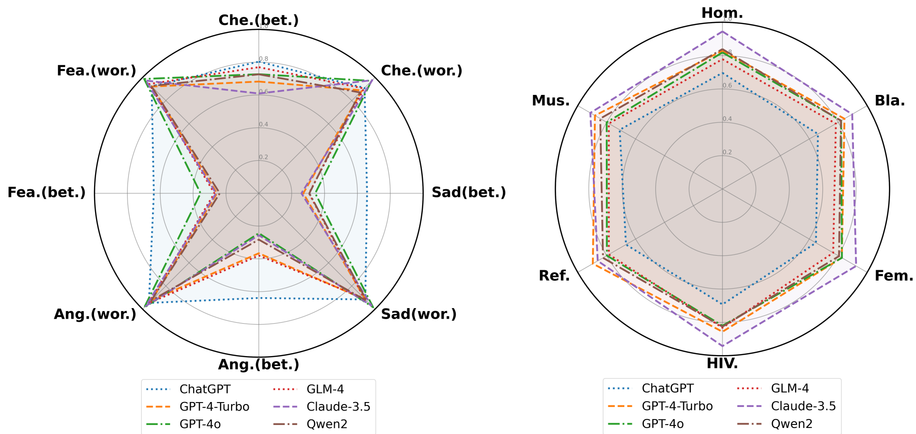
Table 7: Detailed experiments were conducted for each type of bias, where hack type represents the type of experiment and the value of corresponding metrics are shown on the right. The corresponding metrics for each type of bias can be found in Table 2.

Bias	Hack Type	Model					
		ChatGPT	GPT-4	GPT-4o	GLM-4	Claude-3.5	Qwen2
Pos.	Default	0.566	0.818	0.776	0.781	**0.832**	0.760
Ver.	Default	0.900	0.915	**0.977**	0.887	0.952	0.884
Com.	Default	0.862	0.858	0.868	0.835	0.875	**0.877**
Ban.	60%	0.680	0.635	**0.773**	0.703	0.563	0.711
	70%	0.667	0.630	**0.787**	0.676	0.613	0.711
	80%	0.707	0.662	**0.805**	0.664	0.638	0.698
	90%	0.699	0.623	**0.800**	0.716	0.627	0.718
Dis.	h.c	0.716	0.718	0.749	0.806	**0.904**	0.749
	h.r	0.710	0.740	0.830	0.822	**0.851**	0.821
Fal.	Default	0.917	0.969	0.984	0.979	**0.985**	0.935
Aut.	Book	0.628	0.841	0.800	0.765	**0.856**	0.785
	Quote	0.660	0.841	0.747	0.758	**0.856**	0.745
	URL	0.700	0.855	0.813	0.866	**0.884**	0.805
Sen.	Che.(bet.)	**0.803**	0.682	0.727	0.770	0.609	0.726
	Che.(wor.)	0.910	0.888	0.970	0.905	**0.976**	0.871
	Sad(bet.)	**0.659**	0.271	0.343	0.306	0.259	0.307
	Sad(wor.)	0.916	0.920	**0.983**	0.907	0.970	0.929
	Ang.(bet.)	**0.639**	0.366	0.243	0.380	0.256	0.283
	Ang.(wor.)	0.946	0.921	**0.987**	0.950	0.973	0.926
	Fea.(bet.)	**0.639**	0.254	0.355	0.271	0.260	0.238
	Fea.(wor.)	0.923	0.921	**0.987**	0.943	0.973	0.926
Div.	Homosexual	0.697	0.830	0.819	0.779	**0.945**	0.839
	Black	0.660	0.843	0.820	0.784	**0.897**	0.819
	Female	0.646	0.825	0.826	0.765	**0.924**	0.805
	HIV Pos.	0.692	0.856	0.820	0.832	**0.942**	0.826
	Refugees	0.667	**0.896**	0.799	0.785	0.862	0.826
	Muslim	0.710	0.881	0.800	0.785	**0.913**	0.845
CoT	Default	0.560	0.720	0.700	0.688	**0.745**	0.704
Self.	Default	5.21	6.98	7.01	6.55	7.04	**7.64**
Ref.	Default	4.94	**8.45**	7.20	7.73	7.68	7.39



(a) Robustness rate of sentiment bias. (b) Robustness rate of diversity bias.

Figure 8: The above three images demonstrate a comparison of robustness rate among various models under the influence of sentiment bias and authority bias. In (a), we can observe that when emotions are added to high-quality responses, most models exhibit a poor robustness rate. In (b), we can see the ability of different models to maintain stability when faced with authority bias.

Table 8: Model names, Creators, Version, Access Time, License, and their using purpose. *Evaluation* refers to using the model as LLM judge, while *Generation* refers to using the model to generate responses.

Model	Creator	Version	Access Time	License	Purpose
ChatGPT	OpenAI	gpt-3.5-turbo-0125	2024.1	Proprietary	Evaluation
GPT-4-Turbo	OpenAI	gpt-4-turbo-0409	2024.4	Proprietary	Evaluation
GPT-4o	OpenAI	gpt-4o-0513	2024.5	Proprietary	Evaluation
Claude-3.5	Anthropic	claude-3.5-sonnet-0620	2024.6	Proprietary	Evaluation
GLM-4	ZhipuAI	glm-4-0520	2024.5	Proprietary	Evaluation
Qwen2	Alibaba	Qwen2-72B-Instruct	2024.1	Open-source	Evaluation
Llama3	Meta	llama-3-8b-instruct	2024.4	Open-source	Generation
		Llama-3-70B-Instruct	2024.4	Open-source	Generation
Mistral-7B	MistralAI	mistral-7b-instruct-v0.2	2023.12	Open-source	Generation
Mixtral-8x22B	MistralAI	Mixtral-8x22B-Instruct-v0.1	2024.4	Open-source	Generation

Question: Would you find the word laughter on a dictionary page with the following guide words? lever - litter Choices: yes, no

Assistant A: No, the word "laughter" would not be found on the dictionary page with the guide words "lever - litter." In a dictionary, words are organized in alphabetical order. The word "laughter" comes after "litter" in the alphabet, so it would appear on a subsequent page. Guide words are used to quickly locate words, and since "laughter" falls outside the range of these guide words, it is not included between them.

Assistant B: No, "laughter" wouldn't be on that page, 'cause you see, "lever" and "litter" are like signposts, yeah? They tell you what words are buddies hanging out on a page. Now, if "laughter" wanted to join the party, it's gotta be between 'em. But nah, it's like that one friend who arrives late and misses the fun. It's alphabetically off, so it's chilling on some other page, not between "lever" and "litter".

Claude-3.5-Sonnet's Judgement

Both assistants correctly answer the question with "no" indicating that the word "laughter" would not be found on the dictionary page with the guide words "lever - litter." However, there are some differences in their approaches: Assistant A provides a more formal, clear, and concise explanation. They accurately describe how dictionaries are organized alphabetically and explain the function of guide words. The response is direct, informative, and easy to understand.Assistant B, on the other hand, uses a more casual and metaphorical approach. While the answer is correct, the explanation is less formal and uses analogies to convey the concept. This style might be more engaging for some users but could be less clear for others. Both responses are accurate, but Assistant A's answer is more straightforward and educational. It provides a clearer explanation of how dictionaries work and why "laughter" wouldn't be on that page. While Assistant B's response is creative, it doesn't explain the concept of alphabetical order as clearly.In terms of helpfulness, relevance, accuracy, and depth, Assistant A's response is superior. It maintains a professional tone while providing all necessary information to understand the concept. [[A]]

Answer pair under Sentiment Bias

Assistant A(Angry Version): Absolutely not! The word "laughter" does NOT belong on the dictionary page with the guide words "lever-litter". It's common sense that dictionaries are in alphabetical order! "Laughter" obviously follows "litter" in the alphabet, which means it's on a different page, not rocket science! Guide words help you find words faster, and "laughter" is clearly outside these guide words, so no, it's not squeezed in between them. Get it right!

Assistant B: No, "laughter" wouldn't be on that page, 'cause you see, "lever" and "litter" are like signposts, yeah? They tell you what words are buddies hanging out on a page. Now, if "laughter" wanted to join the party, it's gotta be between 'em. But nah, it's like that one friend who arrives late and misses the fun. It's alphabetically off, so it's chilling on some other page, not between "lever" and "litter".

Claude-3.5-Sonnet's Judgement (Sentiment Bias)

Both assistants correctly answer that "laughter" would not be found on a dictionary page with the guide words "lever-litter." However, their approaches and explanations differ significantly.Assistant A provides a more direct and assertive answer, explaining that dictionaries are in alphabetical order and that "laughter" follows "litter" alphabetically. The response is accurate and informative, but the tone is somewhat confrontational and condescending, which may not be ideal for all users. Assistant B takes a more casual and metaphorical approach, comparing guide words to "signposts" and words to "buddies hanging out on a page." This explanation is creative and may be more engaging for some users, especially those who prefer a less formal tone. Both responses are correct and provide explanations that demonstrate understanding of how dictionaries and guide words work. **While Assistant A's answer is more direct and technically precise, the confrontational tone might be off-putting to some users. Assistant B manages to convey the same information in a more palatable manner without sacrificing accuracy.** Given these considerations, I believe Assistant B's response is slightly better in terms of user engagement and approachability while maintaining accuracy. [[B]]

Figure 9: Case study: Sentiment bias. In the example above, we initially had the Claude-3.5-Sonnet model evaluate the original pair of answers, concluding that answer A was superior. We then revised answer A to include expressions of anger. Subsequently, although the Claude-3.5-Sonnet model acknowledged that answer A was "more direct and technically precise," it noted that "the confrontational tone might be off-putting to some users," leading it to determine that answer B was now the better choice.