[Amazon Confidential] Criteria-based Judge Data Annotation Specs

We plan to get data annotations for Nova judge training and evaluation.

- We will provide you a set of 3000 prompts by end of this week.
    - Open Question: What is the data format in which we provide the prompts to AGI-DS?
- AGI-DS/Surge can get two responses for the prompts that will be used for annotations.
    - Model-1: Deepseek-R1-0528 (one of current top model in LMArena)
    - Model-2: Folsom (one of best Nova ckpts in LMSys)
- The annotators will be shown (prompt, response_A, response_B) as input. Kindly randomize the order of responses when providing the data to annotators.
- Output from annotators given (prompt, response_A, response_B):
    - We need the output in the format shown here: Output JSON format
    - A set of 3-10 criteria against which to evaluate the responses. Each criteria preferably 5-20 words.
        - e.g. if it is a coding question, one of the criteria can be "is the code executable?". For questions like "write two paragraphs explaining photosynthesis in less than 50 words", couple of criteria can be "are there two paragraphs in the response?" or "does the response has less than 50 words?"
    - If criteria is boolean, annotate with True/False if each of the responses meet the criterion. Give a brief reasoning (10 - 60 words)
    - If criteria is not boolean, give a score on a scale of 1-5 to what extent each of the responses meet the criterion, 1 being worst and 5 being best. Give a brief reasoning (10 - 60 words)
    - Assign a weight to each criteria, on how important it is to determine the final judgement which is better. Weight should be a number between 0 and 1. Adding the weights over all criteria should sum to 1.
    - Final overall judgement which response is better and a brief reasoning 10-60 words. Adhere to the format below for the final overall judgement.
        - A>B if response A is better
        - B>A if response B is better
        - A=B if the responses are a tie and they satisfy the conditions provided in the prompt
        - A=B (both bad) if both responses are bad e.g. if both the model responses are incorrect

An example is shown below with Input to annotators and Output from annotators

Appendix

Output JSON format

```
{ "criterion_1":
  { "question": "...", // the criterion on which you are
evaluating the responses
    "resp_a": "...", // one of: "true" or "false" or 1-5 score
indicating whether assistant A's response meets the criterion or
not or to what extent
    "resp_b": "...", // one of: "true" or "false" or 1-5 score
indicating whether assistant B's response meets the criterion or
not or to what extent respectively
    "reason": "...", // reason why the two assistants' responses
meet or do not meet the criterion or the assigned score
    "weight": e.g. 0.3, // importance of criteria, should sum to
1 when added over all criteria
  },
  ... // repeat for each 3-10 criteria
  "overall":
  { "judgement": "...", // verdict for overall (one of: A>B,
B>A)
    "reason": "..." // reason for the overall preferred
judgement based on criteria above
  }
}
```

Example Input to annotators

Input doesn't have to be in exact same format as below. Any format is fine as long as there is a prompt response_A, and response_B clearly distinguished with the response order randomized for the two models.

<|User Prompt|>
Suggest a possible solution to the given problem. The company is not able to attract customers due to outdated website design.

<|Start of Assistant A's Response|>
One possible solution could be to invest in a website redesign to make it more modern, user-friendly, and visually appealing. This could involve updating the design, layout, and navigation of the website, as well as optimizing it for mobile devices. Additionally, incorporating features such as online booking, chat support, and personalized recommendations could help attract and retain customers. It may also be beneficial to regularly update content and images to keep the website fresh and engaging.

<|End of Assistant A's Response|>

<|Start of Assistant B's Response|>

To address the issue of retaining customers who are leaving due to non-responsive customer service, the company needs to take a multi-pronged approach that focuses on both immediate improvements and long-term changes to the customer service culture. Here's a step-by-step strategy to tackle this problem:

1. Assess the Current Situation:
- Evaluate the current customer service processes to identify bottlenecks or failures in responsiveness.
- Collect and analyze customer feedback, complaints, and service metrics (e.g., response time, resolution time, customer satisfaction scores).

2. Set Clear Objectives:
- Define what "responsive customer service" means for your company (e.g., average response times, resolution rates).
- Set specific, measurable goals to improve customer service responsiveness (e.g., X% improvement in response time within Y months).

3. Invest in Customer Service Training:
- Implement regular training programs to enhance the skills of customer service representatives, focusing on communication, problem-solving, and empathy.
- Develop a knowledge base and resources that representatives can refer to, ensuring accurate and timely information delivery.

4. Improve Infrastructure and Technology:
- Upgrade customer service software (e.g., ticketing systems, live chat, CRM platforms) to streamline operations and provide quicker responses.
- Accessibly integrate customer interaction channels (phone, email, chat, social media) into a unified platform to provide a consistent and efficient service experience.

5. Implement an Omnichannel Approach:
- Ensure customers can reach out through various channels and receive an equally responsive service across all platforms.
- Use automated responses and chatbots to provide immediate acknowledgment and basic support while queuing issues for human service agents.

6. Empower Customer Service Teams:
- Give service representatives the authority to make decisions and solve problems to avoid unnecessary delays.
- Encourage proactive communication and follow-ups to ensure issues are resolved and customers are satisfied.

7. Monitor Performance and Feedback:
- Implement real-time monitoring and reporting to track responsiveness and customer satisfaction.
- Regularly review customer feedback to identify areas for improvement and adjust strategies accordingly.

8. Recognition and Reward Systems:
- Create incentives for outstanding customer service performance to motivate staff.
- Celebrate successes and share positive customer testimonials to boost morale.

9. Implement a Closed-Loop Feedback System:
- Ensure every customer's complaint or feedback is followed through to resolution.
- Communicate back to the customer about what was done to address their concerns, improving customer trust and loyalty.

10. Long-Term Cultural Change:
- Shift company culture towards customer-centricity, where responsiveness is a core value.
- Engage leadership at all levels to champion customer service excellence.

11. Communicate with Customers:
- Inform customers about the steps being taken to improve customer service responsiveness.
- Set realistic expectations concerning response times and follow-through.

12. Review and Adapt:
- Periodically revisit the strategies put in place to ensure they're working effectively.
- Be prepared to make adjustments as customer needs and the business environment evolve.

By taking these steps, the company can begin to rebuild its reputation for customer service, regain customer trust, and ultimately improve customer retention rates. It's essential that improvements are ongoing and reflect a genuine commitment to delivering the service that customers expect and deserve.
<|End of Assistant B's Response|>

Example Output from annotators

{
"criterion_1": {
"question": "Does the response directly address the outdated website design as the core problem?",
"resp_a": "true",
"resp_b": "false",
"reason": "Assistant A's response focuses on website redesign, explicitly mentioning updates to

design, layout, and navigation. Assistant B's response discusses customer service responsiveness for retention, which is unrelated to website design as specified in the prompt.",
“weight”: 0.3
},
"criterion_2": {
"question": "Does the response explain how the solution will help attract customers?",
"resp_a": "true",
"resp_b": "false",
"reason": "Assistant A links solutions like mobile optimization and added features (e.g., online booking) to attracting and retaining customers. Assistant B focuses on retaining existing customers through service improvements but does not address attracting new customers via website changes.",
“weight”: 0.3
},
"criterion_3": {
"question": "Does the response provide specific, actionable recommendations for improvement?",
"resp_a": "true",
"resp_b": "false",
"reason": "Assistant A offers concrete steps such as updating design elements, optimizing for mobile, and incorporating features like chat support. Assistant B provides detailed steps, but they are for customer service (e.g., training, technology upgrades), not for website design, making them irrelevant to the prompt.",
“weight”: 0.3
},
"criterion_4": {
"question": "Does the response stay on-topic without introducing unrelated issues?",
"resp_a": "true",
"resp_b": "false",
"reason": "Assistant A remains focused solely on website redesign solutions. Assistant B deviates by addressing customer service retention, which is not mentioned in the prompt and is a separate problem.",
“weight”: 0.1
},
"overall": {
"judgement": "[[A>B]]",
"reason": "Assistant A's response fully addresses the prompt by providing relevant, specific solutions for the outdated website design and clearly linking them to customer attraction. Assistant B's response is off-topic, discussing customer service improvements instead of website design, and fails to meet the core requirements of the prompt. Based on all criteria, A is superior as it directly solves the stated problem."
}
}