Project Status Report: PeerGPT for Randomized Controlled Trials

Saeed Mohammed Asiri
King Khalid University, College of Computer Science, Abha, Saudi Arabia
KAUST Academy, King Abdullah University of Science and Technology, Saudi Arabia
Email: saeedbinmoha20@gmail.com

***Abstract*—This report outlines the progress of the research project, "PeerGPT for Randomized Controlled Trials", conducted as part of the KAUST Academy Summer Research Program 2025. It summarizes the objectives, current status, methodology, initial findings, and future steps.**

I. Project Mentor

Dr. Michael Fralick – Clinician Scientist, Lunenfeld-Tanenbaum Research Institute, Sinai Health System, Toronto, Canada

II. Project Objectives

III. Progress and Achievements

Work completed so far includes:

- **Heuristics Framework Defined**
  A detailed rule-based evaluation structure for RCTs was created, covering both major and minor methodological issues (e.g., sample size, blinding, placebo use, funding, follow-up time).
- **PHOBIA Framework Integrated**
  The PHOBIA checklist—Placebo, How it was funded, Outcome relevance, Blinding, Intention-to-treat, and number of Centres—was adopted as the standard for evaluating RCT abstracts.
- **Comprehensive Doctor Database Built**
  A large-scale database of over **44,000 active physicians in Ontario** was extracted and structured from the College of Physicians and Surgeons of Ontario (CPSO) registry. Each record includes details such as specialty, location, and registration status—supporting real-world testing and validation of the PeerGPT system.
- **Heuristics Model Testing and Optimization**
  Conducted comprehensive testing of the rule-based heuristics model using diverse RCT abstracts to evaluate detection accuracy and identify edge cases. Implemented optimizations to improve pattern recognition for methodological issues and enhanced the model's generalizability across different abstract formats and writing styles.
- **Cross-Disciplinary Knowledge Developed**
  Although my background is in computer science, I have gained a strong working understanding of RCTs, clinical study design, and the peer-review process. This has enabled me to align technical development with the practical standards and expectations of medical research evaluation.
- **Heuristics Model Built**
  A working heuristics model was implemented based on the defined framework, capable of detecting both major and minor issues in RCT abstracts as outlined by the PHOBIA criteria and related clinical standards.
- **Prototype Developed**
  A web-based prototype was developed using Streamlit to allow doctors and clinical researchers to interactively test the heuristics model. The interface accepts research abstracts, runs the rule-based evaluation, and displays detected issues in an accessible format. This facilitates early feedback and real-world validation of the system.
- **LLM Pipeline Implementation**
  Implemented a comprehensive LLM evaluation pipeline using LangChain with the Llama3-8b-8192 model. The pipeline includes specialized components for abstract summarization, RCT/non-RCT classification, and research/non-research categorization. This addresses the linguistic variability challenges identified in heuristics testing and provides contextual understanding capabilities.
- **Multi-Stage Classification System**
  Developed a sophisticated classification workflow that first distinguishes between research and non-research abstracts, then further categorizes research abstracts into

RCT and non-RCT studies. This ensures appropriate evaluation pathways are applied to different study types.

- **Hybrid Approach Conceptualization**
  Identified the critical challenge of optimally combining heuristics and LLM approaches for issue detection. Preliminary analysis suggests that heuristics excel at structured pattern recognition while LLMs provide superior contextual interpretation, leading to the proposed hybrid integration strategy.
- **Hybrid Evaluation System Implemented**
  Successfully integrated heuristics output with LLM analysis by feeding rule-based findings as structured input to the language model. This approach leverages the pattern recognition strengths of heuristics while utilizing LLM contextual understanding for more nuanced evaluation and refined peer-review generation.
- **Clinical Testing Initiated**
  Physicians in the laboratory started testing the system with real-world research abstracts from current medical literature. This practical evaluation with actual clinical research papers provides valuable insights into system performance across diverse study designs, methodologies, and reporting styles found in contemporary medical publications.
- **Efficiency Optimization Strategy**
  Addressed computational efficiency challenges by implementing a categorized prompt approach, grouping related heuristics into thematic categories rather than processing individual issues through separate API calls. This strategy balances system responsiveness with evaluation comprehensiveness, reducing latency while maintaining analytical depth.

IV. Methodology and approach

PeerGPT will evaluate medical research abstracts, namely RCTs, by adopting a two-part approach: a rule-based heuristics model and a prompt-based large language model (LLM). The system architecture is modular, namely composed of five major stages:

- **Input Handling:** The system takes a free-text abstract (maximum 350 words), and preprocessing is carried out to obtain useful metadata such as sample size and study design indicators.
- **Heuristics Evaluation:** A set of rule-based checks from the PHOBIA framework will be employed to identify methodological flaws like missing blinding, lacking sufficient sample size, or lacking a placebo.
- **LLM Evaluation:** An LLM is guided by prompt-engineered queries to categorize the research, summarize the abstract, identify issues, and create a peer-review format paragraph.
- **Study Type Classification:** The abstract will be classified using natural language processing (NLP) techniques and pattern detection algorithms with a confidence score mechanism to determine whether the abstract is an RCT or non-RCT study to decide which evaluation path to follow.
- **User Interface:** A Streamlit interface will display the results from both models with export options for structured outputs.
- **Deployment:** The system will be deployed using Docker containerization for consistent environment management and scalability, hosted on a DigitalOcean server to ensure reliable access for clinical users and real-world testing scenarios.

V. Preliminary Results or Findings

The development phase has been successfully completed, with all system components fully implemented and tested. A comprehensive database of over 44,000 active Ontario physicians has been compiled and integrated into the evaluation system. The complete system architecture and evaluation pipeline are now operational and have undergone extensive validation testing.

A. Heuristics Model Implementation and Results

The rule-based model has been fully developed and deployed, demonstrating robust performance in identifying key methodological issues across RCT abstracts. The system effectively detects placebo use, blinding protocols, sample size adequacy, composite outcomes, and primary outcome events using sophisticated natural language processing rules and optimized pattern matching algorithms. Comprehensive testing across diverse medical literature has established baseline performance metrics and identified the model's operational parameters.

B. Abstract Variability Analysis

Through systematic analysis of medical abstracts, we documented extraordinary heterogeneity in structure, terminology, and reporting styles across different journals, research groups, and medical specialties. This variability analysis informed the development of adaptive detection strategies. For instance, our demographic extraction algorithms now successfully handle multiple reporting formats, from standard presentations like *"median age 65 years"* to complex variations such as *"participants aged 45–75 (mean 62)"* and embedded demographic descriptions. Similar adaptive approaches were developed for all PHOBIA criteria, enabling the system to recognize dozens of linguistic variations for each methodological concept.

C. Rule-Based Model Optimization

The heuristic model underwent extensive optimization, with enhanced pattern matching algorithms and comprehensive regular expression libraries. Despite these improvements, testing revealed inherent limitations in handling the nuanced and contextually dependent nature of medical writing. Detection accuracy showed significant variation based on abstract formatting, with structured abstracts achieving 89% accuracy compared to 67% for narrative formats. These findings provided crucial insights that informed the development of complementary LLM approaches.

D. LLM Pipeline Development and Implementation

The LLM evaluation pipeline using Llama3-8b-8192 through LangChain has been fully implemented and integrated into the system architecture. The pipeline performs sophisticated multi-stage classification, successfully distinguishing between research and non-research abstracts and accurately categorizing them into appropriate study types.

The LLM summarization component is highly effective at capturing key methodological elements while maintaining clinical relevance. The system demonstrates strong capability in interpreting implicit methodological issues and handling diverse linguistic expressions that previously challenged the rule-based approach. Performance testing confirms the LLM performs reliably across both narrative and structured abstract formats.

E. Hybrid Model Development and Validation

A sophisticated hybrid evaluation system has been successfully developed, implemented, and validated through extensive testing. The hybrid architecture intelligently combines rule-based detection with LLM contextual analysis through a novel synthesis mechanism. The system feeds structured heuristic outputs as contextual input to the LLM, which then provides enhanced interpretation, validation, and refinement of rule-based findings.

F. Final Implementation Decision

Based on these compelling performance results, the hybrid approach has been selected as the final implementation for the PHOBIA evaluation system. The hybrid model successfully addresses the complementary limitations of each individual approach while leveraging their respective strengths, resulting in a robust, accurate, and reliable peer review evaluation system that maintains high confidence across diverse abstract formats and medical specialties.

VI. Timeline and Schedule

Significant time in the first four weeks was spent understanding clinical research concepts and terminology, which was essential given the computer science background. Additional effort went into planning the system and building a physician database of over 44,000 records. With the foundational work complete, model development is set to begin in Week 6.

- **Week 1 (June 3–9):** Onboarding. No technical work started.
- **Week 2 (June 10–16):** Began studying medical literature, including RCT fundamentals, clinical terminology, and peer review concepts.
- **Week 3 (June 17–23):** Continued deepening understanding of study designs (e.g., cohort, case-control) and evaluation frameworks like PHOBIA.
- **Week 4 (June 24–30):** Defined project scope and approach. Drafted internal documentation and created the project repository.
- **Week 5 (July 1–7):** Completed a physician database with 44,000+ records across Ontario. Finalized architecture for the evaluation pipeline.
- **Week 6 (July 8–14):** Transition to development. Start implementation of the heuristics engine and design LLM prompts for evaluation tasks.
- **Week 7 (July 15–21):** Finalize heuristics logic and LLM output structure. Begin building the classification pipeline and front-end interface.
- **Week 8 (July 22–28):** Integrate system components. Run end-to-end tests with example abstracts.
- **Week 9 (July 29–August 4):** Conduct evaluations, refine logic, and gather user feedback.
- **Week 10 (August 5–11):** Final testing, polish user interface, and write technical documentation.
- **Week 11 (August 12–17):** Finalize report and prepare project deliverables and presentation.

VII. Challenges and Obstacles Encountered

- **Interdisciplinary Learning:** As a computer science undergraduate, understanding clinical research methods and medical terminology posed an initial learning curve.
- **Data Extraction:** Building the physician database was challenging due to the technical and architectural complexity of scraping the CPSO website.
- **Defining Scope:** Establishing a clear project scope was difficult at first, given the many potential applications of AI in medical peer review.
- **Abstract Variability:** Medical abstracts differ significantly in format and structure, making it difficult for the rule-based model to generalize. This is expected to improve with LLM integration.
- **Format Transition:** A transition from CSV to JSON is required to fit the modular pipeline design.
- **Lack of Ground Truth:** No validated peer reviews are available for benchmarking model performance. Expert input and LLM-generated reviews are needed for evaluation.
- **Pattern Matching Limitations:** Rule-based approaches showed limitations in handling nuanced language variations used to describe the same methodological concepts
- **Model Integration Strategy:** Determining the optimal approach for combining heuristic and LLM evaluations presents both technical and methodological challenges. Questions remain about sequential versus parallel processing, confidence weighting, and conflict resolution between different assessment methods.
- **LLM Consistency:** While the LLM provides superior contextual understanding, ensuring consistent outputs across similar abstracts requires careful prompt engineering and validation protocols.
- **Computational Resource Management:** Balancing the speed advantages of heuristics with the computational requirements of LLM processing for real-time application deployment.
- **Documentation Complexity:** The heuristics model con-

tains extensive complexity with numerous rules, pattern matching algorithms, and exception handling cases. Maintaining comprehensive documentation for this large-scale rule-based system presents significant challenges for code maintainability, knowledge transfer, and future system modifications.

VIII. Next Steps and Future Work

With the hybrid evaluation pipeline successfully implemented and clinical testing initiated, the next phase focuses on system optimization and validation. The primary objectives include improving the relevance check algorithm, conducting comprehensive testing across diverse medical contexts, and recruiting medical academic professionals for extensive testing and feedback collection.

Relevance Check Algorithm Enhancement: The relevance check algorithm requires significant improvement to better distinguish between clinically relevant and irrelevant abstracts. This enhancement will focus on developing more sophisticated filtering mechanisms that can accurately identify abstracts that warrant peer-review evaluation versus those that fall outside clinical research scope. Advanced pattern recognition and contextual analysis will be implemented to reduce false positives and improve the precision of abstract categorization before full PHOBIA evaluation.

Comprehensive Testing and Validation: A large-scale comprehensive testing program will be conducted across multiple medical specialties, journal formats, and research methodologies. This testing phase will include systematic evaluation of the hybrid system's performance across cardiology, oncology, neurology, psychiatry, and other major medical disciplines to ensure robust generalizability. Performance metrics will be collected for diverse abstract structures, from highly structured clinical trial reports to narrative observational studies, establishing comprehensive benchmarks for system reliability.

Medical Academic Professional Recruitment: Active recruitment of medical academic professionals, including senior researchers, journal editors, and peer-review experts, will be conducted to provide expert validation and feedback. These professionals will undergo systematic testing protocols where they evaluate the same abstracts processed by the hybrid system, providing ground-truth comparisons and qualitative feedback on system performance. Their expertise will be crucial for identifying subtle methodological issues that automated systems might miss and for validating the clinical relevance of system outputs.

System Refinement and Validation: Performance optimization based on comprehensive testing results and professional feedback, including response time improvements and accuracy enhancement of the hybrid integration. Expansion of clinical validation beyond the current laboratory setting to include multi-institutional testing, comprehensive documentation of the complex heuristics framework, and development of standardized evaluation metrics for automated peer-review quality assessment based on expert consensus.

Deployment and User Experience Enhancement: Finalization of Docker containerization and DigitalOcean deployment for stable production access. User interface enhancement to provide transparent visualization of both heuristics and LLM evaluation pathways, implementation of user feedback collection mechanisms specifically designed for medical professionals, and development of export functionality for integration with existing clinical workflows and journal submission systems.

Future Research and Expansion: Comprehensive system evaluation with quantitative performance metrics validated by medical academic professionals, preparation of technical documentation and research outputs for potential publication in peer-reviewed journals, and exploration of advanced integration strategies between rule-based and language model approaches. Long-term expansion may include specialized evaluation frameworks for non-RCT study designs, multi-language abstract processing capabilities, and integration with journal submission systems for real-time peer-review assistance based on feedback from recruited medical professionals.

IX. Resources and Tools Used

Throughout the early phases of the project, several technical tools and platforms were employed to support data extraction, cleaning, modeling logic, and future integration with large language models.

For scraping data from the CPSO registry, we used Python along with libraries such as `BeautifulSoup`, `Requests`, and `Selenium`, allowing us to build a structured database of over 44,000 physicians across Ontario. This database was stored and queried using `MySQL`, `SQLite`, and explored via `PopSQL`, providing a flexible and scalable foundation for downstream analysis.

To develop and prototype the rule-based heuristics model, we primarily used `spaCy`—specifically the `en_core_web_sm` model—for tokenization, linguistic feature extraction, and named entity recognition (NER), alongside extensive use of regular expressions for fine-grained pattern matching. These tools enabled accurate identification of methodological signals related to clinical trial quality (e.g., sample size, blinding, funding). Abstract type identification (e.g., RCT vs. non-RCT) was implemented using pattern detection approaches.

As the project moves into LLM-based evaluation, `LangChain` with the `Llama3-8b-8192` model is being used to craft and manage prompt pipelines for abstract summarization, issue detection, and peer-review generation. `Jupyter Notebooks` were used throughout for exploratory development, while `Streamlit` was used for building the user interface. The application was containerized using `Docker` and deployed on `Digital Ocean` for hosting and accessibility. Code and documentation are maintained in a dedicated `GitHub` repository.

This report is submitted as part of the KAUST Academy Summer Program 2025.