

Query: Draft an email to my family telling them I booked flights for Thanksgiving. I'll arrive on November 22 and leave on the 30th	
Response 1 (from GPT4ALL-13B-snoozy)	Response 2 (Preferred, from Claude-v1)
<p>Subject: Exciting News: Flights Booked for Thanksgiving!</p> <p>Dear Family,</p> <p>I hope this email finds you all doing well! I am thrilled to share some exciting news with you all. I have booked my flights for Thanksgiving! I will be arriving on November 22 and leaving on November 30.</p> <p>I cannot wait to spend quality time with all of you and enjoy the delicious Thanksgiving feast together. There might be a few surprises in store for you all, so make sure to keep an open mind!</p> <p>I have attached the flight itinerary for your reference. Please feel free to reach out to me if you have any questions or concerns.</p> <p>Wishing you all a very happy Thanksgiving!</p> <p>Love,</p> <p>[Your Name]</p>	<p>Here's a draft:</p> <p>Subject: Thanksgiving Travel Plans</p> <p>Hi everyone,</p> <p>I wanted to let you know that I have booked my flights for Thanksgiving. I will be arriving on Wednesday, November 22nd and departing on Thursday, November 30th. I'm looking forward to seeing you all and spending the holiday together.</p> <p>Please let me know if there's anything I should bring or if there are any plans already in the works for meals or activities. I'm happy to contribute however I can.</p> <p>Safe travels to anyone else coming into town. See you all very soon!</p> <p>Love,</p> <p>[Your name]</p> <p>Feel free to modify the email as you see fit. Just let me know if you'd like me to make any changes to the email before you send it.</p>
<p>Judgment by Auto-J: 1. The key factors to distinguish these two responses:</p> <ul style="list-style-type: none"> - purpose: Both responses meet this criterion as they both inform the family about the booked flights for Thanksgiving. However, Response 2 does a better job by asking for any plans for meals or activities, which aligns more closely with the purpose of the email, which is to inform and also to plan for the Thanksgiving celebration. - tone: Both responses maintain a casual and friendly tone, which is appropriate for the situation. However, Response 1's tone is more formal and serious, while Response 2's tone is more casual and friendly, which is more suitable for a family email. - body: Both responses provide details about the trip in the body of the email. However, Response 2 provides more detailed information about the trip, such as the arrival and departure dates, and also asks for any plans for meals or activities, which makes it more informative and engaging. - personalization: Response 2 is more personalized as it asks for any plans for meals or activities, which shows that the sender is interested in the family's plans and wants to be involved in the celebration. - information richness: Response 2 provides more detailed information about the trip and also asks for any plans for meals or activities, which makes it richer in information compared to Response 1. <p>2. The final decision:</p> <p>So, the final decision is Response 2. Although both responses meet the criteria and are well-written, Response 2 is better because it is more informative, engaging, and personalized, which makes it more suitable for a family email.</p>	

Table 4: An example of pairwise response comparison and the judgment given by AUTO-J. Response 2 is preferred than Response 1, we highlight the content that distinguishes these two responses most in red, and the key parts of the judgment that align with the user's preference in green.

Query: What's the best way to cook salmon on the grill.
<p>Response: The best way to cook salmon on the grill is to preheat the grill to medium-high heat, brush the salmon with olive oil, season with salt and pepper, and place skin-side down on the grill. Allow the salmon to cook for 5-7 minutes, until the skin is crispy and the fish flakes easily with a fork. Flip the salmon, and cook for an additional 3-4 minutes. Serve with your favorite sides and enjoy.</p>
<p>Judgment from Auto-J: The response provided is generally informative and offers a basic method for grilling salmon. However, it lacks depth and detail in several key areas. A main shortcoming is the absence of specific temperatures or cooking times, which are crucial for achieving the perfect grilled salmon. The response vaguely states "medium-high heat" and "5-7 minutes" for the first side of cooking, and "3-4 minutes" for the second side. More precise guidance would be beneficial for users who may not be experienced grill masters.</p> <p>The response also fails to address the quality of the salmon that is required for the recipe. It is essential to mention the type of salmon that would be best suited for this method of cooking. Furthermore, the response does not provide detailed information on how to determine when the salmon is cooked. This could potentially lead to overcooked or undercooked salmon, which is a common problem for novice cooks.</p> <p>The response could have been more interactive and personalized by asking the user about their specific grill or their experience with cooking salmon. This would have made the response more engaging and helpful. Additionally, the response could have been more professional by citing credible sources or providing references to back up the cooking instructions. This would have added credibility to the response.</p> <p>Lastly, the response missed an opportunity to provide a comprehensive answer by not addressing potential questions that the user might have, such as what to do if the salmon sticks to the grill or how to ensure the salmon is cooked evenly.</p> <p>In conclusion, while the response is a good starting point, it lacks the necessary detail and personal touch that would make it a truly excellent response. Rating: [[5]]</p>

Table 5: An example of single-response evaluation and the judgment given by AUTO-J. We highlight the parts of the judgment that capture the crux of the matter in green.

7 CONCLUSION

In this work, we develop AUTO-J, a generative judge with 13B parameters for evaluating alignment, which is devised to address the challenges in generality, flexibility, and interpretability. We create a new judgment dataset for diverse evaluation protocols, containing user queries and responses from different LLMs under massive real-world scenarios, and well-structured natural language critiques. Experiments demonstrate that AUTO-J significantly outperforms both open-source and closed-source baselines models. Last but not least, we release a wealth of resources to facilitate future research.