iMerit

CASE STUDY

FIRST-TO-MARKET
DIALOGUE-OPTIMIZED
LLM

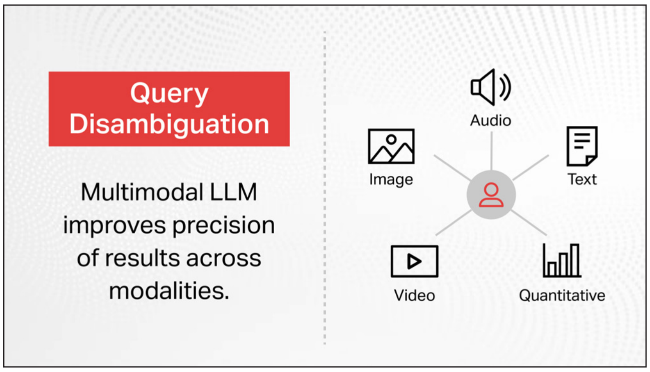
FINE-TUNING MULTIMODAL GEN AI BUSINESS INTELLIGENCE PLATFORM

Top cloud computing company consulted iMerit to augment and improve their LLM-powered business intelligence platform.

THE CHALLENGE

To fine-tune an LLM for a natural-language based business intelligence platform, this top cloud computing company came to iMerit looking for a training data augmentation strategy. The customer needed to enhance their engine's ability to dynamically disambiguate increasingly complex user queries, in order to enable nontechnical stakeholders to interact with data intelligence tools. They had only 3 months to scope the requirements, develop the annotation UI, design the workflows, curate the dataset, and complete the fine-tuning, before a high-stakes public release and showcase at a global AI event. They required an agile, trustworthy partner, with the expertise to perform highly nuanced content generation and linguistic analysis in a short timeframe.

As the customer could not train the model using private end-user data, they relied on a synthetic data strategy, systematically generating and augmenting multimodal synthetic training data units to fine-tune the model for dialogue-optimized information retrieval, narration, and abstractive summarization. The customer needed to fine-tune their model to handle queries across 10 industries, representing a variety of intents, styles, and sometimes anomalous syntactic strategies.



Query Disambiguation

Multimodal LLM improves precision of results across modalities.

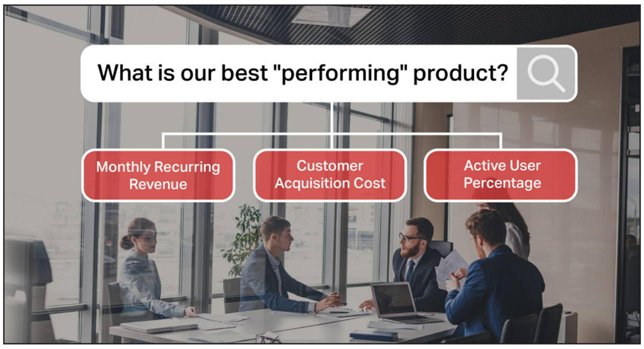
They had a proprietary data annotation tool, but limited capabilities to implement the custom UI and rigorous quality control processes required for nuanced multimodal data augmentation workflows.

THE SOLUTION

After a consultation, iMerit developed a strategy to create domain-specific supervised fine-tuning datasets that would aid their model in handling ambiguity, variation, and domain-specific concepts. To overcome the shortage in usable data, iMerit assembled a team of content specialists to create >50,000 training data units across approximately 10 industries, including healthcare, sports analytics, and advertising. The specialists received custom training in syntactic and semantic analysis, as well as domain-specific concepts, in order to identify the components of a hypothetical business intelligence query, map structured queries (e.g., SQL or noSQL) to multiple natural language paraphrases, and manipulate the linguistic components of the paraphrases to create the augmented corpus.

The project was divided into 8 workflows, each focused on manipulating the corpus by creating paraphrases exploiting different patterns of linguistic ambiguity in the queries. Additional workflows focused on creating novel abstractive summaries of tables, charts, and other data visualizations, to add further value to the product. During the process, iMerit analysts also curated and pruned the synthetic corpus of multimodal training data units, to ensure that they were valid, well-formed, and plausible in the context of the target industries.

To support the multimodal annotation and rigorous quality requirements of the corpus augmentation project, iMerit implemented the project on its proprietary Ango Hub annotation platform. Ango's workflow customization and quality auditing capabilities allowed iMerit to define custom qualitative evaluation rubrics, where iMerit and client quality leads collaborated on scoring the output. The custom workflows and reports enabled stakeholders to detect and intervene in anomalies early, and to ensure mutual trust in the quality of the corpus, reducing overhead, and leading to a successful on-schedule preview release of the product.


What is our best "performing" product?
Monthly Recurring Revenue
Customer Acquisition Cost
Active User Percentage

THE RESULT

The customer met their goals of fine-tuning the model and showcasing the product in a highly publicized release at a global AI conference. The product established an industry precedent for a dialogue-optimized, natural language-based business intelligence platform, and received positive feedback from early adopters.

> *“We couldn’t access our customer data, and had a shortage in usable data. We needed help sourcing queries that were specific enough for our use case.”*
>
> *- Head of Product*

About iMerit

iMerit provides end-to-end data labeling services to Fortune 500 companies in a wide array of industries including agricultural AI, autonomous vehicles, commerce, geospatial, manufacturing, government, financial services, medical AI and technology. iMerit employs more than 5,500 full-time data annotation experts in Bhutan, Europe, India and the United States.

info@imerit.net | imerit.net

LEARN MORE