**LLM-Powered Medical Inquiry Triage Agent – PoC Proposal**

**Prepared by:** Cory Arthus and William Nerbonne  
**Program:** Graduate Certificate in Augmented Intelligence and Generative AI

**Problem Definition**

Pharmaceutical companies receive a high volume of patient inquiries regarding their products. These inquiries vary widely—from dosage and side effects to affordability, availability, and general support questions. In the case of **Ibrance**, a widely prescribed oncology treatment, patient inquiries often require accurate triage to determine whether they are medical in nature or fall into non-medical categories. Currently, triage is handled manually, which is time-consuming, inconsistent, and difficult to scale.

The lack of automation in this space leads to three key challenges:

* **Delayed Response Times:** Manual triage slows down timely responses to patients, potentially impacting their treatment experience.
* **Inconsistency:** Different agents may classify the same inquiry differently, creating variability in routing and responses.
* **Resource Constraints:** As patient awareness and engagement grow, inquiry volume increases—but staffing levels may not scale proportionally.

Given the need for compliance, accuracy, and patient safety, automating this triage process with a transparent, scalable, and bias-aware solution is essential. This Proof of Concept (PoC) focuses exclusively on patient (non-HCP) inquiries related to Ibrance in the United States, with future scalability to HCP inquiries and other products.

**Objectives**

1. **Automated Classification:** Use an LLM to classify patient inquiries as either **Medical** or **Non-Medical** based on predefined intent categories.
2. **Intent Subtyping:** Further categorize each inquiry into specific **intent types** (e.g., "Side Effects", "Product Access", "Affordability", "Dosage").
3. **Static Response Mapping:** Automatically return a **pre-approved static response** for each non-medical intent type.
4. **Evidence-Grounded Medical Drafts:** Generate draft responses for medical inquiries using **Ibrance US Prescribing Information (PI)** only.
5. **Bias Monitoring and Mitigation:** Ensure fair and equitable triage decisions across diverse patient language inputs and demographics.
6. **Scalability Design:** Build a modular system ready to expand to additional products, populations, and inquiry sources.

**Technical Design**

The solution will be developed in Python using Streamlit for rapid UI development. The proposed system consists of the following layers:

1. **User Interface (Streamlit App):**
   * Text input field for the patient’s inquiry.
   * Output displays classification (Medical/Non-Medical), intent type, system response, and confidence level.
   * Feedback option for users to flag incorrect classifications or responses.
2. **Preprocessing Layer:**
   * Normalizes incoming text (spelling, punctuation, etc.).
   * Detects common informal or ambiguous phrasing typical in patient language.
3. **Inference Engine (LLM):**
   * Uses an LLM (e.g., GPT-4, Claude) to:
     + Classify the inquiry as Medical or Non-Medical.
     + Determine the specific intent subtype using a predefined schema.
   * Medical intents trigger a Retrieval-Augmented Generation (RAG) pipeline:
     + Embeddings are generated from the US Ibrance PI documents.
     + The LLM retrieves relevant excerpts and generates a grounded, compliant draft response.
4. **Static Response Handler:**
   * Non-medical intents are linked to static responses from a configured response library (e.g., FAQs on cost or availability).
   * These responses are instantly returned without generation.
5. **Post-Processing & Rule Layer:**
   * Applies rules based on keywords, confidence thresholds, and fallback logic.
   * Ensures that critical medical intents (e.g., side effects) are never misclassified as non-medical.
6. **Bias Detection & Feedback Loop:**
   * Tracks input metadata and classification decisions to detect systematic bias (e.g., differences in how informal vs. formal language is classified).
   * Enables retraining or prompt adjustments based on flagged issues.

**Anticipated Business Value**

1. **Efficiency and Speed:**  
   Dramatically reduces triage time by automating classification and static response delivery, improving the patient experience.
2. **Scalability:**  
   Supports growing inquiry volumes without needing proportional increases in staff.
3. **Consistency and Compliance:**  
   Ensures every response is based on a consistent logic chain and grounded in regulatory-approved PI content or vetted static text.
4. **Bias Control and Trust:**  
   Proactively tracks fairness and provides transparency into the triage process, essential for patient trust and corporate responsibility.
5. **Foundation for Expansion:**  
   Modular design supports future inclusion of HCP inquiries, other brands, and multilingual support.

**Evaluation Metrics**

* **Classification Accuracy:** Agreement between LLM and expert triage on Medical/Non-Medical split and correct intent subtype.
* **Response Quality:** Human review of both static and generated responses for clarity, factuality, and compliance.
* **Bias Metrics:** Analysis of performance consistency across language tone, phrasing, and demographics.
* **Time to Resolution:** Average reduction in triage and response handling time per inquiry.
* **User Feedback:** Satisfaction scores from agents reviewing or using the responses.

**Conclusion**

This LLM-powered triage tool represents a focused, practical application of generative AI within Medical Information. By limiting initial scope to **US-based patient inquiries about Ibrance**, the project becomes immediately actionable while delivering measurable improvements in speed, consistency, and compliance. With bias safeguards and modular extensibility, the solution not only addresses today’s needs but also lays the groundwork for more advanced and equitable AI support in patient engagement.