| **Identified by** | **Reviewer** |
| --- | --- |
| **Limitation Category** | Model Output Quality / Performance |
| **Limitation Description** | The model’s current performance across key dimensions such as faithfulness, completeness, and relevance is not aligned with expectations. Both user feedback and independent evaluation indicate material quality concerns that may limit the model’s effectiveness in production use. |
| **Assessment and Severity** | A joint analysis of model performance—based on **user feedback (n = 696)** and **independent evaluation by the dedicated testing team (n = 299)**—indicates that the model is currently underperforming across key dimensions of **faithfulness, completeness, and relevance**. • **Faithfulness**: ~65% of user responses were rated as fully or mostly accurate, with over 18% flagged as largely inaccurate or completely incorrect. Independent testing found 70% factual accuracy, with 11% rated inaccurate and 19% marked as N/A. • **Completeness**: Around 60% of user feedback marked outputs as complete or covering most key points, while 29% were seen as brief, incomplete, or containing major gaps. The testing team found 69% completeness, 6% incomplete, and 25% N/A. • **Relevance**: Relevance was comparatively stronger, with 68% of users and 85% of testers indicating positive alignment. However, up to 32% of responses were still rated as marginally relevant or irrelevant. Together, these findings highlight that while the model performs adequately in some scenarios, its consistency and reliability do not currently meet production standards. The observed gaps suggest risks to user trust and operational utility. This issue is assessed as a **Level 2** limitation. |
| **Severity Level** | Level 2 |
| **Mitigating Controls** | The model is currently used under a limited user scope with guardrails in place, including restricted use cases and feedback mechanisms. These help reduce exposure but are not sufficient to fully mitigate the observed quality risks in a broader deployment. |
| **Required Remediation Action Items** | 1) Conduct root cause analysis on low-performing outputs to identify common error patterns. 2) Explore prompt engineering or model tuning based on observed failure cases. 3) Expand test coverage to include additional scenarios and input types. 4) Define quantitative quality thresholds and integrate them into the ongoing monitoring plan. 5) Integrate structured user feedback (e.g., thumbs up/down, comments) into the model improvement cycle. |
| **Action Item Due Date** | [Insert Date, e.g., Oct/15/2025] |