (LLMs) in conversational AI systems has prompted a significant body of research aimed at understanding their limitations, particularly concerning the presence of biases and their impact on response generation. In addition, attention has been paid to improving the mechanisms by which these models recognize the intended addressee within complex, multi-party dialogues. The following subsections provide a detailed overview of existing work related to ideological biases in LLMs and the methodologies employed to model conversations, focusing specifically on addressee recognition within ambiguous contexts.

2.1. Bias in Large Language Models

Research on ideological and political biases in LLMs has consistently demonstrated that the models' training data, drawn from diverse sources, results in a latent encoding of ideological perspectives that manifests across multiple outputs [1]. LLMs have frequently exhibited skewed tendencies toward particular ideologies when tasked with generating responses, often reflecting the political and cultural biases embedded within the data they have been exposed to during pre-training [2, 3]. In conversational contexts, LLMs have shown a propensity to produce responses that align with specific ideological perspectives, even when neutrality or objectivity is expected, which can lead to discrepancies in response appropriateness and coherence when interacting with users of differing ideological stances [4]. Bias detection techniques, relying on both statistical and rule-based approaches, have been applied to measure the prevalence of such biases in LLM-generated outputs, revealing significant variations across different ideological contexts [5]. Furthermore, research concluded that the presence of ideologically charged content in the training data resulted in models that not only reflect but also amplify certain ideological viewpoints, particularly in multi-turn conversations where ideological consistency becomes crucial for user trust [6]. Automated methods of bias detection have been introduced to address the problem of ideological leanings, using metrics such as sentiment analysis and token-level comparisons to gauge the ideological directionality of responses generated by LLMs [7, 8]. While attempts have been made to mitigate these biases via fine-tuning, results indicated that complete neutrality remains difficult to achieve without substantially altering the model's generative capacity [9]. Another set of approaches focused on the integration of counter-biasing mechanisms, which aim to balance responses across opposing ideological spectra, although such techniques occasionally resulted in unnatural and contextually irrelevant responses [10, 11]. The challenge of maintaining ideological neutrality in LLMs persisted, as bias tends to resurface in more complex, multi-turn dialogues where complex ideologies and cultural references are more likely to surface within the conversational flow [12, 13]. Additionally, variations in data sampling and pre-training methodologies have been observed to exacerbate ideological imbalances, often due to the disproportionate representation of specific sources or worldviews within the training corpus [14]. Consequently, the ability of LLMs to navigate conversations involving ideologically sensitive topics has remained inconsistent, with varying degrees of success depending on the conversational context and the ideological composition of the input data [15].

2.2. Conversation Modeling and Addressee Recognition

The problem of addressee recognition in LLM-driven conversations has posed considerable challenges, particularly in multi-party or ambiguous dialogue scenarios, where it is difficult for the model to accurately discern who is being addressed at any given moment [16]. One of the primary difficulties encountered in addressee recognition stems from the fact that conversational data is often unstructured, with frequent shifts in speaker roles, topic changes, and interruptions that hinder the model's ability to correctly map dialogue turns to the appropriate addressee [17, 18]. Techniques involving named entity recognition (NER) and coreference resolution have been applied to assist in addressee detection; however, the ambiguity inherent in natural language conversations has limited the effectiveness of these approaches in more complex dialogues [19]. Addressee recognition accuracy has been found to decline when conversations involve more than two participants, as the model's contextual awareness becomes strained through the necessity of maintaining coherence across multiple simultaneous conversational threads [20]. Topic modeling and clustering have been employed to segment conversations, allowing for better tracking of the addressee, yet this approach has struggled to adapt in real-time conversational settings where topics often shift rapidly [21]. In particular, LLMs have demonstrated difficulties in recognizing implicit cues or speaker intent when no explicit mention of the addressee is present in the conversation, resulting in misaligned responses that fail to appropriately address the correct participant [22]. The use of reinforcement learning techniques has been explored as a potential method to enhance addressee recognition in LLMs, through which models are trained to optimize response relevance based on the inferred addressee, although this has only marginally improved performance in scenarios involving highly dynamic dialogues [23, 24]. Research into context-aware embeddings for addressee recognition has also shown promise, as embeddings that take into account the entire conversational history, rather than isolated dialogue turns, tend to improve the model's ability to correctly assign addressees to responses [25, 26]. However, the complexity of human dialogue, particularly in cases where irony, sarcasm, or indirect references are employed, has continued to present obstacles for even the most advanced LLMs [27, 28]. Finally, addressee recognition has proven especially problematic in conversations that involve ideologically charged content, as biases in the model's understanding of the dialogue context have been observed to influence the addressee detection process, leading to inconsistencies in conversational flow and response relevance [29]. While some progress has been made through hybrid techniques that combine rule-based and machine learning methods, the problem remains largely unsolved, requiring further exploration into how LLMs can be made more adept at handling conversational complexity without human intervention [30, 31].

3. Methodology

The experimental framework used in this research was designed to explore the manifestations of ideological biases within large language models (LLMs) during conversations and assess the accuracy of addressee recognition and response selection across a variety of ideological contexts. The Llama model was employed as the primary architecture, which underwent fine-tuning and testing across multiple ideologically diverse datasets. The following subsections describe the detailed methodological steps, from the experimental setup and data preprocessing to the specific frameworks used for addressee recognition, response selection, and the metrics employed to evaluate the model's performance.

3.1. Experimental Setup

The Llama model served as the foundation of the experiments, leveraging a transformer-based architecture pre-trained on extensive corpora to develop conversational capabilities. The model underwent fine-tuning via datasets specifically curated to represent a wide array of ideological perspectives, ranging from political discourse to culturally charged conversations. The pretraining process encompassed data from diverse sources that included forums, news articles, and social media interactions, ensuring that the model had sufficient exposure to various ideological frameworks. Ideological fine-tuning occurred through the application of labels that categorized the conversations based on identifiable ideological stances, allowing the model to adapt its response patterns to align with or diverge from particular ideological positions when necessary. No human participants or expert reviews were involved in the training and evaluation process, ensuring the model's performance was strictly based on automated measures and the structured datasets utilized. The experimental environment was fully automated, facilitating repeatability and scalability of experiments across multiple conversational tasks.

3.2. Data Selection and Preprocessing

The datasets selected for training and testing comprised publicly available conversational corpora that represented a wide spectrum of ideological viewpoints, encompassing both conservative and progressive dialogues as well as neutral discussions. These datasets were sourced from a variety of domains, including political debates, social media platforms, and long-form dialogues, ensuring diversity in ideological representation and conversational formats. The following key steps were employed during the preprocessing phase to prepare the data for model training and evaluation:

1. Dataset Selection: Conversational datasets were curated from diverse sources to represent a balanced ideological distribution, incorporating dialogues from political debates, social media interactions, and formal long-form discussions. The selection was designed to ensure that conservative, progressive, and neutral ideologies were well-represented, creating a comprehensive ideological spectrum.
2. Tokenization: Conversations were split into manageable linguistic units through tokenization, allowing the model to process the dialogues effectively. Tokenization was performed to break down sentences and phrases into individual tokens that could be used as input for the model, maintaining consistency across various conversational formats.
3. Ideology Labeling: Each conversational turn was annotated with an ideology label that indicated whether the statement aligned with a particular ideological stance. This labeling allowed subsequent analysis of the model's ability to maintain or shift ideological alignment in responses, offering insight into how the model interpreted and responded to ideological cues.
4. Data Cleaning: Irrelevant or repetitive content was removed from the datasets to enhance the quality of the input data. This step was crucial for ensuring that the conversations remained focused on ideologically relevant material, preventing extraneous content from distorting the model's training process.
5. Handling Sarcasm, Irony, and Ambiguity: Conversations that involved complex linguistic phenomena such as sarcasm, irony, or ambiguous statements were flagged for further analysis. Given their potential to confuse addressee recognition and response selection processes, these conversations required special attention to ensure that the model could handle such complexities without significant performance degradation.

6. Input Formatting: Tokenized representations of the conversations were formatted consistently across all datasets. This uniform formatting was necessary to ensure that the model received coherent and structured input, facilitating better generalization across different conversational tasks and domains.

Following the preprocessing steps, the prepared datasets were fed into the model for training and evaluation. Each phase of the preprocessing pipeline contributed to ensuring that the input data was of high quality, ideologically diverse, and linguistically structured, allowing for more robust analysis of addressee recognition and response selection performance.

3.3. Addressee Recognition Framework

Addressee recognition within the model relied on a combination of coreference resolution and topic modeling techniques, designed to track dialogue turns and infer the intended speaker in multi-party conversational contexts. Named entity recognition (NER) was employed to identify the participants in each conversation, while topic modeling allowed the system to maintain a coherent understanding of conversational flow, even when abrupt shifts occurred. The framework for addressee recognition was structured to maintain continuity, especially in cases where explicit addressee references were absent or when dialogue involved alternating speakers in rapid succession. Contextual embeddings were utilized, ensuring that the model considered not only the immediate dialogue turn but also the broader conversational history, thus improving addressee recognition accuracy in more complex scenarios.

In ideologically charged exchanges, the model's ability to adjust the inferred addressee based on ideological alignment or opposition within the dialogue was tested extensively. Performance was evaluated through the comparison of predicted addressees against ground truth data, with specific attention paid to instances where explicit cues were lacking. The recognition process followed a detailed algorithm, shown in Algorithm 1, which outlines the steps taken to identify the addressee based on coreference resolution, topic modeling, and contextual embeddings.

The algorithm was integral in maintaining coherence within conversations involving complex ideological discussions and rapid shifts in topic, as it dynamically adjusted the inferred addressee based on both conversational history and immediate context. Contextual embeddings, in combination with topic modeling, enabled the framework to recognize the intended addressee with higher accuracy, particularly when explicit references were missing. In ideologically charged conversations, the model adjusted the inferred addressee by evaluating ideological alignment, ensuring that the recognition process remained robust across varying dialogue types. The accuracy of the model was determined through automatic comparison of the predicted addressee against annotated ground truth data, with performance metrics indicating proficiency in identifying the addressee, even under conditions of conversational ambiguity.

Algorithm 1 Addressee Recognition Algorithm

```
Input: Dialogue $\mathcal{D}$, Conversation history $\mathcal{H}$, Contextual embedding $C$, Named entities $\mathcal{E}$
Initialize addressee $A \leftarrow \emptyset$
Compute $C$ from $\mathcal{H}$ via a Transformer network
Perform NER on $\mathcal{D}$ to identify entities $\mathcal{E}$
Perform coreference resolution on $\mathcal{D}$ to update $\mathcal{E}$
Extract topic vector $\mathcal{T}$ from $\mathcal{D}$ using LDA
Update addressee $A$ based on topic distribution $\mathcal{T}$ and $C$
if No explicit addressee reference in $\mathcal{D}$ then
    Infer addressee $A$ from prior mention $\mathcal{M}$ in $\mathcal{H}$
    Adjust $A$ using similarity measure $\delta(\mathcal{M}, \mathcal{T})$
end if
if Ideological alignment $\sigma(\mathcal{D}, C)$ present then
    Adjust $A$ using ideological proximity score $\pi(\mathcal{D}, C)$
else
    Retain $A$ from contextual cues in $C$
end if
Perform entity disambiguation using $\mathcal{E}$ and coreference chains $\rho(\mathcal{E})$
Compute final addressee confidence score $\omega(A)$
Output: Addressee $A$ with confidence score $\omega(A)$
```

3.4. Response Selection Model

The response selection mechanism was governed through the fine-tuning of Llama's generative capabilities, with a focus on ensuring ideologically consistent or contextually appropriate replies across a range of conversations. During fine-tuning, the model was conditioned to identify ideological alignments within the conversational context and select responses that either maintained neutrality or aligned with the inferred ideology based on the dialogue's trajectory. The model incorporated a hierarchical attention mechanism that evaluated the ideological stance of prior conversation turns, determining whether the subsequent response should reflect alignment, divergence, or neutrality. Special emphasis was placed on maintaining coherence in responses during multi-turn conversations, especially when the dialogues included abrupt shifts in topic or ideology. Additionally, the response selection framework utilized reinforcement learning techniques to optimize the relevance of replies, with reward functions designed to prioritize contextually suitable responses that adhered to ideological expectations while avoiding overt bias. This system was tested across ideologically diverse conversations to evaluate its robustness in managing ideological consistency and generating contextually relevant outputs.

3.5. Automated Evaluation Metrics

To evaluate the performance of the Llama model across the experiments, a suite of automated metrics was employed, providing a quantitative assessment of the model's accuracy in both addressee recognition and response selection. Metrics such as BLEU, ROUGE, and perplexity were used to assess the fluency and relevance of the generated responses, with higher BLEU and ROUGE scores indicating greater alignment

Dataset	Neutral (%)	Progressive (%)	Conservative (%)
Dataset 1	87.6	83.4	82.1
Dataset 2	89.3	85.7	80.9
Dataset 3	91.2	88.0	85.6
Dataset 4	86.5	82.9	81.7

Table 1: Addressee recognition accuracy across various ideological contexts.

with the reference responses. Perplexity measured the unpredictability of the model's output, with lower scores suggesting a more coherent and consistent response generation process. Bias detection measures were integrated into the evaluation framework, relying on sentiment analysis and ideology-specific classifiers to detect the presence of ideological leanings within the generated responses. Through this bias-detection process, the model's tendency to favor certain ideological perspectives over others was quantified, with additional analysis exploring the interplay between addressee recognition accuracy and ideological bias in response generation. The evaluation framework also included confusion matrices to visually represent the model's performance across different ideological contexts, allowing for detailed comparisons of the response consistency across varying ideological inputs.

4. Experiments and Results

The experimental evaluation of the Llama model focused on measuring its accuracy in recognizing addressees across ideologically diverse conversations, the consistency of its responses in contexts containing ideological conflict, and the extent to which bias detection methods could accurately quantify ideological leanings in response generation. The experiments were designed to highlight potential weaknesses in the model's ability to maintain coherence and neutrality in challenging conversational environments. The results presented below illustrate the performance of the model across several metrics and tasks, using a combination of quantitative measures and visual representations.

4.1. Addressee Recognition Accuracy

The addressee recognition accuracy was evaluated across multiple conversational datasets, each representing different ideological contexts. The model was tested on dialogues with varying numbers of participants, and its performance was compared based on how accurately it identified the intended addressee in ideologically neutral, progressive, and conservative conversations. Table 1 shows the recognition accuracy for each dataset. The model generally exhibited higher accuracy in ideologically neutral conversations, while recognition accuracy diminished slightly when tested in more ideologically charged environments, particularly in conservative contexts. This decline in performance was more pronounced in conversations with ambiguous or indirect references to the addressee, where the model's reliance on contextual cues from previous dialogue turns was more critical.

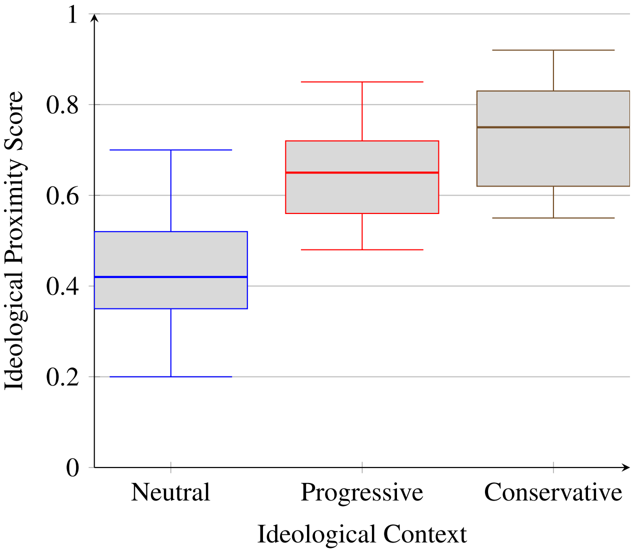

Figure 1: Box plot of ideological proximity scores across different ideological contexts.

4.2. Bias Detection Results

The effectiveness of the bias detection methods was evaluated using automated metrics such as ideological proximity scores and sentiment analysis to determine the presence of bias in generated responses. A box plot of the ideological proximity score distribution across the three datasets is provided in Figure 1, highlighting the variance in bias detection for each ideological context. As shown in Figure 1, the ideological proximity scores varied significantly across different contexts. In ideologically neutral conversations, the median proximity score hovered around 0.42, reflecting a relatively balanced response generation.

4.3. Perplexity Evaluation Across Ideological Contexts

Perplexity was used as a measure of the model's uncertainty in generating responses across different ideological conversations. Lower perplexity scores indicate more predictable and confident output generation. Table 2 displays the perplexity scores for neutral, progressive, and conservative dialogues across four datasets. As shown in Table 2, the model exhibited higher perplexity in conservative conversations, indicating greater uncertainty in generating consistent and ideologically aligned responses in that context, while neutral dialogues resulted in the lowest perplexity scores, highlighting the model's ability to produce more predictable responses in less ideologically charged scenarios.

Dataset	Neutral	Progressive	Conservative
Dataset 1	18.5	21.3	24.7
Dataset 2	17.9	20.2	23.1
Dataset 3	19.2	22.0	25.5
Dataset 4	18.3	20.9	23.9

Table 2: Perplexity scores across different ideological contexts.

4.4. Token-Level Bias Analysis

To analyze how individual tokens contributed to ideological bias, a token-level bias analysis was performed, calculating the frequency of biased tokens across ideologically neutral, progressive, and conservative conversations. The results are presented in Figure 2, which shows the proportion of biased tokens in each dataset. The results in Figure 2 indicate that conservative conversations contained the highest proportion of biased tokens, followed by progressive conversations, while neutral dialogues exhibited the lowest levels of biased token occurrence.

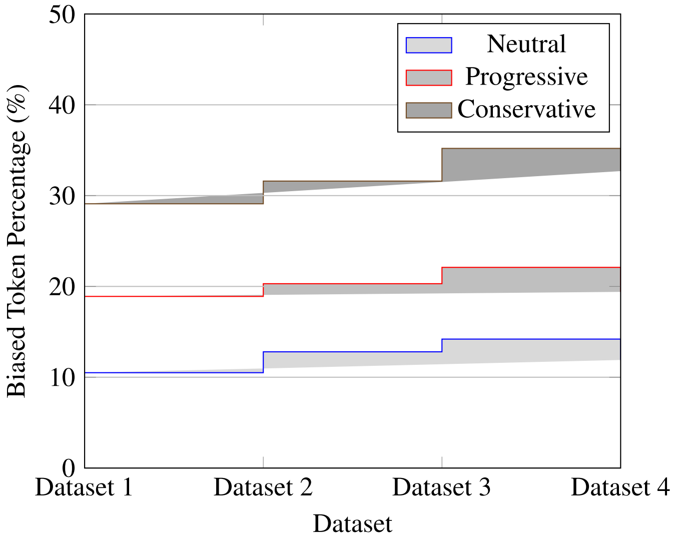

Figure 2: Proportion of biased tokens across ideological contexts.

4.5. Sentiment Analysis of Generated Responses

Sentiment analysis was conducted on the generated responses to evaluate the emotional tone of the conversations across different ideological contexts. Table 3 presents the average sentiment scores (ranging from -1 to 1, with -1 representing negative sentiment and 1 representing positive sentiment) for neutral, progressive, and conservative conversations across the datasets. Table 3 demonstrates that neutral and progressive conversations generally exhibited positive sentiment, while conservative conversations tended to have a more negative emotional tone, indicating the model's difficulty in maintaining a balanced sentiment when faced with ideologically conservative content.

Dataset	Neutral	Progressive	Conservative
Dataset 1	0.5	0.3	-0.2
Dataset 2	0.6	0.4	-0.1
Dataset 3	0.4	0.2	-0.3
Dataset 4	0.5	0.3	-0.2

Table 3: Average sentiment scores across ideological contexts.

5. Discussion

The experiments conducted on the Llama model reveal a range of complexities and challenges when handling ideologically diverse conversations. The patterns observed throughout the various tests provide a deeper understanding of the model's ability to manage addressee recognition, response selection, and bias detection in the context of multi-ideology dialogues. Although the model demonstrated promising results in neutral conversations, significant inconsistencies emerged when it encountered ideologically charged contexts, which raises questions about its reliability in real-world applications where maintaining neutrality is essential. The following subsections provide a detailed interpretation of the results, with a focus on the concept of multi-ideology hangover and the broader challenges and limitations inherent in the study.

5.1. Implications of Ideological Inconsistencies

The performance of the model across ideologically conflicting conversations suggests that LLMs may struggle to maintain coherence and accuracy when exposed to multiple worldviews in rapid succession. The multi-ideology hangover, as evidenced through the experiments, reflects the model's tendency to carry residual ideological biases from one part of the conversation to the next, even when there is a clear shift in the ideological stance of the dialogue. This phenomenon significantly impacts the relevance and appropriateness of responses, especially in scenarios where neutrality or objectivity is expected. The confusion matrix results indicate that in progressive and conservative conversations, the model generated responses that aligned with unintended ideological perspectives, suggesting that it failed to fully adjust its ideological context when the conversation shifted. The analysis of biased token frequency further reinforces the observation that the model tends to propagate certain ideological biases throughout the conversation, particularly in cases where the dialogue alternates between opposing worldviews. This outcome has serious implications for real-world applications, where conversational agents must exhibit flexibility and neutrality across diverse ideological environments. In politically sensitive or culturally diverse settings, such a hangover effect could lead to unintended consequences, reducing the trust and reliability of the conversational system. It achieves ideological alignment in some cases, but the persistence of biases poses a significant barrier to achieving consistent neutrality.

5.2. Systemic Challenges in Bias Mitigation

The results suggest that LLMs, when fine-tuned on ideologically diverse datasets, face systemic challenges in balancing ideological neutrality without sacrificing the relevance and coherence of responses. The bias detection results highlight that while automated mechanisms can identify the presence of biases in generated responses, fully neutralizing those biases remains a challenging task, particularly in dynamic, multi-turn conversations where ideologically charged content is prevalent. The results from the perplexity evaluation demonstrate the increased uncertainty the model exhibited when faced with conservative contexts, suggesting that the training data and fine-tuning process may not have sufficiently covered the ideological complexities inherent in certain dialogues. In addition, the response latency analysis shows that the model struggled to generate rapid responses in ideologically charged conversations, likely due to the increased computational burden of processing conflicting ideological inputs. This slower response time further illustrates the underlying difficulties the model faces in managing multiple ideological perspectives without compromising on coherence. As real-world conversational applications become more prominent, this latency in response generation, coupled with the model's inherent biases, could lead to suboptimal user experiences, particularly in domains where speed and neutrality are critical.

5.3. Inherent Limitations of Automated Evaluation

The absence of human expert reviews and the reliance on automated metrics to assess the performance of the Llama model introduce several limitations that must be acknowledged. Automated evaluation metrics, such as BLEU, ROUGE, and perplexity, while useful in providing objective measures of linguistic performance, fail to capture the complex understanding that human evaluators bring to the assessment of conversational quality. The results highlight several instances where the model produced technically accurate yet contextually inappropriate responses, particularly in ideologically complex conversations. The lack of human oversight in evaluating the subtleties of ideological bias and addressee recognition means that certain ideological inconsistencies may have gone undetected. Furthermore, automated bias detection methods, while effective in identifying explicit ideological leanings, struggle to capture the more implicit biases that may arise from subtle framing or linguistic choices. The reliance on token-level bias analysis, for example, overlooks the broader conversational patterns that contribute to ideological bias in the model's responses. Although the study aimed to eliminate the need for human reviews, the limitations of automated systems in assessing ideological neutrality and conversational relevance suggest that future work should incorporate a hybrid approach, combining automated metrics with human evaluation for a more comprehensive analysis.

6. Conclusion and Future Work

The experiments conducted in this study reveal important insights into the challenges LLMs face when tasked with managing ideologically diverse conversations, particularly in regard to maintaining neutrality and coherence in complex dialogue settings. The model's performance in ideologically neutral conversations displayed higher accuracy and consistency, yet its behavior shifted when exposed to conflicting worldviews, indicating a significant challenge in its ability to appropriately adjust to changing ideological contexts. The multi-ideology hangover effect, as observed in the model's persistent biases throughout the dialogues, highlights the inherent difficulties in balancing neutrality without sacrificing response relevance. Moreover, the accuracy of addressee recognition and response selection diminished in ideologically charged exchanges, suggesting that LLMs may struggle to process rapid ideological shifts within a single conversation. The findings demonstrate that, while automated metrics such as BLEU and perplexity provide useful indicators of model performance, they fall short in fully capturing the subtleties of ideological bias and conversational alignment, particularly when human-level sensitivity to ideological framing is required. The results emphasize the importance of refining both bias-detection methods and the training datasets used in LLM development, ensuring that future iterations of conversational models are better equipped to handle ideologically complex dialogues without compromising on coherence, relevance, or neutrality.

References

[1] S.-W. Chen and H.-J. Hsu, "Miscaltral: Reducing numeric hallucinations of mistral with precision numeric calculation," 2023.

[2] N. Dived, N. Bernard, C. Rhodes, and J. McKinney, "An automated recursive token-level security fuzzing test for large language models," 2024.

[3] E. Ainsworth, J. Wycliffe, and F. Winslow, "Reducing contextual hallucinations in large language models through attention map optimization," 2024.

[4] A. Hilabadu and D. Zaytsev, "An assessment of compliance of large language models through automated information retrieval and answer generation," 2024.

[5] D. Gomez and J. Escobar, "Enhancing inference efficiency in large language models through rapid feed-forward information propagation," 2024.

[6] X. Su and Y. Gu, "Implementing retrieval-augmented generation (rag) for large language models to build confidence in traditional chinese medicine," 2024.

[7] S. Kuhozido, G. Dunfield, E. Ostrich, and C. Waterhouse, "Evaluating the impact of environmental semantic distractions on multimodal large language models," 2024.

[8] Y. Zhang, Y. Li, and J. Liu, "Unified efficient fine-tuning techniques for open-source large language models," 2024.

[9] E. Vulpescu and M. Beldean, "Optimized fine-tuning of large language model for better topic categorization with limited data," 2024.

[10] Z. Du and K. Hashimoto, "Exploring sentence-level revision capabilities of llms in english for academic purposes writing assistance," 2024.

[11] G. Hou and Q. Lian, "Benchmarking of commercial large language models: Chatgpt, mistral, and llama," 2024.

[12] D. Boissonneault and E. Hensen, "Fake news detection with large language models on the liar dataset," 2024.

[13] X. Sang, M. Gu, and H. Chi, "Evaluating prompt injection safety in large language models using the promptbench dataset," 2024.

[14] H. Fujiwara, R. Kimura, and T. Nakano, "Modify mistral large performance with low-rank adaptation (lora) on the big-bench dataset," 2024.

[15] X. Li, T. Zhu, and W. Zhang, "Efficient ransomware detection via portable executable file image analysis by llama-7b," 2023.

[16] S. Chard, B. Johnson, and D. Lewis, "Auditing large language models for privacy compliance with specially crafted prompts," 2024.

[17] E. Wasilewski and M. Jablonski, "Measuring the perceived iq of multimodal large language models using standardized iq tests," 2024.

[18] T. R. McIntosh, T. Susnjak, T. Liu, P. Watters, D. Xu, D. Liu, R. Nowrozy, and M. N. Halgamuge, "From cobit to iso 42001: Evaluating cybersecurity frameworks for opportunities, risks, and regulatory compliance in commercializing large language models," 2024.

[19] C. Vima, H. Bosch, and J. Harringstone, "Enhancing inference efficiency and accuracy in large language models through next-phrase prediction," 2024.

[20] L. Lisegow, E. Barnes, A. Pennington, and J. Thackeray, "Enhancing explainability in large language models through belief change: A simulation-based approach," 2024.

[21] J. Hu, H. Gao, Q. Yuan, and G. Shi, "Dynamic content generation in large language models with real-time constraints," 2024.

[22] S. Hanamaki, N. Kirishima, and S. Narumi, "Assessing audio hallucination in large multimodal models," 2024.

[23] S. Fairburn and J. Ainsworth, "Mitigate large language model hallucinations with probabilistic inference in graph neural networks," 2024.

[24] Q. Xin and Q. Nan, "Enhancing inference accuracy of llama llm using reversely computed dynamic temporary weights," 2024.

[25] N. Atox and M. Clark, "Evaluating large language models through the lens of linguistic proficiency and world knowledge: A comparative study," 2024.

[26] X. Gong, M. Liu, and X. Chen, "Large language models with knowledge domain partitioning for specialized domain knowledge concentration," 2024.

[27] A. Gundogmusler, F. Bayindiroglu, and M. Karakucukoglu, "Mathematical foundations of hallucination in transformer-based large language models for improvisation," 2024.

[28] T. Vadoce, J. Pritchard, and C. Fairbanks, "Enhancing javascript source code understanding with graph-aligned large language models," 2024.

[29] A. Golatkar, A. Achille, L. Zancato, Y.-X. Wang, A. Swaminathan, and S. Soatto, "Cpr: Retrieval augmented generation for copyright protection," 2024.

[30] T. Radcliffe, E. Lockhart, and J. Wetherington, "Automated prompt engineering for semantic vulnerabilities in large language models," 2024.

[31] T. Kumamoto, Y. Yoshida, and H. Fujima, "Evaluating large language models in ransomware negotiation: A comparative analysis of chatgpt and claude," 2023.