A cohort builder is a GUI based application that allows users to identify groups of individuals who meet certain inclusion & exclusion criteria. For example the GDC cohort builder allows users to identify samples across studies that meet specific demographic, treatment, outcome, etc parameters (ref - GDC Analysis Center). Cohorting is often the first step in identifying market segments for designing specific outreach programs.
Most Cohorting applications are filter based ie users

Aim - Build a chorting application that allows users to identify subgroups using a LLM-based text-to-query tool

As the developer your job is to achieve the following—

- Parse user intent & Convert natural language to backend queries
- Ensure the UX is topnotch and that the user is not confused
- Test the system such that you are sure that system is behaving as expected.

Tasks—

- Who is the ideal user persona
- Define a method to parse user intent & handle ambiguity or incomplete queries?
  e.g., “show elderly patients”
- How would you test that the system is not hallucinating?
- Define a framework to evaluate the performance i.e. how well it compares with gold standard filter based methods.
- Sketch or describe the user flow - bonus point

Assumptions—

- All data is already harmonised
- You also have programmatic access to the harmonised data to cross check responses from the system
- You can use any LLM

Success Metrics

- The system visually updates filter status (e.g., number of patients matching criteria) live as users refine or type queries.
- Users can iteratively adjust or clarify their queries with minimal friction, ideally resolving ambiguities within 2–3 interactions.

- Interface elements (filters, chips, summaries) clearly display inclusion/exclusion criteria so users never feel “lost” in query construction.
- **Bonus**: Show visual cues (e.g., confidence bars, icons, or highlights) indicating how well the LLM understood and mapped the user’s intent.
- **Bonus**: Users can instantly preview cohort size and summary statistics (age distribution, gender, treatment count, etc.) without switching context.