Firstshift

tursio

SCOGEN:

Domain Specific Generative AI to Democratize Supply Chain Insights

Executive Summary

Generative AI has shown a lot of promise in enterprise domains like marketing, customer support etc. Most of the applications in those areas have mostly followed a pattern of using unstructured data to answer queries or generate new content. The typical approach there is to use retrieval augmented generation (RAG) based applications that create vector embeddings and provide a query interface to answer questions or generate new content (marketing emails, etc.).

Democratizing Supply Chain insights and recommendations

We believe that if deployed thoughtfully, Generative AI can play a significant role in the Supply Chain domain as well – in democratizing Supply Chain Insights to a broader set of stakeholders within an enterprise. However, accuracy and recall of these insights are of primary importance as decisions driven by these insights have significant business impact.

Imagine a Supply Chain Planner having access to a COPILOT that she or he can count on to answer ad-hoc questions that they run into as they are addressing Supply Chain issues –questions like “Why is my inventory in Chicago DC lower than expected for Air Jordan Shoes?”, “What has been the demand trend for running shorts in the Northeast?”.

Why does a purely RAG based approach not cut it in Supply Chain Planning?

In a domain like Supply chain planning, most of the data is in structured data repositories, such as ERP, CRM, and data warehouses, that have large and complex schemas with new data coming in all the time. Therefore, planners need to quickly zoom in into the relevant portions of data, or the planning points, and analyze them in multiple ways, e.g., ask “what”, “where”, and “why” questions. The traditional approach is to bake some of these analysis into static dashboards that planners can explore interactively. Unfortunately, this leads to creating too many dashboards that are unwieldy for getting the insights and continued use of spreadsheets for offline analysis. Generative AI can transform this process by interpreting user intent in natural language and helping navigate and analyze the planning points quickly.

Applying RAG to structured data repositories requires copying data from those repositories into vector databases – a challenge for the supply chain domain where data is large and constantly evolving. Furthermore, multi-tenant environments make it harder to move data around since the SaaS provider already has strict data privacy agreements in place.

Also, the cost of RAG infrastructure grows linearly with the size of the data making it difficult for large scale supply chain deployments.

Prompting language models to interpret natural language questions into structured queries (text-to-SQL) is another popular approach. It requires passing the database schemas to the Large Language Models (LLMs) for generating the SQL. Unfortunately, this only works for simpler schemas and questions that clearly refer to the appropriate schema objects. Supply chain schemas are way more complex, and planners hardly understand how the underlying schemas have been implemented. Text-to-SQL can therefore end up with very low accuracies, less than 50% in our experiments, which is unacceptable.

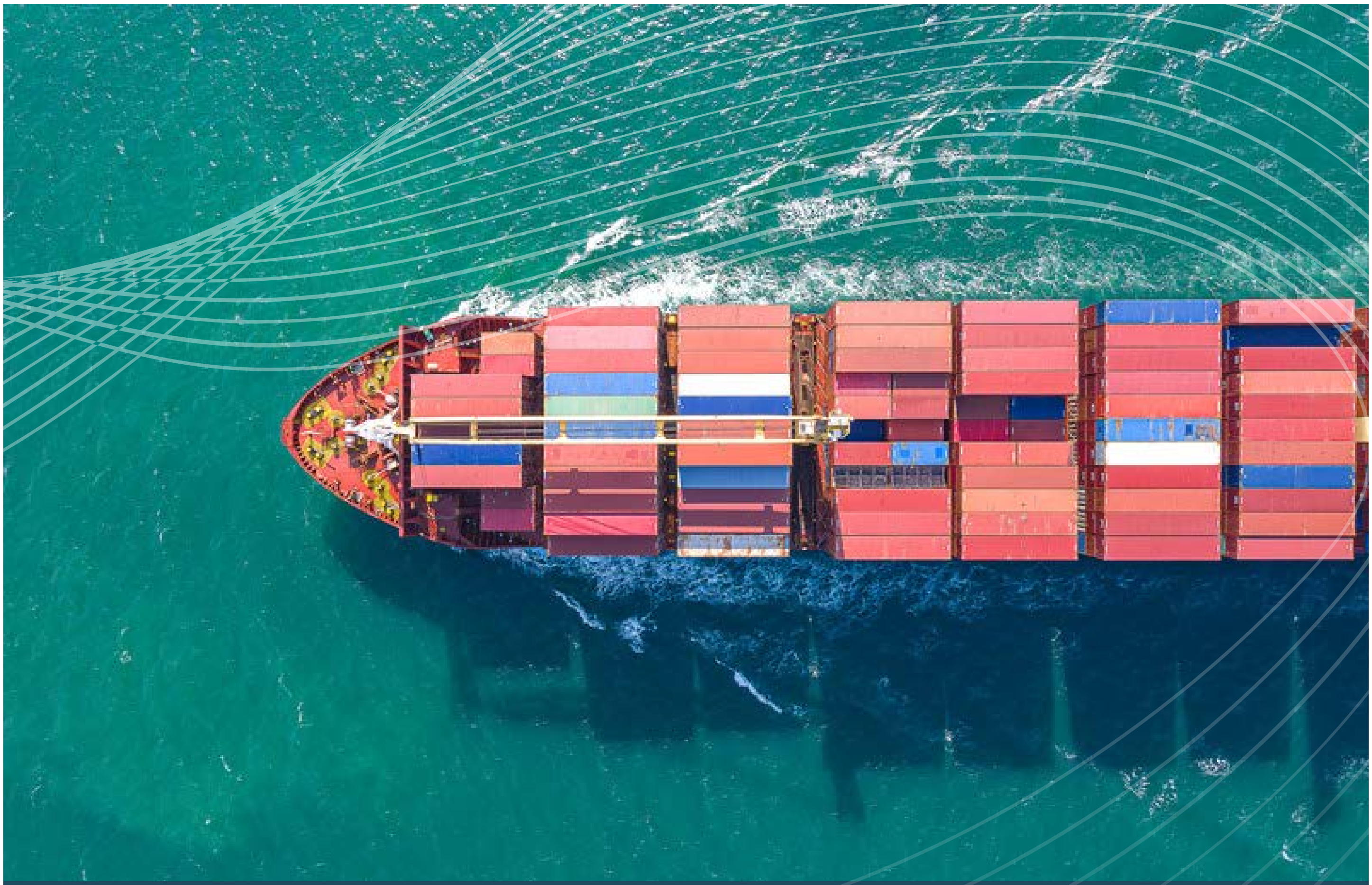
FirstShift and Tursio have partnered to build domain-specific generative AI for supply chain and operations (SCOGEN). The core principles of our design are four-fold: (1) leverage specialized data structures to interpret and respond to questions, (2) disambiguate the intent based on domain-specific vocabulary, (3) push down questions as query operators into the underlying structured data repositories, and (4) optimize for interactive performance and low cost. We describe each of these in more detail below.

Specialized data models: Supply chain data involves specialized data structures that have been carefully implemented by domain experts. Both RAG and text-to-SQL completely miss these data models and fail to contextualize the questions based on what exists in the structured data repositories. In SCOGEN, we train small models that constrain the interpretation to well-defined data models, the relationships between them, and derived data models on top of them.

Domain-specific compiler: Supply chain vocabulary is well understood by the planners, but it is quite different from the object names in the underlying schema. It includes dimensions and measures of interest, querying patterns, time trends, expected insights, visual charts, and navigation/exploration hints. Incorporating this domain knowledge is crucial for the planner productivity. SCOGEN small models leverage supply chain domain knowledge to compile natural language questions into semantically correct data operations.

Database-native runtime: Once the natural language has been interpreted accurately, SCOGEN pushes them down into existing databases for execution. This means SCOGEN runs the interpreted queries as well-defined operator trees over one or more tables, schemas, or data sources. It involves building the operator trees, rewriting them for the best possible execution, and deciding what to run where. Database-native runtime is important for multi-tenant environments where data cannot leave the tenant boundaries.

Performance and cost optimizer: Supply chain queries are complex, and they process large volumes of data. Therefore, it is critical to optimize them for performance and cost. In SCOGEN, we have built an optimizer to identify and reuse portions of computations that are shared across different questions, thus avoiding both LLM and data processing latencies/costs. Doing this consistently (zero errors) and at scale (arbitrarily complex query trees) is non-trivial. Our optimizer was able to bring down latencies from 30 seconds to under 2 seconds, a 15x speedup, and costs from linear to near flat.



By taking a domain-specific approach for supply chain, we have been able to make generative AI practical for planners. With SCOGEN, they can now worry less about learning the tools and focus more on getting the insights and making the right decisions.

firstshift.ai | **tursio.ai**