Table of Contents

# B. CÓDIGO FUENTE PRINCIPAL

## B.1 Repositorio del Proyecto

El código fuente completo del sistema RAG para recuperación semántica de documentación técnica de Microsoft Azure está disponible en el repositorio de GitHub del proyecto.

### B.1.1 Ubicación del Repositorio

**Repositorio GitHub:** [Pendiente de publicación]

### B.1.2 Estructura del Repositorio

El repositorio contiene la implementación completa del sistema, organizada en los siguientes directorios principales:

SupportModel/  
├── src/ # Código fuente principal  
│ ├── apps/ # Aplicaciones Streamlit modulares  
│ ├── core/ # Componentes centrales del sistema  
│ │ ├── qa\_pipeline.py # Pipeline principal de Q&A  
│ │ └── reranker.py # Reranking con CrossEncoder  
│ ├── data/ # Procesamiento de datos  
│ │ ├── embedding.py # Gestión de modelos de embedding  
│ │ ├── processing.py # Procesamiento de documentos  
│ │ └── extract\_links.py # Extracción de enlaces  
│ ├── evaluation/ # Framework de evaluación  
│ │ ├── metrics/ # Métricas especializadas  
│ │ └── comparison.py # Comparación de modelos  
│ ├── services/ # Servicios del sistema  
│ │ ├── auth/ # Autenticación APIs  
│ │ ├── storage/ # ChromaDB y almacenamiento  
│ │ └── answer\_generation/ # Generación de respuestas RAG  
│ └── ui/ # Interfaces de usuario  
├── Docs/ # Documentación del proyecto  
│ ├── Finales/ # Documentación final de tesis  
│ │ ├── capitulo\_\*.md # Capítulos de la tesis  
│ │ ├── anexo\_\*.md # Anexos detallados  
│ │ └── Contenidos.md # Tabla de contenidos  
│ ├── Analisis/ # Scripts de análisis  
│ │ ├── analyze\_metrics\_v2.py # Análisis de métricas  
│ │ ├── verify\_\*\_statistics.py # Verificación de estadísticas  
│ │ └── wilcoxon\_\*.py # Tests estadísticos  
│ └── README.md # Documentación de estructura  
├── colab\_data/ # Notebooks de Google Colab  
│ ├── Cumulative\_Ticket\_Evaluation.ipynb # Notebook principal  
│ ├── lib/ # Librerías modulares para Colab  
│ └── \*.parquet # Embeddings pre-calculados (ignorados)  
├── external\_helpers/ # Scripts auxiliares  
│ ├── check\_chromadb\_\*.py # Verificación de ChromaDB  
│ ├── create\_questions\_\*.py # Población de colecciones  
│ └── verify\_questions\_\*.py # Validación de datos  
├── tests/ # Tests unitarios  
├── data/ # Datos experimentales (ignorados en Git)  
│ ├── cumulative\_results\_\*.json # Resultados experimentales  
│ ├── \*.csv # Ground truth y datasets  
│ └── \*.pt # Modelos entrenados  
├── requirements.txt # Dependencias del proyecto  
├── .gitignore # Archivos ignorados por Git  
└── ARCHIVOS\_IGNORADOS.md # Documentación de archivos ignorados

### B.1.3 Componentes Principales

#### B.1.3.1 Pipeline Principal (src/core/)

* **qa\_pipeline.py**: Pipeline principal de pregunta-respuesta con métricas
* **reranker.py**: CrossEncoder con normalización sigmoid para reranking

#### B.1.3.2 Procesamiento de Datos (src/data/)

* **embedding.py**: Gestión de múltiples modelos de embedding (Ada, MPNet, MiniLM, E5-Large)
* **processing.py**: Segmentación y limpieza de documentos técnicos
* **extract\_links.py**: Extracción y normalización de enlaces de ground truth

#### B.1.3.3 Framework de Evaluación (src/evaluation/)

* **metrics/**: Módulos especializados para métricas de recuperación y RAG
* **comparison.py**: Comparación sistemática entre modelos de embedding

#### B.1.3.4 Servicios del Sistema (src/services/)

* **storage/chromadb\_utils.py**: Utilidades para ChromaDB y gestión vectorial
* **answer\_generation/ragas\_evaluation.py**: Evaluación RAG con RAGAS framework
* **auth/**: Gestión de autenticación para APIs (OpenAI, Google)

#### B.1.3.5 Aplicaciones Streamlit (src/apps/)

* **cumulative\_metrics\_results\_matplotlib.py**: Visualización de resultados experimentales
* **comparison\_page.py**: Comparación interactiva de modelos
* **main\_qa\_app.py**: Interfaz principal de consultas

#### B.1.3.6 Documentación Organizada (Docs/)

* **Finales/**: Documentación final de la tesis (capítulos y anexos)
* **Analisis/**: Scripts de análisis y verificación estadística
* **README.md**: Documentación de la estructura del proyecto

#### B.1.3.7 Scripts de Análisis (Docs/Analisis/)

* **analyze\_metrics\_v2.py**: Análisis comprehensivo de métricas de rendimiento
* **verify\_document\_statistics.py**: Verificación de estadísticas del corpus
* **wilcoxon\_detailed\_analysis.py**: Tests estadísticos de significancia

#### B.1.3.8 Notebooks Experimentales (colab\_data/)

* **Cumulative\_Ticket\_Evaluation.ipynb**: Notebook principal de evaluación experimental
* **lib/**: Librerías modulares para evaluación en Google Colab
* Implementación completa del pipeline de evaluación multi-modelo

### B.1.4 Tecnologías y Dependencias

El proyecto utiliza las siguientes tecnologías principales:

* **Python 3.8+**: Lenguaje de programación principal
* **ChromaDB 0.5.23**: Base de datos vectorial
* **Sentence-Transformers 5.0.0**: Modelos de embedding
* **OpenAI API 1.93.0**: Modelo Ada y evaluación RAG
* **Streamlit 1.46.1**: Interfaz de usuario web
* **Transformers 4.44.0**: Arquitecturas de modelos de lenguaje

### B.1.5 Reproducibilidad

El repositorio incluye:

1. **Configuración de ambiente** completa (requirements.txt)
2. **Scripts de verificación** para validar configuración
3. **Datos de ejemplo** para testing rápido
4. **Documentación detallada** de instalación y uso
5. **Notebooks ejecutables** en Google Colab

### B.1.6 Instrucciones de Acceso

Para acceder al código fuente completo:

1. **Clonar el repositorio:**

* git clone [URL\_del\_repositorio]  
  cd SupportModel

1. **Configurar el ambiente:**

* pip install -r requirements.txt

1. **Ejecutar la aplicación:**

* streamlit run src/apps/main\_qa\_app.py

### B.1.7 Licencia y Términos de Uso

El código fuente se distribuye bajo los términos establecidos para investigación académica, con las siguientes consideraciones:

* **Uso académico**: Permitido para investigación y educación
* **Datos de Microsoft**: Sujeto a términos de uso de Microsoft Learn
* **Modelos propietarios**: OpenAI Ada requiere API key válida
* **Atribución**: Citar apropiadamente en trabajos derivados

### B.1.8 Contacto y Soporte

Para consultas sobre el código fuente, implementación o extensiones:

* **Autor**: Harold Gómez
* **Institución**: [Institución académica]
* **Email**: [Email de contacto]

### B.1.9 Nota sobre Versiones

El código corresponde a la versión utilizada para la evaluación experimental reportada en este trabajo (julio 2025). Versiones posteriores pueden incluir mejoras y optimizaciones adicionales.