# CREATING AN AI LLM TOBACCO DATABASE

Kevin Burd[1], Bryan Burd[1] and Ryan Selby[2] - [1]Chemular Inc, Hudson MI, US; [2]Generative AI Solutions, Vancouver, Canada

Poster #86

Tobacco Titan™

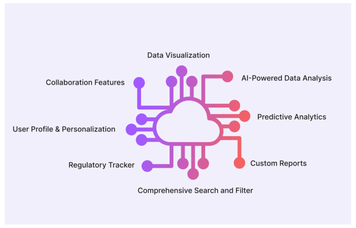
CHEMULAR

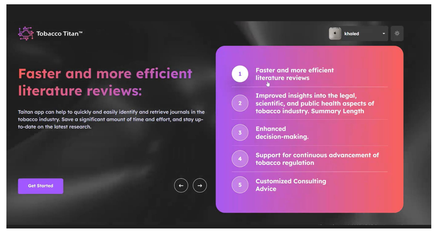
## ABSTRACT

In the rapidly evolving landscape of tobacco regulation, staying abreast of the latest research, policies, and legal developments is paramount for regulatory professionals. The vast and continuously expanding body of literature on tobacco products presents a unique challenge in accessing, organizing, and harnessing the information available. We propose the establishment of an Artificial Intelligence (AI)-driven Large Language Model (LLM) database specifically tailored to support regulatory professionals in navigating and leveraging published literature on tobacco products. The AI LLM database aims to address the gaps and inefficiencies in accessing and utilizing tobacco products literature by leveraging the power of machine learning and natural language processing techniques. We will provide an overview of the key components and benefits of the proposed database, highlighting its potential to revolutionize the way regulatory professionals engage with tobacco products literature. The AI LLM database will offer a solution by employing advanced algorithms to curate, categorize, and analyze the extensive literature on tobacco products. The database can identify patterns, detect trends, and uncover relationships within the literature, providing regulatory professionals with a deeper understanding of tobacco products' legal, scientific, and public health aspects. Such insights can assist in developing effective regulatory frameworks, evaluating the impact of existing policies, and identifying areas for further research. By harnessing the power of AI, machine learning, and natural language processing, the proposed database has the potential to transform how regulatory professionals access, analyze and utilize published literature. This innovative tool can enhance decision-making, facilitate evidence-based policy development, and support the continuous advancement of tobacco regulation in an increasingly complex and dynamic landscape.

## INTRODUCTION

The AI LLM Tobacco database is a customizable cutting-edge, web-based application designed to serve as a one-stop hub for all pertinent tobacco industry research. Our web-based application harnesses the power of advanced AI to transform how individuals and organizations interact with research and market intelligence within the global tobacco industry. Designed to be a powerful, user-centric tool, our application caters to entities seeking to enhance their decision-making, analyze the market, improve existing offerings, or inspire groundbreaking innovations.




At its core, the application is an interactive gateway into a vast trove of precise, timely, and valuable information. Its key features offer an immersive, multifaceted service comprising six vital components:

**AI-Assisted Research Outlines**: Leveraging machine learning, the application intelligently extracts main headlines and key takeaways from a wide array of research studies based on user-requested topics. Users are empowered with a succinct yet comprehensive overview, saving them the time and effort of sifting through entire papers.

**Smart Keyword Searches**: When a user enters keywords not directly linked to the results of a certain study, the AI algorithm scours the database to find these terms' appearances and context within various research papers. The application thus gives a broader, more nuanced understanding of the topic at hand.

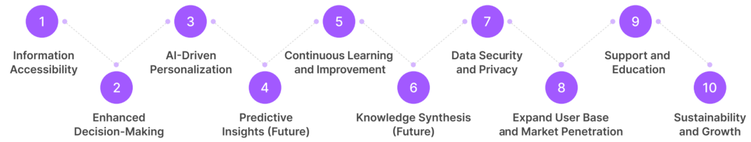
**Comprehensive Research Summary**: The application performs both the above functions across all relevant studies to a user's topic, delivering a synthesized summary of prior research. It's like having your personal AI research assistant, organizing and presenting information tailored to your needs.

**AI-Driven Comparative Analysis**: The application's intelligent engine offers a comparison of results from different research papers, drawing out points of agreement and highlighting areas of disagreement, providing a balanced view of the topic.

**Interpretation of Results**: The application's AI capabilities extend to interpreting research findings, articulating convergence and divergence in the data in an easy-to-understand format.

**Predictive Insights for Development Opportunities**: Utilizing AI's predictive abilities, the application identifies potential areas of development, intervention niches, unexplored gaps, and underserved needs. These insights pave the way for future action, guiding strategic decision-making processes.

## OBJECTIVES



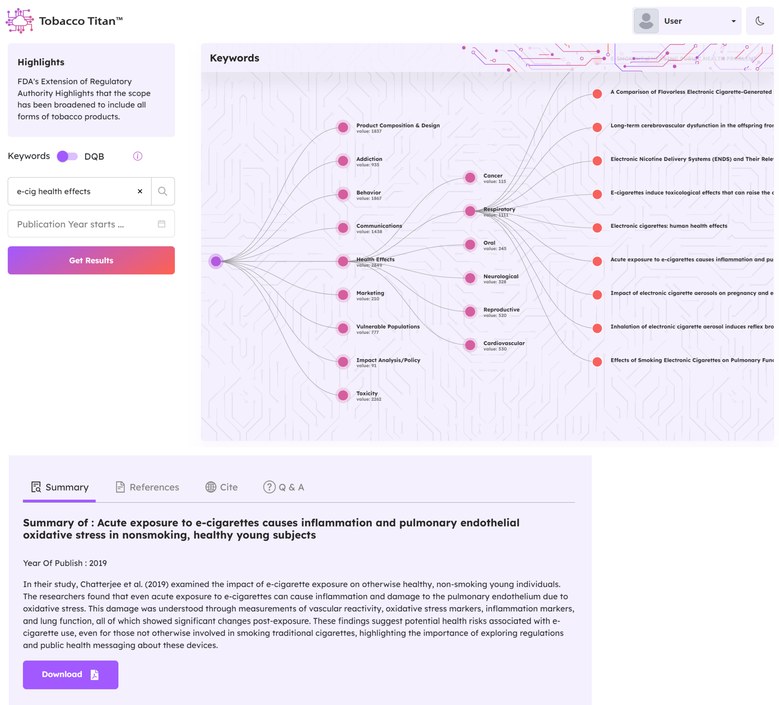

We seek to centralize vast information resources for easy access and enable data-driven decision-making through AI-powered insights. The app will offer personalized experiences by learning user preferences and will also feature predictive capabilities for future market trends. Continual AI refinements will provide ever-improving insights, while strong data security protocols will ensure user trust. Our focus extends to rapid market penetration, comprehensive user support, and long-term sustainability by frequently updating research data.

## MAIN FEATURES

- **An AI-powered Search Service:** Search bar uses artificial intelligence to deliver more relevant results, improve user experience, and offer features like natural language processing, personalization, and predictive typing. It enhances search efficiency and user satisfaction.
- **AI-Powered Data Analysis:** This feature would analyze the contents of the database to uncover insights, correlations, trends, and patterns. It should have the capability to handle natural language processing to understand the context of the documents.
- **AI-driven Question and Answer extraction:** Involves identifying the main questions within text, pre-processing, training a model to recognize key questions and their corresponding answers, facilitating efficient access to critical information within documents or conversations.
- **AI-driven Research Paper Summarization:** Involves extracting text, preprocessing, training a model to identify important features, and generating concise summaries, simplifying complex papers for quick understanding by researchers.

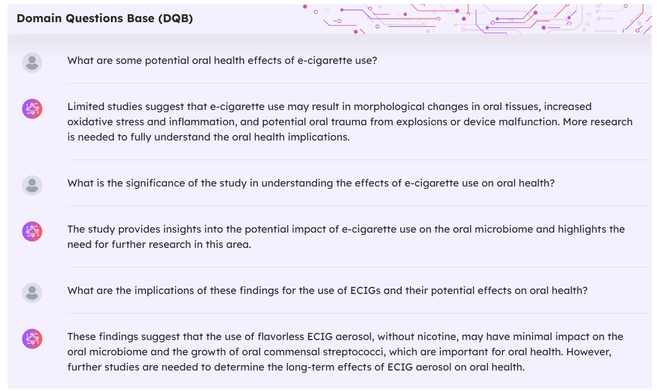
## SYSTEM OVERVIEW – KEYWORD SEARCH

Categories and subcategories were modeled off FDA Center for Tobacco Products (CTP) research priorities. Below is an example of a search using the keywords “e-cig health effects”. The values listed under each subcategory reflect a combination of 1) the number of times the keyword was identified in an article and 2) the number of publications in which the keywords appeared. The user can expand specific subcategories by clicking on the circles. In addition, a summary of each article is provided.




## SYSTEM OVERVIEW – DOMAIN QUESTIONS BASE

There is also an additional Domain Question Base (DQB) function in which the AI database can populate questions and answers for each publication. Below is an example of autogenerated questions from Isik Andrikopoulos, G., Farsalinos, K., & Poulas, K. (2019). Electronic nicotine delivery systems (ENDS) and their relevance in oral health. Toxics, 7(4), 61.




## ADDRESSED CHALLENGES

- **Data Collection and Quality:** Access to comprehensive and up to date tobacco and nicotine-related literature and documents can be challenging. In addition to ensuring that the data is of high quality, free from errors and well structured is crucial for accurate analysis.
- **Data Volume:** The volume of tobacco-related documents and literature available on the web can be overwhelming. Handling and processing large datasets efficiently is essential for web-based AI applications.
- **Text Complexity:** Tobacco literature often includes scientific jargon, medical terminology, and legal language. Understanding and processing this complex text requires advanced natural language processing (NLP) techniques.
- **Data Heterogeneity:** Documents may come in various formats, including PDFs, images, HTML, and plain text. Developing AI models that can handle diverse data formats and extract meaningful information is a challenge.
- **Multilingual Content:** Tobacco literature may be available in multiple languages. AI applications must be capable of handling multilingual content to provide comprehensive insights.
- **Semantic Understanding:** Extracting meaningful information, relationships, and context from tobacco documents and literature requires advanced semantic analysis to capture nuances and subtleties in the text.
- **Search and Retrieval Efficiency:** Developing efficient search and retrieval algorithms to provide users with relevant information quickly is a significant challenge, especially with large and diverse datasets.
- **Scalability and Performance:** As the volume of tobacco-related literature and documents grows, AI applications must be scalable and performant to handle the increasing workload.
- **Copyright and License Compliance.** If an AI web application compiles and aggregates data from multiple sources, including copyrighted ones, it must respect copyright laws. Proper attribution and compliance with any licensing terms are essential. In addition, should avoid producing content that infringes on copyrighted materials or plagiarizes existing works. Some content may be available under licenses that permit certain uses, such as Creative Commons licenses. AI applications must adhere to the terms of these licenses when using content covered by them.

## FUTURE WORK

- **Predictive Analytics:** This feature would use AI to forecast future trends based on existing data, giving users a way to anticipate potential changes in the industry and plan accordingly.
- **Regulatory Tracker:** This feature would keep track of policy changes and regulatory updates, alerting users to any changes that might affect the tobacco industry.
- **Custom Reports**: Users should be able to generate reports based on their searches and analyses, with options to customize the format and content of the report to suit their needs.
- **User Profile and Personalization**: Each user should be able to create a profile where they can save their searches, favorite documents, and customize their app experience according to their preferences and needs.
- **Data Visualization:** The ability to transform raw data and insights into easily digestible charts, graphs, and other visualizations to help users understand the information better.
- **Collaboration Features:** Users should have the option to share their findings, reports, and favorite documents with others, fostering a collaborative environment.

76th TSRC Conference 2023, Norfolk, VA USA