Automated Technical Information Extraction from Scientific Literature Using Artificial Intelligence

By: **Wesley Baum**, Orion Qualls & Dr. Alexey *Suslov National High Magnetic Field Laboratory*

Introduction

- Scientific literature contains vast amounts of technical information, but manual extraction is:
  - Time-consuming
  - Prone to human error
- Research output continues to grow exponentially, making efficient information processing increasingly challenging
- Modern Large Language Models (LLMs) enable:
  - Extraction and organization of technical data from research papers
  - Large-scale analysis previously impractical or financially unfeasible
  - Utilization of context and semantic information lost to traditional hard-coded systems
- Our project demonstrates this capability by:
  - Focusing on quantum computing research challenges
  - Identifying commercially available cryogenic transistors and their parameters
  - Using a state-of-the-art OpenAI LLM to:
    - Automatically scan research papers
    - Identify specific technical components
    - Verify commercial availability
- Benefits of our approach:
  - Saves valuable research time
  - Creates foundation for similar applications across other scientific fields
  - Enables technical component identification previously impossible with traditional methods

Methods

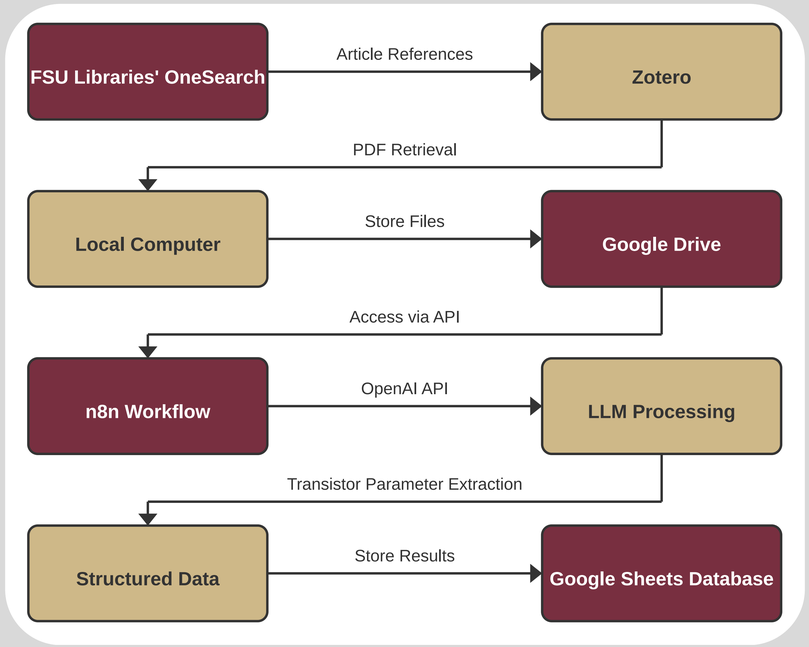
- Data Collection:
  - FSU Libraries' OneSearch platform to find and export a .ris file of articles
  - Zotero for:
    - reference management and deduplication
    - Automatic PDF retrieval and download for unique entries
- Storage & Processing:
  - Google Drive storage repository
  - Google Drive API for systematic file access/processing
- Information Extraction:
  - OpenAI API through n8n workflows
  - Specialized prompt for transistor model identification
- Data Organization:
  - Automated organization in Google Sheets
  - Structured database with standardized fields
- Pipeline Benefits:
  - Streamlines manual search processes
  - Ensures comprehensive literature coverage
  - Maintains extraction consistency

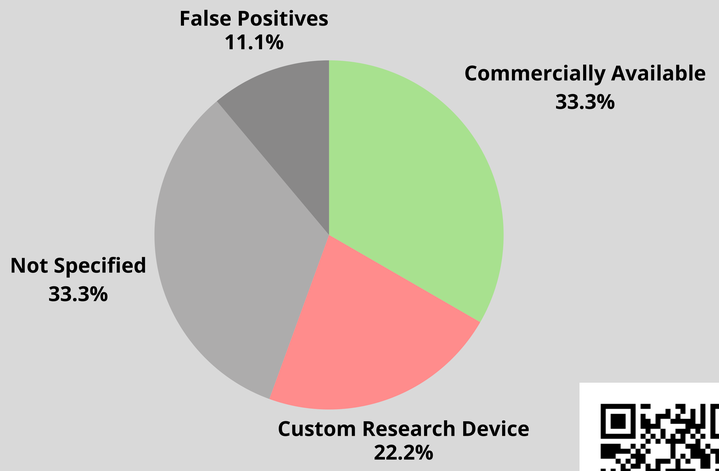
Preliminary Results

- Analysis Summary:
  - Examined 20 research papers on cryogenic transistors
  - 10 papers contained factual specific transistor information
  - 6 papers lacked details or focused on other aspects
  - 2 papers had false positives (the AI extracted operational amplifiers, not actual transistors)
- Identified Transistor Categories:
  - Commercial off-the-shelf (COTS):
    - Examples: LNC4_16B, ATF-54143, FHX13LG, BFP640, ATF-35143, and ATF34143
    - Readily available from major manufacturers
  - Custom-fabricated research devices:
    - From Fraunhofer IAF, Chalmers University, etc.
  - Unspecified transistors:
    - Focus on performance metrics rather than models
- Common Technology Types:
  - HEMTs (High Electron Mobility Transistors)
  - SiGe HBTs (Silicon-Germanium Heterojunction Bipolar Transistors)
- Key Conclusions:
  - Several COTS transistors work reliably at cryogenic temperatures
  - Significant gap between commercial and custom-fabricated devices
  - Many papers omit model information (reporting issue)
- Future Work:
  - Expand the database with a larger paper sample
  - Develop parameter verification methods

Future Directions

- Expand Search Methodology
  - Implement a more targeted and extensive search pattern
- Process full corpus (several thousand articles) beyond initial sample, pending budget approval
- Enhance Model Selection
  - Transition to more cost-effective models based on approved budget
  - Evaluate performance trade-offs between different OpenAI models
  - Optimize token usage to manage API costs
- Refine Extraction Approach
  - Modify prompts to extract only transistor models without parameter details
- Develop research agent to separately retrieve datasheet parameters from manufacturer sources
- Processing Pipeline Improvements
  - Possibly further develop n8n workflow automation for API batch processing
- Document complete methodology for educational purposes and reproducibility





References: