### 1. Please provide a brief overview of your solution/application's AI capabilities (distinguish Gen AI vs. other if applicable)? What benefits will it bring to the business?

**Answer:** Our solution utilizes Large Language Models (LLMs) for extracting ground truth, key terms, and key term definitions. It offers enhanced accuracy and efficiency in data extraction, leading to improved decision-making and reduced manual effort.

### 2. Please list the underlying AI service(s) used if known (e.g., OpenAI) or write "Do not know."

**Answer:** We are using Azure OpenAI services.

### 3. What AI model is being used? Is it supervised, reinforced, or unsupervised? Please describe why this was chosen.

**Answer:** We use Large Language Models, which are typically fine-tuned on supervised data for performance. This approach was chosen for its ability to handle complex language tasks effectively.

### 4. Is the model used as batch or real time? If real-time, how have you optimized the model to ensure scaling and concurrency needs?

**Answer:** The model is used in real-time. We optimize it by employing auto-scaling and load balancing to handle increased requests efficiently.

### 5. Describe Model CI/CD process code, model/data versioning, experiment tracking, release mechanism.

**Answer:** We use version control systems for code and data, and we track experiments using a logging framework. Our release mechanism follows a CI/CD pipeline with automated testing and deployment.

### 6. How do you plan to monitor performance and data drift in production?

**Answer:** We will implement performance monitoring tools and conduct regular evaluations to detect data drift, ensuring the model remains accurate over time.

### 7. Describe your model training process - which environment (UAT/Prod) is it being trained at?

**Answer:** The model is trained in a controlled UAT environment before being deployed to production to ensure reliability.

### 8. What accuracy metrics are being used and recorded?

**Answer:** For our AI solution, we utilize several accuracy metrics to evaluate model performance:

1. **Completeness & Correctness:**
   * **Correctness** measures the percentage of accurate predictions against ground truth.
   * **Completeness** assesses how comprehensively the model captures required information.
2. **Precision & Recall:**
   * **Precision** is the proportion of correctly identified terms out of all predicted.
   * **Recall** is the proportion of correctly identified terms out of total relevant items.
3. **F1 Score:** The harmonic mean of precision and recall, balancing both metrics.
4. **ROUGE-L Score:** Measures overlap between extracted terms and ground truth, indicating completeness.
5. **Count-Based Metrics:**
   * **Count-Based Recall** and **Precision** ensure all expected items are identified and only correct items are captured.
6. **Performance Benchmarks:** Predefined thresholds monitor metric performance during fine-tuning.

### 9. Do you retrain the model? If yes, describe frequency, environments, toolsets.

**Answer:**

### 10. Does training or inference require specialized hardware - GPU / TPU?

**Answer:**

### 11. How are you logging model explainability/feature importance?

**Answer:** We log explainability metrics and feature importance using tools that integrate with our model to provide insights into predictions.

### 12. Has the MRM approval been secured?

**Answer:** Yes, MRM approval has been secured.

### 13. What are your sources of data? How is the source data quality assured?

**Answer:** Our data sources include internal databases and publicly available datasets. Quality is assured through regular validation checks and data cleaning processes.

**14. List all of the sourced Private Data items**

**Answer:**

### 14. Does your solution leverage Foundation (Gen AI) Models?

**Answer:** Yes, our solution leverages foundational models like those from Azure OpenAI.

### 15. What measures do you have in place to prevent data leakage and ensure confidentiality during model training and inference?

**Answer:** We implement strict access controls, data anonymization techniques, and encryption protocols to protect sensitive data.

### 16. Please describe your architecture, aligning with the published patterns - RAG, NQL to SQL, or complex LLM chains or Agent based.

**Answer:** Our architecture follows a RAG (Retrieval-Augmented Generation) pattern, which combines retrieval methods with generative capabilities for enhanced results.

### 17. How are you addressing Responsible AI concerns in communicating with LLM?

**i.e. prompt injections, sensitive/PII leakage, off-topic discussion, profanity/toxic prevention and hallucinations."**

**Answer:** We implement input validation, sensitive content filtering, and monitoring for biases and inaccuracies to address Responsible AI concerns.

### 18. How does business consume the generated data and ensure that reviews are done prior to the business use?

**Answer:** Generated data is reviewed by domain experts before being utilized in business processes to ensure accuracy and relevance.

### 19. Are you using a vector database? If yes, describe technology, purpose, which index/matching algorithm?

**Answer:**

### 20. How is testing conducted for generated data? i.e. Should cover functional, performance, and accuracy.

**Answer:** Testing is conducted through unit tests for functionality, performance benchmarks, and validation against ground truth data for accuracy.

### 21. Is the application monitored for performance & accuracy on a regular basis? Describe the evaluation metrics generated.

**Answer:** Yes, we monitor performance regularly using metrics like accuracy, precision, recall, latency, and throughput.

### 22. Is the LLM model out-of-the-box or fine-tuned? If fine-tuned, describe how training data is generated.

**Answer:** The model is fine-tuned on curated datasets generated from internal and external sources, ensuring no confidential information is included.

### 23. Describe how the model is fine-tuned and established for optimized inference.

**Answer:** We use techniques like LoRA or QLoRA for fine-tuning to minimize parameter updates and reduce inference costs without sacrificing performance.

### 24. LLM usage can become costly as usage increases, and also performance can degrade. Please describe how you plan to limit costs and improve performance over time?

**Answer:** We will monitor usage patterns, implement caching, use batch processing, and explore model distillation to manage costs while maintaining performance.