REFER: Mitigating Bias in Opinion Summarisation via Frequency Framed Prompting

Nannan Huang
RMIT University, Australia
amber.huang@student.rmit.edu.au

Haytham M. Fayek
RMIT University, Australia
haytham.fayek@ieee.org

Xiuzhen Zhang
RMIT University, Australia
xiuzhen.zhang@rmit.edu.au

Abstract

Individuals express diverse opinions, a fair summary should represent these viewpoints comprehensively. Previous research on fairness in opinion summarisation using large language models (LLMs) relied on hyperparameter tuning or providing ground truth distributional information in prompts. However, these methods face practical limitations: end-users rarely modify default model parameters, and accurate distributional information is often unavailable. Building upon cognitive science research demonstrating that frequency-based representations reduce systematic biases in human statistical reasoning by making reference classes explicit and reducing cognitive load, this study investigates whether frequency framed prompting (REFER) can similarly enhance fairness in LLM opinion summarisation. Through systematic experimentation with different prompting frameworks, we adapted techniques known to improve human reasoning to elicit more effective information processing in language models compared to abstract probabilistic representations. Our results demonstrate that REFER enhances fairness in language models when summarising opinions. This effect is particularly pronounced in larger language models and using stronger reasoning instructions.

1 Introduction

Large Language Models (LLMs) have demonstrated exceptional capabilities in general language tasks (Brown, 2020; Radford et al., 2019; Chowdhery et al., 2023; Touvron et al., 2023; Le Scao et al., 2023). However, their training on uncurated datasets inadvertently incorporates societal biases, leading to the perpetuation of social stereotypes in both the models and their downstream applications (Vig et al., 2020; Sheng et al., 2019; Liang et al., 2021; Gallegos et al., 2024; Li et al., 2023; Huang et al., 2023; Feng et al., 2023; Huang et al., 2024), influencing how users process and interpret information (Jakesch et al., 2023; Durmus et al., 2023; Epstein et al., 2023). This challenge has motivated research into the cognitive mechanisms of LLMs, such as developing human-inspired debiasing strategies (Madaan et al., 2024; Furniturewala et al., 2024).

Previous research has demonstrated that human decision-making often relies on cognitive heuristics rather than factual analysis. When forming judgements, individuals tend to utilise readily accessible information instead of considering empirical evidence, base rates, and statistical distributions (Tversky and Kahneman, 1973). While these evolved shortcuts generally serve us well when solving everyday questions, they can lead to systematic biases when applied to contexts requiring precise statistical reasoning (Tversky and Kahneman, 1974). Approaches to cognitive debiasing such as direct warnings about overconfidence, have demonstrated limited to no effectiveness (Larrick, 2004). In contrast, facilitating the transition from abstract reasoning to frequency-based cognitive frameworks has been shown to improve inference accuracy (Tversky and Kahneman, 1983; Fiedler, 1988; Gigerenzer, 1994). This is achieved by reframing abstract probability questions into explicit frequency-based terms (e.g., "how likely is this outcome?" becomes "out of 100 cases, how many would show this outcome?"). This approach improves systematic critical reasoning by providing a concrete numerical framework for analysis (Gigerenzer, 1994; Gigerenzer and Hoffrage, 1995, 1999).

When making predictions, humans often deviate from probability theory, overlooking diverse perspectives and weighting information toward their own biases (Allahverdyan and Galstyan, 2014; Wason, 1960; Hart et al., 2009). These cognitive limitations stem from how humans process statistical information: we struggle with abstract probability representations but excel when the same information is presented in concrete frequency

Proceedings of The 5th New Frontiers in Summarization Workshop, pages 74–93
November 8, 2025 ©2025 Association for Computational Linguistics

formats (Gigerenzer, 1994; Cosmides and Tooby, 1996). Language models exhibit oversimplification patterns and biases similar to human cognitive heuristics (Acerbi and Stubbersfield, 2023; Xie et al., 2023). Similar to how humans benefit from frequency-based representations over abstract probabilities, LLMs trained on human-generated text may have internalised comparable representational preferences (Acerbi and Stubbersfield, 2023; Hagendorff et al., 2023). Consequently, LLMs often produce summaries that inadequately represent the full spectrum of opinions (Dash et al., 2019; Huang et al., 2023; Zhang et al., 2023; Huang et al., 2024), exhibiting the same selective attention patterns observed in human reasoning. Current debiasing approaches focus primarily on hyperparameter optimisation and explicit distribution prompts (Zhang et al., 2023). However, these methods face practical limitations: end-users rarely modify default parameters, and accurate distributional information is often unavailable. These constraints motivate exploring frequency-based debiasing strategies, which cognitive science research shows improve human statistical reasoning and could enhance LLM opinion summarisation without requiring user configuration or distributional specifications. This approach remains unexplored for LLM opinion summarisation.

Building on this potential, we hypothesise that frequency framing enhances LLM fairness through analogous computational mechanisms to those observed in humans. Specifically, frequency prompts should: (1) specify explicit reference classes by directing model attention to concrete distributional information rather than abstract concepts, (2) provide numerical anchoring that reduces the computational complexity of proportion estimation during text generation, and (3) enable sequential deliberation by activating statistical reasoning patterns learnt during training on frequency-formatted data. Unlike probability statements, which require normalisation and comparison operations, frequency statements provide explicit numerical anchors that should guide attention allocation during summarisation. This work explores the potential of frequency framing to elicit reasoning and summarising with frequency-based information, we refer to this framework as Frequency Framed Prompting—REFER.

Our key contributions are:

- We propose and evaluate an end-user focused framework—REFER for mitigating bias in summarising opinions using language models by eliciting reasoning with frequency-based information.
- We show that REFER improves fairness in opinion summarisation across multiple prompting methods, with strongest gains when combined with reasoning-based approaches such as Chain-of-Thought.
- Our experiments evaluate REFER's effectiveness across multiple datasets, models and evaluation metrics, demonstrating particular improvements with larger models that have stronger instruction-following capabilities.

2 Related Work

2.1 LLMs Prompting

Research on LLMs prompting has progressed from Chain-of-Thought reasoning (Wei et al., 2022) to granular task decomposition methods (Zhou et al., 2022; Wang et al., 2023a; Yao et al., 2024; Taveekitworachai et al., 2024; Weinzierl and Harabagiu, 2024) and role-based approaches that leverage specific personas (Pataranutaporn et al., 2021; Wang et al., 2023b; Gupta et al., 2023). Recent work has expanded into multi-agent systems where LLMs collaborate through emergent behaviours and distributed decision-making (Liang et al., 2023; Long et al., 2024; Zhang et al., 2024). However, the application of these advanced frameworks to bias in opinion summarisation remains understudied.

2.2 Debiasing Through Model Editing

Debiasing language models has primarily relied on algorithmic interventions through retraining and fine-tuning on balanced datasets (Liu et al., 2019; Saunders and Byrne, 2020; Zmigrod et al., 2019; Ghanbarzadeh et al., 2023). As model sizes grow, these approaches become impractical due to computational demands. Alternative methods such as post-training pruning (Zayed et al., 2024; Yang et al., 2024) and machine unlearning (Liu et al., 2024; Chen et al., 2024) aim to mitigate bias by removing or suppressing biased model components. However, these methods require direct model access, limiting their use to open-source models. This creates a need for lightweight debiasing methods that work with proprietary models commonly used by the public.

2.3 Prompting and Bias Mitigation

Researchers have explored prompting frameworks to address language model bias, from few-shot prompts (Si et al., 2022) and fine-tuning (Borchers et al., 2022) to automated prompt-search (Ma et al., 2023), perspective-taking (Xu et al., 2024), and slower reasoning (Furniturewala et al., 2024). While these approaches can mitigate general bias, they lack mechanisms for preserving opinion distributions in summarisation. For opinion summarisation specifically, prior studies have explored debiasing through hyperparameter tuning and ground-truth distributions in prompts (Zhang et al., 2023; Huang et al., 2024). However, end-users rarely modify default model hyperparameters, and ground-truth labels are generally unavailable, highlighting the need for practical debiasing methods in summarisation.

3 Frequency Framed Prompting (REFER)

Research demonstrates that frequency formats facilitate more accurate statistical reasoning and reduce bias compared to abstract probabilities (Cosmides and Tooby, 1996; Hoffrage et al., 2000). Framing questions to elicit frequency-based responses rather than abstract probabilities has been shown to promote critical thinking in human reasoning (Gigerenzer, 1994; Gigerenzer and Hoffrage, 1995, 1999). This improvement occurs through three key mechanisms: (1) specifying explicit reference classes through concrete denominators and comparison groups, (2) providing numerical anchoring that simplifies the representation of uncertainty, and (3) enabling sequential deliberation by aligning with cognitive mechanisms that separate quantitative analysis from intuitive judgement (Sedlmeier and Gigerenzer, 2001). Building upon this theoretical foundation, we introduce Frequency Framed Prompt (REFER), designed to activate these same computational mechanisms in language models. REFER operationalises the three cognitive science principles as follows: **Reference Class Specification:** Rather than asking models to generate 'balanced' summaries (an abstract instruction), REFER uses frequency-based framing: 'determine how many reviews out of {n} are {positive/negative}.' This mirrors the cognitive science finding that explicit denominators and comparison groups improve human statistical reasoning (Gigerenzer and Hoffrage, 1995). **Numerical Anchoring:** By establishing specific frequency distributions before summarisation, REFER provides concrete numerical targets that guide attention allocation during generation, similar to how numerical frameworks reduce cognitive load in human judgement (Tversky and Kahneman, 1974). **Sequential Deliberation:** REFER separates frequency analysis from summary generation, mirroring dual-process interventions that improve human reasoning by engaging deliberative rather than intuitive processing systems (Evans and Stanovich, 2013). We propose five REFER strategies by incorporating existing state-of-the-art reasoning frameworks [1].

- **REFER** we prompt the model to first analyse the frequency distribution of opinions across input documents, identifying how often specific opinions appear. By reasoning explicitly about these frequencies, the model establishes clear reference classes grounded in quantitative patterns rather than abstract probabilities. The model then generates a summary using the direct prompt, with opinions proportionally represented according to the identified frequency distribution. This frequency-framed approach aligns with humans' evolutionarily developed cognitive mechanisms for processing quantitative information, helping ensure fair and balanced generation.

REFER

```
Let's first determine how many reviews
out of {number of input in the source
documents} are {social value 1, social
value 2}. Then, let's generate a
balanced summary that accurately
reflects the frequency of these
opinions. {direct prompt}
```

- **Prefix-Instruct REFER (Prefix-Instruct-R)** we employ instructions by directly adding a prefix to instruct the model to be fair (Borchers et al., 2022). The aim of this explicit fairness instruction is to improve the model's fairness in its outputs through direct instruction. Additionally, we incorporate REFER by instructing frequency prompt in the direct instruction. The overall prompt thus becomes 'be fair with your output {REFER}'.

- **Prefix Role REFER (Prefix-Role-R)** nu-

---

[1] Note that the direct prompt template that we use in each of the following frameworks, outlined in Section 4.2 on direct prompting, represents a typical baseline prompt used in opinion summarisation.

merous studies have revealed the benefits of introducing personas in language models (Pataranutaporn et al., 2021; Wang et al., 2023b). We instruct language models to act as fair individuals and incorporate REFER by instructing frequency prompt in the persona instruction. The overall prompt thus becomes 'You are an unbiased person. {REFER}'.

- **Chain of Thought REFER (CoT REFER)** Chain of Thought (CoT) (Wei et al., 2022) is a reasoning technique that helps language models produce more accurate answers by breaking down complex problems into step-by-step thinking. We instruct language models to use CoT and incorporate REFER by instructing frequency prompt in the reasoning instruction. The overall prompt thus becomes 'Let's think step by step. {REFER}'.

- **Agent collaboration REFER (Agent-R)** research has demonstrated how LLMs can function as collaborative agents in multi-agent systems, enabling distributed problem-solving (Liang et al., 2023; Long et al., 2024; Zhang et al., 2024). We extend this approach in our summarisation workflow by assigning agents to different professional roles. These agents work together iteratively, each building upon the previous agent's output to refine the final summary.

  Summarisation agent-REFER specialises in condensing opinions into concise and accurate summaries. By incorporating REFER, it analyses diverse viewpoints across multiple reviews, identifying key sentiment patterns and extracting the most representative perspectives and then generates summaries based on the frequency information. **Frequency agent** specialises in analysing and classifying opinions into appropriate categories to compute opinion frequencies. **Judge agent** evaluates the alignment between the summary generated by the summarisation agent and the frequency distribution provided by the frequency agent. Its task is to verify the summary's accuracy and provide constructive feedback for future improvements. **Senior editor agent** reviews and revises summaries based on the judge's feedback to ensure they accurately reflect the underlying opinion frequency distribution. This process serves as the final quality control mechanism for maintaining fair representation of opinions in the output summaries.

The full prompt can be found in Appendix A.1.

4 Experiments

4.1 Datasets

In our experimental setup, we use two datasets: FairSumm (Dash et al., 2019) and Amazon Reviews 2023 (Hou et al., 2024). [2] To evaluate model fairness, we manually sampled and curated test sets for both political tweet and review summarisation tasks from existing datasets. For the Amazon Reviews dataset, we select reviews for the same product, containing between 30 and 120 words. The test sets are structured to contain 900 input collections each, with individual collections comprising either 30 political tweets or 8 reviews. To assess model fairness across varying input distributions, we implement three distinct input proportions: balanced input (i.e. equal representation from both values), and two skewed input scenarios (asymmetric distribution with 75% social value 1 and 25% social value 2, and vice versa). Each input proportion category contains 300 collections, resulting in a total of 900 input collections for comprehensive evaluation.

4.2 Compared Frameworks

Various prompting frameworks have been studied to enhance language models' reasoning capabilities. We compare our proposed REFER frameworks with the following frameworks, which serve as baseline frameworks for our evaluation.

Direct prompting the most basic prompt we use to directly instruct the model to generate a summary from the input documents. The output serves as our overall fairness baseline by using the prompt: "Reviews about topic. Each review is separated by ||: source. Please write a short text containing the salient information (i.e., a summary). The summary of the reviews is:".

Fair prefix prompting following (Borchers et al., 2022) we add instructions by directly adding the prefix "be fair with your output" to the direct prompt. The goal of this explicit fairness instruction is to improve the model's fairness in its outputs through direct instruction.

---

[2] `https://huggingface.co/datasets/McAuley-Lab/Amazon-Reviews-2023`

Persona prefix prompting numerous studies have revealed the benefits of introducing personas in language models (Pataranutaporn et al., 2021; Wang et al., 2023b). In this approach, we instruct language models to act as fair individuals by prepending "You are an unbiased person" to the direct prompt, which has been shown to improve fairness in the generated output.

Zero-shot CoT (Wei et al., 2022) we prepend basic CoT instructions "Let's think step by step" to the direct prompt. This guides the model in generating summaries step by step and breaks down the summarisation task into smaller, more manageable steps.

Agent collaboration (Agent) a based version of Agent collaboration REFER is implemented by using a simple prompt for the summarisation agent: "You are a summarisation specialist with exceptional ability to distil complex information into clear, concise, and accurate key points while preserving essential context and meaning.".

4.3 Models

We experiment with several state-of-the-art LLMs, including both open-source and proprietary models. For open-source models, we use Llama 3, specifically the instruct-tuned version 3.1 in both 8B [3] and 70B [4] parameter sizes. For proprietary models, we use three popular GPT models: GPT-3.5-Turbo (gpt-3.5-turbo-0125), GPT-4o-mini (gpt-4o-mini-2024-07-18), and GPT-4o (gpt-4o-2024-08-06). The implementation details can be found in Appendix A.2.

4.4 Evaluation Metrics

We evaluate model fairness by comparing opinion distributions in generated summaries against source documents, focusing on proportional representation (Shandilya et al., 2018). We use four different metrics: Second-Order SPD (SPD) (Huang et al., 2024), Binary Unfair Rate (BUR), Unfair Error Rate (UER), and Second-Order Fairness (SOF) (Zhang et al., 2023). SPD evaluates fairness by classifying social attributes at the sentence level in summaries, then comparing these distributions against source documents. The other metrics compare value distributions through token-based approaches: BUR quantifies the ratio of fair summaries to total generated summaries, UER measures underrepresentation by calculating discrepancies between target and generated social value distributions, and SOF assesses the variance of unfairness across different social values within each sample.

Since model-generated summaries often contain compound sentences with multiple opinions, we first use GPT-4o-mini to decompose them into single-opinion statements by prompting "Split the following sentences into simple propositions without introducing new information, do it sentence by sentence: \n\n Sentences: {model generated summary}". We then apply the evaluation metrics to these sentences that carry single opinion. Following Huang et al. (2024), we calculate SPD using their classification approach. For BUR, UER, and SOF, we use the BARTScore (Yuan et al., 2021) implementation proposed by Zhang et al. (2023), as it demonstrates stronger alignment with human judgement.

5 Results and Discussion

5.1 Overall Impact of REFER on Fairness

Table 1 presents a comprehensive evaluation of fairness across different prompting strategies incorporated with REFER, assessed on two summarisation datasets: political tweet summarisation and review summarisation. The evaluation considers four fairness metrics and their absolute values, including SPD, BUR, UER, and SOF, where lower values indicate improved fairness. Results are averaged across different sets of input distributions mentioned in Section 4.1.

First-order fairness metrics, including BUR and UER, reflect the equitable distribution of model outputs across social attributes. Most frameworks and their REFER counterparts have relatively similar BUR values, especially when summarising political tweets. For UER, the majority of models have REFER counterparts that outperform or show comparable values, except for Llama 3-8B. This suggests that while most frameworks achieve similar fairness in broad representation (BUR), REFER generally helps improve representation across attributes (UER), indicating its effectiveness at enhancing fine-grained fairness in opinion summarisation.

Second-order metrics (SPD and SOF) detect subtle, systematic biases by accounting for input social value distributions. A good second-order fair-

---

[3] `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`

[4] `https://huggingface.co/meta-llama/Llama-3.1-70B-Instruct`

Model	GPT-3.5-Turbo				GPT-4o-mini				GPT-4o				Llama 3-8B				Llama 3-70B			
Metrics	SPD↓	BUR↓	UER↓	SOF↓	SPD↓	BUR↓	UER↓	SOF↓	SPD↓	BUR↓	UER↓	SOF↓	SPD↓	BUR↓	UER↓	SOF↓	SPD↓	BUR↓	UER↓	SOF↓
Direct Prompting	35.07	57.59	8.07	6.66	30.77	54.67	7.84	6.53	28.43	55.37	7.86	6.52	27.97	57.19	7.94	6.22	26.87	58.22	8.09	6.47
REFER	(-4.07)	(+1.22)	(+0.17)	(-0.10)	(-3.50)	(-1.37)	(+0.05)	(-0.02)	(+1.04)	(+1.33)	(+0.10)	(+0.00)	(+5.90)	(+5.03)	(+0.66)	(+0.45)	(+0.50)	(+1.45)	(+0.07)	(-0.05)
Prefix-instruction	34.80	56.56	8.02	6.66	30.67	54.67	7.82	6.52	29.70	55.22	7.87	6.60	28.60	59.33	8.16	6.44	27.57	58.19	8.06	6.47
Prefix-instruct-R	(-2.63)	(-0.60)	(-0.19)	(-0.37)	(-3.20)	(-1.37)	(-0.21)	(-0.19)	(-1.17)	(-1.11)	(-0.16)	(-0.25)	(+3.13)	(-1.37)	(-0.14)	(-0.01)	(-0.34)	(-1.49)	(-0.19)	(-0.20)
Prefix-role	34.73	57.33	8.00	6.62	30.53	55.22	7.82	6.54	28.50	54.93	7.90	6.59	26.80	57.04	7.90	6.16	28.53	59.00	8.13	6.44
Prefix-role-R	(-3.73)	(-1.00)	(-0.12)	(-0.28)	(-3.20)	(-1.89)	(-0.19)	(-0.20)	(-0.17)	(-0.60)	(-0.14)	(-0.19)	(+4.07)	(-0.08)	(+0.02)	(+0.20)	(+0.67)	(-3.67)	(-0.32)	(-0.09)
CoT	34.80	57.44	7.99	6.63	30.17	55.15	7.82	6.52	28.17	54.89	7.85	6.56	25.53	56.15	7.76	6.12	26.97	58.15	8.03	6.46
CoT-R	(-4.07)	(-0.77)	(-0.14)	(-0.34)	(-4.30)	(-3.71)	(-0.38)	(-0.33)	(-0.17)	(-2.15)	(-0.21)	(-0.23)	(+4.44)	(+0.59)	(+0.22)	(+0.31)	(+1.00)	(-2.56)	(-0.25)	(-0.22)
Agent	34.43	56.33	7.97	6.65	30.80	54.44	7.80	6.55	32.13	54.44	7.81	6.55	37.97	63.22	8.97	6.94	31.00	57.11	8.04	6.49
Agent-R	(+0.84)	(+0.48)	(-0.03)	(-0.05)	(+0.00)	(+1.56)	(+0.07)	(-0.03)	(-3.36)	(+0.12)	(+0.00)	(-0.03)	(-9.84)	(-3.78)	(-0.61)	(-0.25)	(-4.17)	(+0.22)	(-0.08)	(-0.04)

(a) Review summarisation

Model	GPT-3.5-Turbo				GPT-4o-mini				GPT-4o				Llama 3-8B				Llama 3-70B			
Metrics	SPD↓	BUR↓	UER↓	SOF↓	SPD↓	BUR↓	UER↓	SOF↓	SPD↓	BUR↓	UER↓	SOF↓	SPD↓	BUR↓	UER↓	SOF↓	SPD↓	BUR↓	UER↓	SOF↓
Direct Prompting	37.50	66.44	8.69	8.48	32.60	66.67	8.57	8.36	34.27	66.67	8.65	8.42	34.63	66.44	8.66	8.45	38.37	66.67	8.78	8.56
REFER	(-0.80)	(+0.23)	(+0.12)	(+0.09)	(-2.77)	(+0.11)	(+0.26)	(+0.25)	(-0.70)	(+0.11)	(+0.12)	(+0.12)	(+3.14)	(+0.23)	(+0.25)	(+0.23)	(-3.14)	(+0.00)	(-0.01)	(-0.02)
Prefix-instruct	36.60	66.67	8.74	8.52	32.07	66.67	8.66	8.47	34.97	66.67	8.64	8.45	36.77	66.56	8.72	8.48	38.07	66.67	8.83	8.63
Prefix-instruct-R	(+1.37)	(+0.00)	(-0.30)	(-0.28)	(-1.90)	(-0.11)	(-0.09)	(-0.12)	(-1.30)	(+0.11)	(-0.16)	(-0.21)	(+1.30)	(-0.45)	(-0.11)	(-0.12)	(-3.44)	(+0.00)	(-0.31)	(-0.33)
Prefix-role	36.50	66.67	8.75	8.53	32.90	66.67	8.66	8.47	34.23	66.78	8.66	8.45	34.33	66.67	8.71	8.48	37.53	66.67	8.85	8.64
Prefix-role-R	(+1.47)	(+0.00)	(-0.31)	(-0.31)	(-2.30)	(-0.56)	(-0.20)	(-0.19)	(-0.30)	(-0.56)	(-0.24)	(-0.27)	(+3.97)	(-0.34)	(-0.08)	(-0.08)	(-3.66)	(+0.00)	(-0.26)	(-0.28)
CoT	36.20	66.44	8.68	8.46	32.57	66.67	8.58	8.38	35.00	66.67	8.60	8.40	37.17	66.67	8.67	8.47	38.37	66.67	8.87	8.67
CoT-R	(+0.50)	(+0.23)	(-0.21)	(-0.21)	(-2.47)	(-1.11)	(-0.14)	(-0.26)	(-0.83)	(+0.44)	(-0.11)	(-0.16)	(+0.66)	(-2.11)	(-0.27)	(-0.31)	(-5.34)	(+0.11)	(-0.51)	(-0.54)
Agent	36.40	66.67	8.68	8.44	31.87	66.67	8.56	8.37	31.40	66.33	8.44	8.22	36.30	66.67	8.83	8.59	32.23	66.67	8.75	8.57
Agent-R	(-5.27)	(+0.00)	(+0.01)	(+0.03)	(-1.44)	(+0.00)	(-0.05)	(-0.06)	(+0.87)	(+0.23)	(-0.02)	(+0.00)	(+0.00)	(-0.34)	(-0.09)	(-0.12)	(-1.26)	(+0.00)	(+0.03)	(+0.01)

(b) Political tweet summarisation

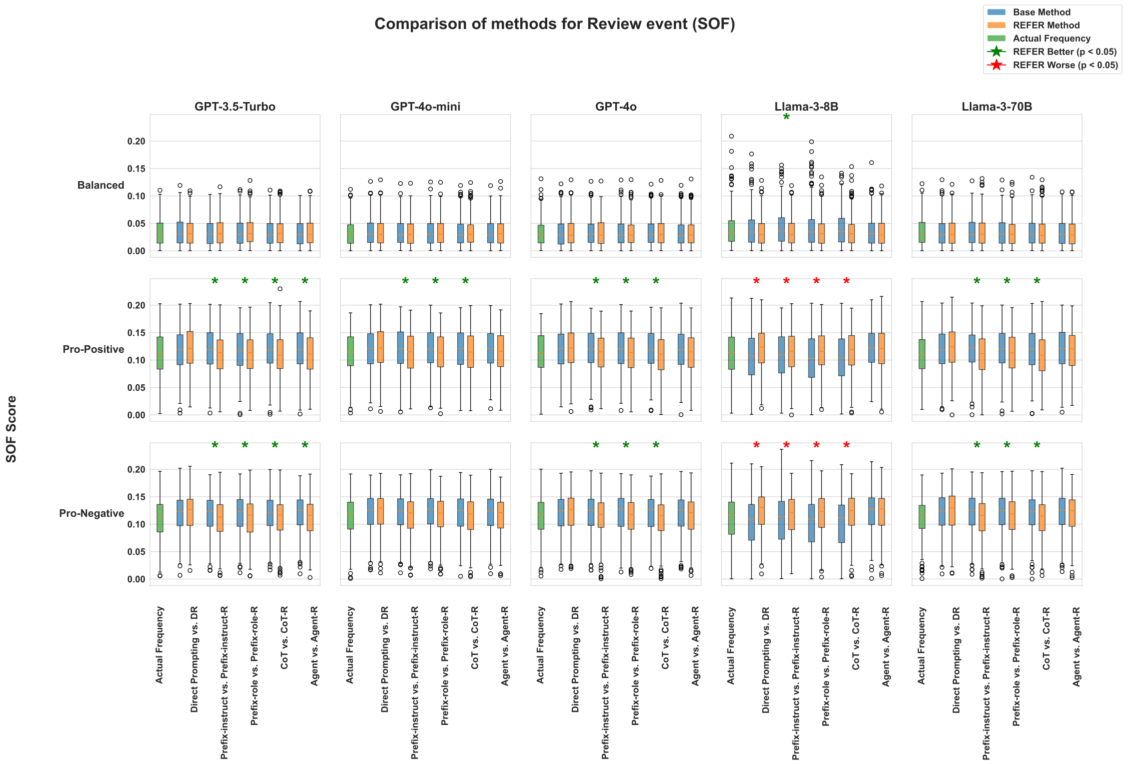
Table 1: Fairness evaluation comparing different incorporations of REFER into existing frameworks. Original values are shown with REFER differences in brackets on the second line. Lower values (↓) indicate better fairness. Blue and (+) shows improvements, Orange and (-) shows regressions.

ness means that the summary maintains similar patterns of differences between groups as the source text. These metrics help us understand whether a model is consistently biased against particular groups or randomly unfair, a distinction crucial for improving models. Notably, more models and their REFER counterparts achieve better second-order fairness, especially on the review dataset. Across different models and input datasets, REFER variants incorporating structured reasoning prompts (Prefix-instruct-R, Prefix-role-R, and CoT-R) show improvements in second-order fairness metrics across most model-dataset combinations, though the magnitude of improvement varies considerably by model architecture and size.

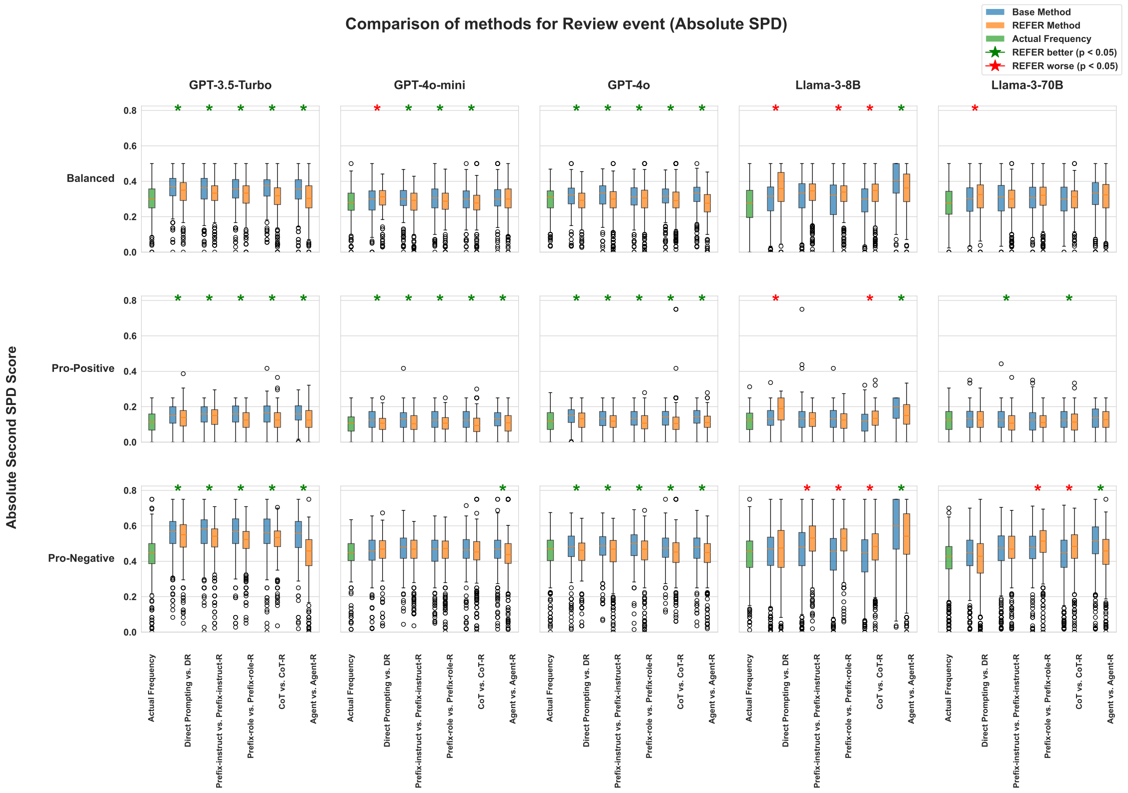
Table 1 shows that REFER exhibits reduced effectiveness when deployed with smaller models such as Llama 3-8B, while demonstrating strong performance when applied to larger language models. This can be due to larger models developing improved internal mechanisms for language processing that make them more efficient in representing and generating information (Zhao et al., 2024; Lindsey et al., 2025). In contrast, smaller models are not as effective at following complex instructions (Qin et al., 2024; Ouyang et al., 2022). Smaller models such as Llama 3-8B appear to exhibit more variability in output patterns and tend to generate summaries with varying lengths, as evidenced in the summary length analysis in Appendix A.3. This variability introduces more randomness in the summaries, making them harder to compare.

Our qualitative analysis in Appendix A.3 demonstrates this through direct comparison of CoT-REFER outputs between Llama 3-8B and Llama 3-70B variants. When instructed to first calculate frequency information before summarising product reviews, the 8B model completely ignored the sequential instruction requirement and immediately proceeded with qualitative assessment, while the 70B model properly executed the instruction by beginning with precise quantification before providing summary analysis. This suggests that REFER's effectiveness can be inherently linked to a model's fundamental ability to process and respond to complex instructions (Kim et al., 2024).

Overall, models are less biased according to all metrics when summarising reviews compared to political tweets. REFER-enhanced prompting im-



(a) Review summarisation - SOF



(b) Review summarisation - SPD

Figure 1: Comparison of SOF and SPD metrics for review summarisation. Green bars represent the oracle prompt by including actual frequency information. Blue bars are the base frameworks and orange bars are the REFER counterparts associated with them. When a REFER framework is statistically significantly better (lower value) than its base framework, the pair is highlighted using a green star on top. If a base framework is better, then it is highlighted using a red star.

proves fairness, with Prefix-instruct-R, Prefix-role-R and CoT-R being particularly effective, exhibiting balanced performance across both first-order and second-order fairness metrics. These findings suggest REFER's benefit across model reasoning processes when generating summaries, highlighting potential to improve LLM-generated content across varied architectures and datasets. Due to performance issues identified in the analysis above, we exclude Llama 3-8B from further analysis.

5.2 Second-Order Fairness Patterns

In this section, we examine second-order fairness in opinion summarisation in greater detail by examining different input proportions rather than using aggregated results, and conducting statistical analyses. While first-order metrics such as BUR and UER measure whether all groups are represented, second-order metrics such as SPD and SOF examine whether the proportional relationships between groups are preserved. This distinction is crucial because a model that consistently underrepresents minority opinions by the same margin across different contexts exhibits systematic bias, which is more problematic than random variations in representation.

Using datasets and input proportions from Section 4.1, we also include an oracle result with actual frequency prompts: '{number 1} and {number 2} out of {number of input} are {social value 1, social value 2}, generate a balanced summary reflecting this distribution. {direct prompt}'. The oracle prompt, which provides exact frequency information, establishes a theoretical upper bound for fairness performance. We report results using base frameworks and their REFER counterparts, with Mann-Whitney U tests for statistical significance. In visualisations, significantly better REFER results are highlighted with green stars. Review dataset results appear in Figure 1, showing patterns similar to the political tweets dataset, with full results in Appendix A.4.

In our analysis, the prompt incorporating actual frequencies proved most fair according to both metrics—an unsurprising result that serves as an upper bound for fairness achievement. REFER demonstrated statistically significant effectiveness when combined with Chain-of-Thought reasoning, prefix-based instructions, and persona-driven prompting, yielding the best overall performance across most LLMs and proving particularly effective with skewed input distributions. Chain-of-Thought REFER explicitly forces models to work through frequency calculations before summarisation, creating deliberative pathways that mirror human statistical reasoning. Prefix-instruction and Prefix-role REFER leverage instruction-following capabilities while providing numerical anchors for fairness assessment. Conversely, Agent collaboration REFER showed mixed results, suggesting that REFER is most effective when executed through unified reasoning processes rather than the added complexity of multi-party collaboration. This finding aligns with the original theoretical development of frequency-based reasoning frameworks, which were grounded in individual cognitive processes rather than distributed reasoning paradigms (Tversky and Kahneman, 1983; Fiedler, 1988; Gigerenzer, 1994; Gigerenzer and Hoffrage, 1995, 1999).

Notably, transitioning to REFER produced larger shifts in fairness metrics than choosing between different base prompting frameworks. As visualised in Figure 1, the variance amongst base frameworks (blue bars) is considerably smaller than the improvements achieved through REFER implementation (orange bars). This suggests that frequency-based reasoning may be more fundamental to fairness than the specific prompting strategy employed.

6 Conclusion

This study introduces REFER (Frequency Framed Prompting), a cognitive science-inspired debiasing framework for opinion summarisation. Drawing from research showing humans make more accurate statistical judgements with concrete frequencies rather than abstract probabilities, we demonstrate that frequency-based prompting enhances fairness in LLM-generated summaries. Our evaluation reveals that REFER improves model fairness across multiple datasets, models, and metrics, particularly when combined with stronger reasoning instructions and in larger models with superior instruction-following capabilities. Notably, REFER improvements often exceed differences between base prompting frameworks, suggesting frequency-based reasoning plays a greater role than specific prompting strategies in addressing fairness. REFER offers a lightweight, end-user accessible solution that works with both open-source and proprietary models without requiring retraining or hyperparameter tuning, addressing critical gaps in existing computationally demanding debiasing approaches.

Limitations

This study specifically focuses on model fairness rather than addressing dataset bias. Given that the fairness characteristics of the models' training data are unknown, our analysis is confined to examining the fairness of the generated summarisation outputs. The prompts employed in this study were manually crafted without extensive optimisation procedures. While our primary focus was on evaluating the discrete steps, eliciting frequency-based responses and analysing components of the summarisation process rather than prompt optimisation, we acknowledge that determining the optimal prompt formulation remains an ongoing challenge. Given computational and resource constraints, we conducted selective testing of representative prompting frameworks rather than an exhaustive evaluation of all possible approaches. Nevertheless, our approach presents a generalisable framework for utilising LLMs to achieve efficient debiasing. Future research directions could incorporate automated prompt generation techniques to further enhance the effectiveness of our methodology.

This study draws parallels between human cognitive processes and LLM information processing based on observed behavioural similarities. However, we acknowledge that these comparisons are fundamentally limited by our incomplete understanding of both human cognition and LLM internal mechanisms. The frequency-based reasoning improvements we observe in LLMs may arise from entirely different computational processes than those underlying human statistical reasoning, despite producing similar outcomes. Our analogy to human cognitive science serves as a useful framework for developing prompting strategies rather than making claims about mechanistic equivalence between human and artificial intelligence.

Ethical Considerations

This study followed ethical principles and guidelines. The authors of this paper by no means suggest that language models are intentionally biased. This study builds on extensive literature demonstrating that LLMs exhibit various forms of bias due to patterns in their training data. While these biases are not intentionally designed into the systems, they nonetheless have real impacts on users and society. Our work aims to provide practical methods for mitigating these well-documented biases in the specific context of opinion summarisation. We highly encourage readers to investigate and evaluate the findings for themselves. Overall, the goal of our research is to promote awareness of bias in summarising social media text since it is critical to understand what is summarised and whether it represents actual public opinions. Our work contributes to understanding the biases of summarisation models when summarising social media text, which is crucial for ethical use.

Our approach relies on predefined labels in datasets to measure bias. These labels are assigned based on established policies. However, if the labelling policy itself is inaccurate, our procedure might measure bias incorrectly. Therefore, we recommend using our technique only with datasets that have undergone careful review and construction to ensure accurate labelling.

References

Alberto Acerbi and Joseph M Stubbersfield. 2023. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120.

Griffin Adams, Alexander R Fabbri, Faisal Ladhak, Eric Lehman, and Noémie Elhadad. 2023. From sparse to dense: Gpt-4 summarization with chain of density prompting. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2023, page 68.

Armen E Allahverdyan and Aram Galstyan. 2014. Opinion dynamics with confirmation bias. *PloS one*, 9(7):e99557.

Conrad Borchers, Dalia Sara Gala, Benjamin Gilburt, Eduard Oravkin, Wilfried Bounsi, Yuki M Asano, and Hannah Rose Kirk. 2022. Looking for a handsome carpenter! debiasing gpt-3 job advertisements. *arXiv preprint arXiv:2205.11374*.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Ruizhe Chen, Jianfei Yang, Huimin Xiong, Jianhong Bai, Tianxiang Hu, Jin Hao, Yang Feng, Joey Tianyi Zhou, Jian Wu, and Zuozhu Liu. 2024. Fast model debias with machine unlearning. *Advances in Neural Information Processing Systems*, 36.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Leda Cosmides and John Tooby. 1996. Are humans good intuitive statisticians after all? rethinking some conclusions from the literature on judgment under uncertainty. *cognition*, 58(1):1–73.

Abhisek Dash, Anurag Shandilya, Arindam Biswas, Kripabandhu Ghosh, Saptarshi Ghosh, and Abhijnan Chakraborty. 2019. Summarizing user-generated textual content: Motivation and methods for fairness in algorithmic summaries. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–28.

Esin Durmus, Karina Nguyen, Thomas I Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, et al. 2023. Towards measuring the representation of subjective global opinions in language models. *arXiv preprint arXiv:2306.16388*.

Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R Frank, Matthew Groh, Laura Herman, Neil Leach, et al. 2023. Art and the science of generative ai. *Science*, 380(6650):1110–1111.

Jonathan St BT Evans and Keith E Stanovich. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241.

Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11737–11762, Toronto, Canada. Association for Computational Linguistics.

Klaus Fiedler. 1988. The dependence of the conjunction fallacy on subtle linguistic factors. *Psychological research*, 50(2):123–129.

Shaz Furniturewala, Surgan Jandial, Abhinav Java, Pragyan Banerjee, Simra Shahid, Sumit Bhatia, and Kokil Jaidka. 2024. Thinking fair and slow: On the efficacy of structured prompts for debiasing language models. *arXiv preprint arXiv:2405.10431*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, pages 1–79.

Somayeh Ghanbarzadeh, Yan Huang, Hamid Palangi, Radames Cruz Moreno, and Hamed Khanpour. 2023. Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models. *arXiv preprint arXiv:2307.10522*.

Gerd Gigerenzer. 1994. Why the distinction between single-event probabilities and frequencies is important for psychology (and vice versa). In *Subjective probability*, pages 129–161. Wiley.

Gerd Gigerenzer and Ulrich Hoffrage. 1995. How to improve bayesian reasoning without instruction: frequency formats. *Psychological review*, 102(4):684.

Gerd Gigerenzer and Ulrich Hoffrage. 1999. Overcoming difficulties in bayesian reasoning: a reply to lewis and keren (1999) and mellers and mcgraw (1999).

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned llms. *arXiv preprint arXiv:2311.04892*.

Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in chatgpt. *Nature Computational Science*, 3(10):833–838.

William Hart, Dolores Albarracín, Alice H Eagly, Inge Brechan, Matthew J Lindberg, and Lisa Merrill. 2009. Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychological bulletin*, 135(4):555.

Ulrich Hoffrage, Samuel Lindsey, Ralph Hertwig, and Gerd Gigerenzer. 2000. Communicating statistical information.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Nannan Huang, Haytham Fayek, and Xiuzhen Zhang. 2024. Bias in opinion summarisation from pretraining to adaptation: A case study in political bias. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1041–1055, St. Julian's, Malta. Association for Computational Linguistics.

Nannan Huang, Lin Tian, Haytham Fayek, and Xiuzhen Zhang. 2023. Examining bias in opinion summarisation through the perspective of opinion diversity. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 149–161, Toronto, Canada. Association for Computational Linguistics.

Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2023. Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 CHI conference on human factors in computing systems*, pages 1–15.

Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, et al. 2024. The biggen bench: A principled benchmark for fine-grained evaluation of language models with language models. *arXiv preprint arXiv:2406.05761*.

Richard P Larrick. 2004. Debiasing. *Blackwell handbook of judgment and decision making*, pages 316–338.

Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan Tn. 2023. Building real-world meeting summarization systems using large language models: A practical perspective. *arXiv preprint arXiv:2310.19233*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2023. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. On the biology of a large language model. *Transformer Circuits Thread*.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2019. Does gender matter? towards fairness in dialogue systems. *arXiv preprint arXiv:1910.10486*.

Zheyuan Liu, Guangyao Dou, Zhaoxuan Tan, Yijun Tian, and Meng Jiang. 2024. Towards safer large language models through machine unlearning. *arXiv preprint arXiv:2402.10058*.

Do Xuan Long, Duong Ngoc Yen, Anh Tuan Luu, Kenji Kawaguchi, Min-Yen Kan, and Nancy F Chen. 2024. Multi-expert prompting improves reliability, safety, and usefulness of large language models. *arXiv preprint arXiv:2411.00492*.

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2023. Fairness-guided few-shot prompting for large language models. *Advances in Neural Information Processing Systems*, 36:43136–43155.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Pat Pataranutaporn, Valdemar Danry, Joanne Leong, Parinya Punpongsanon, Dan Novy, Pattie Maes, and Misha Sra. 2021. Ai-generated characters for supporting personalized learning and well-being. *Nature Machine Intelligence*, 3(12):1013–1022.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. Infobench: Evaluating instruction following ability in large language models. *arXiv preprint arXiv:2401.03601*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. *arXiv preprint arXiv:2004.04498*.

Peter Sedlmeier and Gerd Gigerenzer. 2001. Teaching bayesian reasoning in less than two hours. *Journal of experimental psychology: general*, 130(3):380.

Anurag Shandilya, Kripabandhu Ghosh, and Saptarshi Ghosh. 2018. Fairness of extractive text summarization. In *Companion Proceedings of the The Web Conference 2018*, pages 97–98.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2022. Prompting gpt-3 to be reliable. *arXiv preprint arXiv:2210.09150*.

Yuting Tang, Ratish Puduppully, Zhengyuan Liu, and Nancy Chen. 2023. In-context learning of large language models for controlled dialogue summarization: A holistic benchmark and empirical analysis. In *Proceedings of the 4th New Frontiers in Summarization Workshop*, pages 56–67.

Pittawat Taveekitworachai, Febri Abdullah, and Ruck Thawonmas. 2024. Null-shot prompting: rethinking prompting large language models with hallucination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13321–13361.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157):1124–1131.

Amos Tversky and Daniel Kahneman. 1983. Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, 90(4):293.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, et al. 2023b. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Peter C Wason. 1960. On the failure to eliminate hypotheses in a conceptual task. *Quarterly journal of experimental psychology*, 12(3):129–140.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Maxwell Weinzierl and Sanda Harabagiu. 2024. Tree-of-counterfactual prompting for zero-shot stance detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–880.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2023. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. *arXiv preprint arXiv:2305.13300*.

Rongwu Xu, Zi'an Zhou, Tianwei Zhang, Zehan Qi, Su Yao, Ke Xu, Wei Xu, and Han Qiu. 2024. Walking in others' shoes: How perspective-taking guides large language models in reducing toxicity and bias. *arXiv preprint arXiv:2407.15366*.

Nakyeong Yang, Taegwan Kang, Stanley Jungkyu Choi, Honglak Lee, and Kyomin Jung. 2024. Mitigating biases for instruction-following language models via bias neurons elimination. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9061–9073.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Abdelrahman Zayed, Gonçalo Mordido, Samira Shabanian, Ioana Baldini, and Sarath Chandar. 2024. Fairness-aware structured pruning in transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22484–22492.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Ö Arik. 2024. Chain of agents: Large language models collaborating on long-context tasks. *arXiv preprint arXiv:2406.02818*.

Yusen Zhang, Nan Zhang, Yixin Liu, Alexander Fabbri, Junru Liu, Ryo Kamoi, Xiaoxin Lu, Caiming Xiong, Jieyu Zhao, Dragomir Radev, et al. 2023. Fair abstractive summarization of diverse perspectives. *arXiv preprint arXiv:2311.07884*.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571*.

A Appendix

A.1 Agent REFER Prompt

Summarisation Agent-REFER

`You are a summarisation specialist with exceptional ability to distil complex information into clear, concise, and accurate key points whilst preserving essential context and meaning. Let's first determine how many reviews out of number of input in the source documents are social value 1, social value 2. Then, let's generate a balanced summary that accurately reflects this distribution of opinions. source input`

`First, provide the counts in this format: social value 1 #number, social value 2 #number`

`Please write a short text containing the salient information that reflects this distribution. The summary of the reviews is:`

Frequency Agent

`You are a classification expert with superb ability to analyse, categorise, and sort any type of information with exceptional precision and accuracy. Classify how many reviews out of {number of input in the source documents} are {social value 1, social value 2}. Reviews to analyse: {source input}`

`Output exactly in format: {social value 1 #number, social value 2 #number}`

Judge Agent

`You are a validation expert specialising in quality control of summaries. You compare summaries against source opinion frequency distributions, precisely identify discrepancies, and provide targeted recommendations for alignment and improvement. {source input}`

`Opinion frequency distribution: {frequency classification by the Frequency Agent}`

`Summary: {summary generated by the Summarisation Agent}`

`Compare the summary against the opinion frequency distribution, checking for bias and misrepresentation.`

Senior Editor Agent

`You are a senior editor who expertly revises summaries based on detailed validation feedback, ensuring perfect alignment with source opinion frequency distributions while maintaining clarity and conciseness. You operate as the final quality control gate, transforming flagged summaries into highly accurate versions. {source input}`

`Opinion frequency distribution: {frequency classification by the Frequency Agent}`

`Summary: {summary generated by the Summarisation Agent}`

`Validation feedback: {validation and feedback generated by the Judge Agent}`

`Revise the summary to align with the opinion frequency distribution while maintaining clarity and balance.`

A.2 Implementation Details

We adopt zero-shot prompting for our experiments, as previous studies demonstrated that decoder-only instruction-tuned models perform effectively as zero-shot abstractive summarisers (Tang et al., 2023; Laskar et al., 2023; Adams et al., 2023). For open-source models, we utilised the model implementations and weights available from Hugging Face (Wolf et al., 2020), while proprietary models were accessed through their respective APIs. The experiments with open-source models were conducted using four NVIDIA A100 (40GB) GPUs. The hyperparameters for models employ strictly controlled settings during inference, including constrained output length with maximum new tokens of 256, low temperature of 0.001 for more reproducible output, and modest repetition mitigation via repetition penalty of 1.1, facilitating consistent and reproducible summarisation of review distributions.

A.3 Summary Qualitative Analysis

As visualised in Figure 2, the Llama 3-8B model exhibited the highest median word count among all models. Its interquartile range spanned approximately 95 to 205 words, indicating substantial variability in summary length within the central distribution. The extensive spread observed in the Llama3-8B distribution suggests a lack of consistency in generating summaries of uniform length, distinguishing it from other models which demonstrated comparatively more constrained distributions.

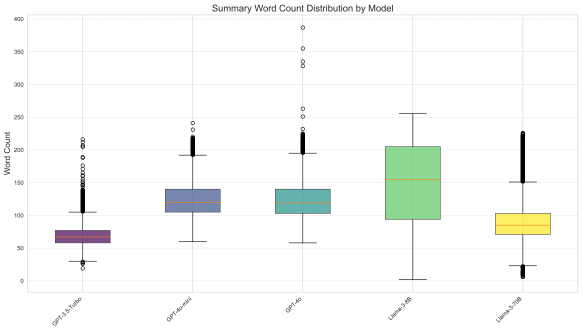


Figure 2: Summary word count distributions across different language models.

We also conducted a qualitative analysis across outputs generated using Llama3 and its 8B and 70B variants for instruction following capability. We present an example using the CoT-REFER prompt. Below are the original outputs from the two language models when instructed to first calculate frequency information before summarising product reviews:

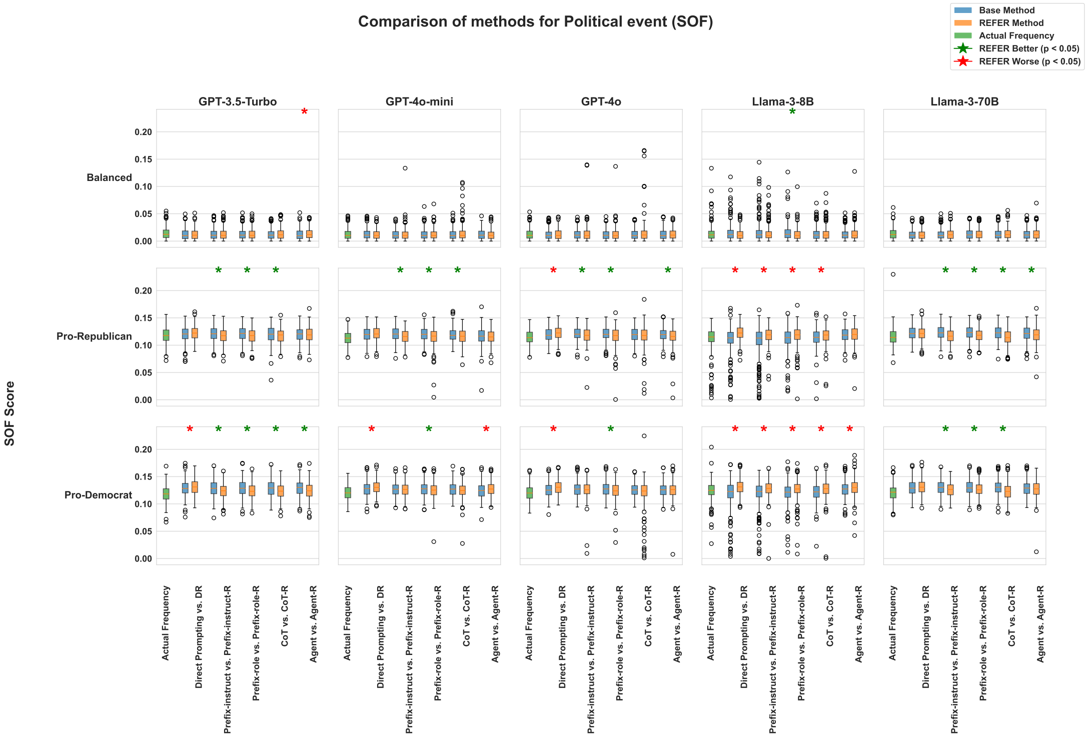
> **Llama3-8B Output**
>
> "Overall, this product has mixed reviews. While some users found it effective, easy to install, and appreciated its affordable price and value, others reported problems with water flow and taste, and some returned it due to poor performance or installation difficulty. This indicates the product may work well for some but not for everyone, highlighting the importance of reading multiple reviews before deciding to purchase."

> **Llama3-70B Output**
>
> "Out of 8 reviews, 4 are positive and 4 are negative. Positive reviewers praise the product for being affordable, improving water taste and flow rate, and providing good value for the price. However, some reviewers experienced issues such as slow water flow, poor taste, and compatibility problems, leading them to return or replace the product. Overall, the product has mixed results, and potential buyers should be aware of the varying experiences reported by customers."

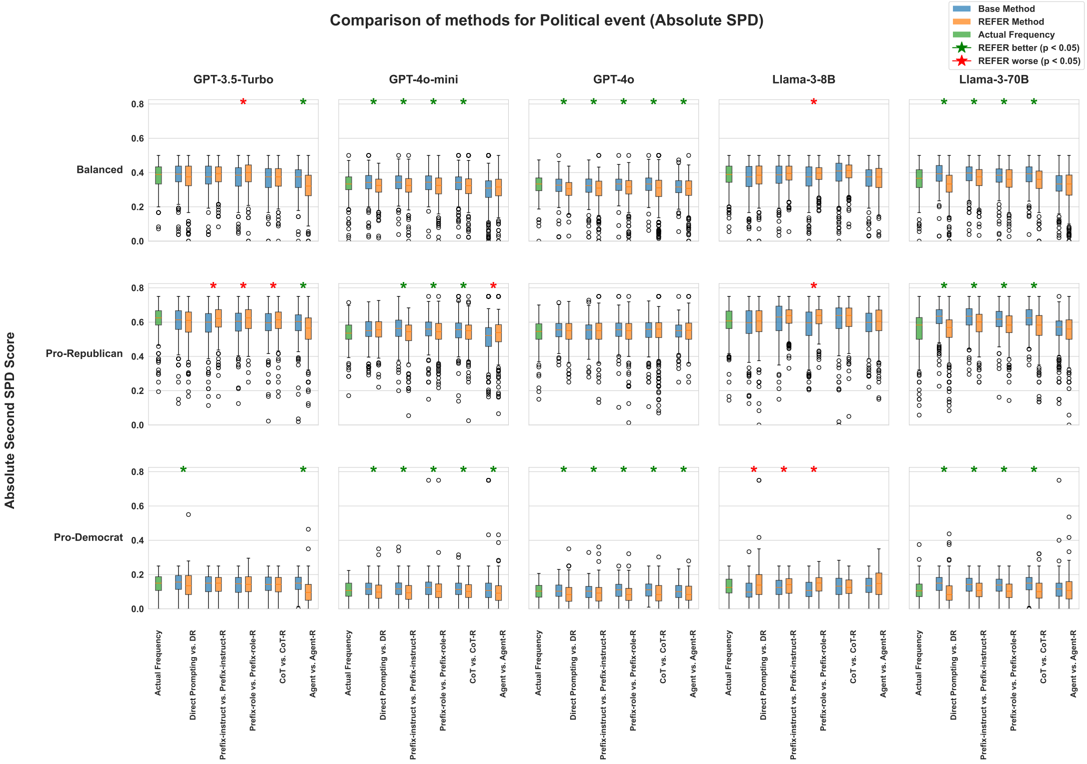
The analysis of these model outputs provides compelling evidence of how parameter scaling affects instruction-following capabilities aligns with previous studies (Qin et al., 2024; Ouyang et al., 2022). The original instruction explicitly requested both models to first calculate frequency information before providing a summary of product reviews. As clearly demonstrated in the outputs and comparative table 2, the 70B model successfully executed this sequential instruction, while the 8B parameter model did not. The 70B model begins with precise quantification—"Out of 8 reviews, 4 are positive and 4 are negative"—establishing the numerical distribution before proceeding to qualitative analysis. This suggests proper adherence to the instruction sequence. In contrast, the 8B model ignore the frequency calculation entirely, immediately beginning with a qualitative assessment: "Overall, this product has mixed reviews." Both models capture similar content points regarding the product's performance, mentioning affordability, taste improvement, installation experience, and flow issues. However, the fundamental difference lies in their instruction-following capabilities rather than content comprehension. The larger model not only follows instructions more precisely but also presents information in a more structured manner with clear delineation between frequency analysis and subjective assessment.

A.4 Second Order Fairness Full Results

The raw results for both the political tweets summarisation and review summarisations are reported in Tables 4, 3, 6, and 5. REFER counterparts that perform better than the base framework are highlighted in green, while base framework results that are better are highlighted in red. The effects of REFER are more profound when input documents are skewed.



(a) Political tweet summarisation - SOF



(b) Political tweet summarisation - SPD

Figure 3: Comparison of SOF and SPD metrics for political tweet summarisation. Green bars represent the oracle prompt by including actual frequency information. Blue bars are the base frameworks and orange bars are the REFER counterparts associated with them. When a REFER framework is statistically significantly better (lower value) than its base framework, the pair is highlighted using a green star on top. If a base framework is better, then it is highlighted using a red star.

Aspect	**Llama3-70B**	**Llama3-8B**
Instruction Following	Explicitly follows the instruction to calculate frequency first: “Out of 8 reviews, 4 are positive and 4 are negative.”	Fails to follow the instruction sequence, omitting frequency analysis entirely.
Frequency Analysis	Quantifies the distribution precisely (4 positive, 4 negative)	No quantification provided
Summary Introduction	Begins with quantified data before transitioning to summary	Immediately begins with qualitative assessment: “Overall, this product has mixed reviews.”
Example of Positive Aspects	“Positive reviewers praise the product for being affordable, improving water taste and flow rate, and providing good value for the price.”	“While some users found it effective, easy to install, and appreciated its affordable price and value...”
Example of Negative Aspects	“However, some reviewers experienced issues such as slow water flow, poor taste, and compatibility problems, leading them to return or replace the product.”	“...others reported problems with water flow and taste, and some returned it due to poor performance or installation difficulty.”

Table 2: Comparison of Instruction Following Between Llama3 8B and 70B Parameter Models

Model	GPT-3.5-Turbo	GPT-4o-mini	GPT-4o	Llama-3-8B	Llama-3-70B
Balanced					
Direct Prompting	0.04	0.03	0.03	0.04	0.04
REFER	0.03 ↑(0.6%)	0.03 ↓(-0.4%)	0.03 ↓(-1.4%)	0.03 ↑(2.0%)	0.03 ↑(1.6%)
Prefix-instruct	0.03	0.03	0.03	0.03	0.04
Prefix-instruct-R	0.04 ↓(-3.7%)	0.03 ↑(0.9%)	0.03 ↑(3.0%)	0.03 ↑(1.3%)	0.04 ↓(-0.4%)
Prefix-role	0.03	0.03	0.03	0.04	0.04
Prefix-role-R	0.04 ↓(-2.0%)	0.03 ↓(-0.6%)	0.03 ↑(0.9%)	0.03 ↑(6.8%)	0.03 ↑(2.7%)
CoT	0.03	0.03	0.03	0.04	0.03
CoT-R	0.04 ↓(-4.4%)	0.03 ↓(-0.6%)	0.03 ↓(-0.7%)	0.03 ↑(9.4%)	0.03 ↓(-1.0%)
Agent	0.03	0.03	0.03	0.04	0.03
Agent-R	0.03 ↓(-1.7%)	0.03 ↓(-1.4%)	0.03 ↓(-0.1%)	0.04 ↑(1.1%)	0.03 ↑(0.3%)
Pro-Positive					
Direct Prompting	0.12	0.12	0.12	0.11	0.12
REFER	0.12 ↓(-2.2%)	0.12 ↓(-1.3%)	0.12 ↓(-1.4%)	0.12 ↓(-8.8%)	0.12 ↓(-3.0%)
Prefix-instruct	0.12	0.12	0.12	0.12	0.12
Prefix-instruct-R	0.11 ↑(7.8%)	0.11 ↑(5.4%)	0.11 ↑(6.6%)	0.12 ↑(2.3%)	0.11 ↑(7.6%)
Prefix-role	0.12	0.12	0.12	0.12	0.12
Prefix-role-R	0.11 ↑(7.1%)	0.11 ↑(6.2%)	0.11 ↑(6.5%)	0.12 ↓(-0.8%)	0.11 ↑(6.9%)
CoT	0.12	0.12	0.12	0.11	0.12
CoT-R	0.11 ↑(7.5%)	0.11 ↑(5.8%)	0.11 ↑(8.8%)	0.12 ↓(-5.4%)	0.11 ↑(9.1%)
Agent	0.12	0.12	0.12	0.13	0.12
Agent-R	0.11 ↑(6.9%)	0.12 ↑(3.7%)	0.11 ↑(5.1%)	0.12 ↑(4.9%)	0.12 ↑(2.6%)
Pro-Negative					
Direct Prompting	0.12	0.12	0.12	0.11	0.12
REFER	0.12 ↓(-1.9%)	0.12 ↓(-1.7%)	0.12 ↓(-1.4%)	0.12 ↓(-10.0%)	0.12 ↓(-2.1%)
Prefix-instruct	0.12	0.12	0.12	0.12	0.12
Prefix-instruct-R	0.11 ↑(7.6%)	0.11 ↑(4.4%)	0.11 ↑(6.2%)	0.12 ↑(0.7%)	0.11 ↑(7.4%)
Prefix-role	0.12	0.12	0.12	0.11	0.12
Prefix-role-R	0.11 ↑(7.3%)	0.11 ↑(4.7%)	0.11 ↑(5.8%)	0.12 ↓(-5.6%)	0.11 ↑(6.0%)
CoT	0.12	0.12	0.12	0.11	0.12
CoT-R	0.11 ↑(7.0%)	0.11 ↑(4.9%)	0.11 ↑(8.9%)	0.12 ↓(-8.9%)	0.11 ↑(7.4%)
Agent	0.12	0.12	0.12	0.13	0.12
Agent-R	0.11 ↑(7.1%)	0.11 ↑(4.2%)	0.11 ↑(4.4%)	0.12 ↑(3.6%)	0.12 ↑(2.1%)

Table 3: SOF—Reviews results for all input conditions with improvements highlighted. REFER counterparts better than the base framework are indicated with ↑, worse results are indicated with ↓.

Model	GPT-3.5-Turbo	GPT-4o-mini	GPT-4o	Llama-3-8B	Llama-3-70B
Balanced					
Direct Prompting	0.01	0.01	0.01	0.01	0.01
REFER	0.01 ↓(-1.9%)	0.01 ↑(4.7%)	0.01 ↓(-2.4%)	0.01 ↑(6.4%)	0.01 ↓(-0.8%)
Prefix-instruct	0.01	0.01	0.01	0.01	0.01
Prefix-instruct-R	0.01 ↓(-3.6%)	0.01 ↓(-6.2%)	0.01 ↓(-10.4%)	0.01 ↑(3.0%)	0.01 ↓(-5.7%)
Prefix-role	0.01	0.01	0.01	0.01	0.01
Prefix-role-R	0.01 ↓(-0.2%)	0.01 ↓(-0.4%)	0.01 ↓(-3.3%)	0.01 ↑(13.6%)	0.01 ↓(-0.2%)
CoT	0.01	0.01	0.01	0.01	0.01
CoT-R	0.01 ↓(-8.6%)	0.01 ↓(-16.2%)	0.02 ↓(-27.6%)	0.01 ↓(-1.8%)	0.01 ↓(-4.1%)
Agent	0.01	0.01	0.01	0.01	0.01
Agent-R	0.01 ↓(-11.2%)	0.01 ↑(7.3%)	0.01 ↓(-5.5%)	0.01 ↑(3.8%)	0.01 ↓(-5.4%)
Pro-Republican					
Direct Prompting	0.12	0.12	0.12	0.12	0.12
REFER	0.12 ↓(-2.2%)	0.12 ↓(-1.4%)	0.12 ↓(-2.2%)	0.12 ↓(-2.9%)	0.12 ≡(0.0%)
Prefix-instruct	0.12	0.12	0.12	0.12	0.12
Prefix-instruct-R	0.12 ↑(3.3%)	0.12 ↑(3.8%)	0.12 ↑(2.1%)	0.12 ↑(1.6%)	0.12 ↑(4.9%)
Prefix-role	0.12	0.12	0.12	0.12	0.12
Prefix-role-R	0.12 ↑(3.7%)	0.12 ↑(3.8%)	0.12 ↑(3.0%)	0.12 ≡(0.0%)	0.12 ↑(3.7%)
CoT	0.12	0.12	0.12	0.12	0.12
CoT-R	0.12 ↑(1.8%)	0.12 ↑(2.9%)	0.12 ↑(1.8%)	0.12 ↑(3.9%)	0.12 ↑(6.6%)
Agent	0.12	0.12	0.12	0.12	0.12
Agent-R	0.12 ↑(0.3%)	0.12 ↑(1.1%)	0.12 ↑(2.8%)	0.12 ↑(0.4%)	0.12 ↑(1.8%)
Pro-Democrat					
Direct Prompting	0.13	0.13	0.13	0.13	0.13
REFER	0.13 ↓(-2.3%)	0.13 ↓(-2.7%)	0.13 ↓(-3.7%)	0.13 ↓(-3.3%)	0.13 ↓(-0.9%)
Prefix-instruct	0.13	0.13	0.13	0.13	0.13
Prefix-instruct-R	0.12 ↑(4.3%)	0.13 ↑(0.7%)	0.13 ↑(1.1%)	0.13 ↑(0.7%)	0.13 ↑(3.0%)
Prefix-role	0.13	0.13	0.13	0.13	0.13
Prefix-role-R	0.12 ↑(3.8%)	0.12 ↑(2.1%)	0.12 ↑(2.2%)	0.13 ↑(0.4%)	0.13 ↑(2.4%)
CoT	0.13	0.13	0.13	0.13	0.13
CoT-R	0.12 ↑(4.0%)	0.12 ↑(1.5%)	0.12 ↑(4.0%)	0.12 ↑(2.8%)	0.12 ↑(6.1%)
Agent	0.13	0.12	0.13	0.13	0.13
Agent-R	0.12 ↑(3.6%)	0.13 ↓(-2.7%)	0.13 ↓(-0.2%)	0.13 ↑(0.8%)	0.13 ↑(1.2%)

Table 4: SOF—Political tweets results for all input conditions with improvements highlighted. REFER counterparts better than the base framework are indicated with ↑, worse results are indicated with ↓, and no change is indicated with ≡.

Model	GPT-3.5-Turbo	GPT-4o-mini	GPT-4o	Llama-3-8B	Llama-3-70B
Balanced					
Direct Prompting	0.36	0.29	0.32	0.29	0.29
REFER	0.34 ↑(6.5%)	0.30 ↓(-4.5%)	0.29 ↑(8.8%)	0.35 ↓(-18.0%)	0.31 ↓(-6.0%)
Prefix-instruct	0.35	0.30	0.32	0.31	0.29
Prefix-instruct-R	0.32 ↑(7.8%)	0.28 ↑(6.9%)	0.28 ↑(10.7%)	0.32 ↓(-4.8%)	0.29 ↑(2.0%)
Prefix-role	0.35	0.30	0.31	0.29	0.30
Prefix-role-R	0.31 ↑(10.7%)	0.28 ↑(6.0%)	0.29 ↑(6.6%)	0.32 ↓(-8.7%)	0.30 ↓(-0.2%)
CoT	0.35	0.29	0.31	0.29	0.29
CoT-R	0.30 ↑(14.3%)	0.27 ↑(7.8%)	0.29 ↑(4.9%)	0.32 ↓(-13.7%)	0.30 ↓(-2.6%)
Agent	0.35	0.30	0.32	0.39	0.32
Agent-R	0.29 ↑(15.0%)	0.29 ↑(3.3%)	0.27 ↑(16.0%)	0.35 ↑(11.1%)	0.30 ↑(3.7%)
Pro-Positive					
Direct Prompting	0.15	0.13	0.15	0.13	0.13
REFER	0.14 ↑(8.7%)	0.11 ↑(16.7%)	0.13 ↑(13.2%)	0.18 ↓(-31.2%)	0.13 ↑(4.1%)
Prefix-instruct	0.15	0.13	0.13	0.13	0.13
Prefix-instruct-R	0.14 ↑(7.9%)	0.11 ↑(14.6%)	0.12 ↑(10.3%)	0.13 ↑(2.7%)	0.11 ↑(13.8%)
Prefix-role	0.16	0.13	0.13	0.13	0.12
Prefix-role-R	0.13 ↑(19.2%)	0.11 ↑(16.7%)	0.11 ↑(15.6%)	0.12 ↑(7.9%)	0.12 ↑(5.7%)
CoT	0.16	0.13	0.13	0.11	0.13
CoT-R	0.13 ↑(20.1%)	0.10 ↑(22.3%)	0.11 ↑(14.3%)	0.13 ↓(-16.8%)	0.12 ↑(8.1%)
Agent	0.16	0.13	0.14	0.19	0.14
Agent-R	0.13 ↑(19.3%)	0.11 ↑(12.8%)	0.11 ↑(19.7%)	0.15 ↑(18.3%)	0.13 ↑(7.4%)
Pro-Negative					
Direct Prompting	0.55	0.45	0.47	0.44	0.42
REFER	0.54 ↑(3.3%)	0.46 ↓(-1.8%)	0.45 ↑(4.4%)	0.46 ↓(-5.1%)	0.39 ↑(6.3%)
Prefix-instruct	0.55	0.46	0.48	0.45	0.44
Prefix-instruct-R	0.52 ↑(5.7%)	0.45 ↑(3.4%)	0.44 ↑(8.3%)	0.51 ↓(-13.6%)	0.45 ↓(-1.5%)
Prefix-role	0.55	0.45	0.48	0.42	0.46
Prefix-role-R	0.51 ↑(7.3%)	0.45 ↑(1.1%)	0.44 ↑(8.8%)	0.51 ↓(-21.6%)	0.49 ↓(-6.2%)
CoT	0.54	0.45	0.47	0.41	0.42
CoT-R	0.51 ↑(5.7%)	0.44 ↑(2.9%)	0.44 ↑(7.2%)	0.46 ↓(-12.0%)	0.46 ↓(-9.9%)
Agent	0.54	0.46	0.47	0.57	0.49
Agent-R	0.43 ↑(20.2%)	0.43 ↑(5.2%)	0.43 ↑(7.2%)	0.52 ↑(8.0%)	0.43 ↑(13.2%)

Table 5: SPD—Review results for all input conditions with improvements highlighted. REFER counterparts better than the base framework are indicated with ↑, worse results are indicated with ↓.

Model	GPT-3.5-Turbo	GPT-4o-mini	GPT-4o	Llama-3-8B	Llama-3-70B
Balanced					
Direct Prompting	0.38	0.34	0.33	0.36	0.39
REFER	0.37 ↑(3.3%)	0.32 ↑(5.0%)	0.30 ↑(7.3%)	0.38 ↓(-3.8%)	0.33 ↑(16.4%)
Prefix-instruct	0.38	0.34	0.32	0.38	0.39
Prefix-instruct-R	0.38 ↓(-2.4%)	0.32 ↑(5.7%)	0.30 ↑(6.7%)	0.39 ↓(-2.5%)	0.36 ↑(5.8%)
Prefix-role	0.37	0.34	0.33	0.37	0.38
Prefix-role-R	0.39 ↓(-5.5%)	0.32 ↑(5.9%)	0.31 ↑(7.1%)	0.39 ↓(-6.5%)	0.36 ↑(5.1%)
CoT	0.36	0.34	0.33	0.39	0.39
CoT-R	0.37 ↓(-0.7%)	0.32 ↑(5.4%)	0.30 ↑(9.6%)	0.40 ↓(-2.7%)	0.35 ↑(8.8%)
Agent	0.36	0.30	0.32	0.37	0.33
Agent-R	0.32 ↑(12.9%)	0.31 ↓(-3.7%)	0.30 ↑(5.5%)	0.37 ↑(1.0%)	0.32 ↑(3.6%)
Pro-Republican					
Direct Prompting	0.60	0.55	0.55	0.59	0.62
REFER	0.59 ↑(1.3%)	0.55 ↑(0.2%)	0.54 ↑(2.2%)	0.59 ↓(-1.1%)	0.54 ↑(12.9%)
Prefix-instruct	0.59	0.56	0.55	0.61	0.62
Prefix-instruct-R	0.61 ↓(-4.4%)	0.53 ↑(5.5%)	0.54 ↑(1.2%)	0.63 ↓(-3.0%)	0.59 ↑(5.5%)
Prefix-role	0.59	0.55	0.56	0.58	0.61
Prefix-role-R	0.61 ↓(-3.8%)	0.54 ↑(3.1%)	0.54 ↑(2.9%)	0.63 ↓(-8.4%)	0.58 ↑(6.1%)
CoT	0.59	0.55	0.55	0.61	0.62
CoT-R	0.61 ↓(-3.1%)	0.54 ↑(2.8%)	0.54 ↑(1.7%)	0.62 ↓(-1.4%)	0.57 ↑(8.8%)
Agent	0.59	0.51	0.55	0.59	0.55
Agent-R	0.55 ↑(6.6%)	0.53 ↓(-3.0%)	0.55 ↓(-0.1%)	0.60 ↓(-1.3%)	0.55 ↑(0.4%)
Pro-Democrat					
Direct Prompting	0.15	0.12	0.10	0.11	0.14
REFER	0.13 ↑(10.5%)	0.10 ↑(11.3%)	0.09 ↑(15.5%)	0.15 ↓(-32.4%)	0.10 ↑(31.8%)
Prefix-instruct	0.14	0.12	0.10	0.12	0.14
Prefix-instruct-R	0.14 ↓(-1.8%)	0.10 ↑(17.3%)	0.09 ↑(10.8%)	0.13 ↓(-6.6%)	0.11 ↑(19.5%)
Prefix-role	0.14	0.12	0.11	0.11	0.14
Prefix-role-R	0.14 ↓(-2.3%)	0.11 ↑(11.7%)	0.09 ↑(19.7%)	0.14 ↓(-27.3%)	0.10 ↑(24.1%)
CoT	0.14	0.12	0.11	0.13	0.14
CoT-R	0.14 ↑(1.8%)	0.10 ↑(11.7%)	0.09 ↑(13.6%)	0.13 ↑(4.6%)	0.11 ↑(25.0%)
Agent	0.14	0.12	0.10	0.14	0.12
Agent-R	0.10 ↑(30.9%)	0.10 ↑(18.9%)	0.09 ↑(8.2%)	0.15 ↓(-7.6%)	0.12 ↑(0.5%)

Table 6: SPD—Political tweets results for all input conditions with improvements highlighted. REFER counterparts better than the base framework are indicated with ↑, worse results are indicated with ↓.