Table of Contents

# E. RESULTADOS DETALLADOS POR MÉTRICA

## E.1 Introducción

Este anexo presenta el análisis exhaustivo de todas las métricas evaluadas durante la investigación experimental, basándose en los datos verificables contenidos en cumulative\_results\_1753578255.json y wilcoxon\_test\_results.csv. Los resultados corresponden a la evaluación ejecutada el 26 de julio de 2025, procesando 11 preguntas de prueba distribuidas entre 4 modelos de embedding diferentes.

## E.2 Configuración Experimental

### E.2.1 Parámetros de Evaluación Verificados

{  
 "config": {  
 "num\_questions": 11,  
 "models\_evaluated": 4,  
 "reranking\_method": "crossencoder",  
 "top\_k": 10,  
 "generate\_rag\_metrics": true  
 },  
 "data\_verification": {  
 "is\_real\_data": true,  
 "no\_simulation": true,  
 "no\_random\_values": true,  
 "rag\_framework": "RAGAS\_with\_OpenAI\_API",  
 "reranking\_method": "crossencoder\_reranking"  
 }  
}

**Características del Corpus:** - **Total documentos indexados:** 187,031 chunks técnicos - **Ground truth validado:** 2,067 pares pregunta-documento - **Duración total evaluación:** 774.78 segundos (12.9 minutos) - **Framework de evaluación:** RAGAS con API de OpenAI

## E.3 Resultados por Modelo

### E.3.1 Ada (OpenAI text-embedding-ada-002)

#### E.3.1.1 Especificaciones Técnicas

* **Dimensiones:** 1,536
* **Proveedor:** OpenAI
* **Método de acceso:** API

#### E.3.1.2 Métricas de Recuperación Pre-Reranking

| Métrica | Valor | Desviación Estándar |
| --- | --- | --- |
| Precision@1 | 0.000 | ±0.000 |
| Precision@2 | 0.000 | ±0.000 |
| Precision@3 | 0.000 | ±0.000 |
| Precision@4 | 0.000 | ±0.000 |
| **Precision@5** | **0.055** | **±0.000** |
| **Recall@5** | **0.273** | **±0.000** |
| **F1@5** | **0.100** | **±0.000** |
| **NDCG@5** | **0.126** | **±0.000** |
| **MAP@5** | **0.125** | **±0.000** |
| **MRR** | **0.125** | **±0.000** |

#### E.3.1.3 Métricas de Recuperación Post-Reranking

| Métrica | Valor | Cambio vs Pre-Reranking |
| --- | --- | --- |
| **Precision@5** | **0.055** | Sin cambios |
| **Recall@5** | **0.273** | Sin cambios |
| **F1@5** | **0.100** | Sin cambios |
| **NDCG@5** | **0.162** | **+28.6%** |
| **MAP@5** | **0.125** | Sin cambios |
| **MRR** | **0.125** | Sin cambios |

#### E.3.1.4 Métricas RAG Especializadas

| Métrica | Valor | Interpretación |
| --- | --- | --- |
| **Faithfulness** | **0.482** | Consistencia factual moderada |
| **BERTScore Precision** | **0.740** | Alta precisión semántica |
| **BERTScore Recall** | **0.724** | Buen recall semántico |
| **BERTScore F1** | **0.732** | Balance semántico sólido |

### E.3.2 MPNet (multi-qa-mpnet-base-dot-v1)

#### E.3.2.1 Especificaciones Técnicas

* **Dimensiones:** 768
* **Especialización:** Question-Answering
* **Método de acceso:** Sentence-Transformers local

#### E.3.2.2 Métricas de Recuperación Pre-Reranking

| Métrica | Valor | Desviación Estándar |
| --- | --- | --- |
| **Precision@5** | **0.055** | **±0.000** |
| **Recall@5** | **0.273** | **±0.000** |
| **F1@5** | **0.100** | **±0.000** |
| **NDCG@5** | **0.108** | **±0.000** |
| **MAP@5** | **0.113** | **±0.000** |
| **MRR** | **0.082** | **±0.000** |

#### E.3.2.3 Métricas de Recuperación Post-Reranking

| Métrica | Valor | Cambio vs Pre-Reranking |
| --- | --- | --- |
| **Precision@5** | **0.055** | Sin cambios |
| **Recall@5** | **0.273** | Sin cambios |
| **F1@5** | **0.100** | Sin cambios |
| **NDCG@5** | **0.189** | **+75.0%** |
| **MAP@5** | **0.113** | Sin cambios |
| **MRR** | **0.082** | Sin cambios |

#### E.3.2.4 Métricas RAG Especializadas

| Métrica | Valor | Comparación vs Ada |
| --- | --- | --- |
| **Faithfulness** | **0.518** | **+7.5%** mejor |
| **BERTScore Precision** | **0.746** | **+0.8%** mejor |
| **BERTScore Recall** | **0.731** | **+1.0%** mejor |
| **BERTScore F1** | **0.739** | **+1.0%** mejor |

### E.3.3 MiniLM (all-MiniLM-L6-v2)

#### E.3.3.1 Especificaciones Técnicas

* **Dimensiones:** 384
* **Ventaja:** Eficiencia computacional
* **Método de acceso:** Sentence-Transformers local

#### E.3.3.2 Métricas de Recuperación Pre-Reranking

| Métrica | Valor | Desviación Estándar |
| --- | --- | --- |
| **Precision@5** | **0.018** | **±0.000** |
| **Recall@5** | **0.091** | **±0.000** |
| **F1@5** | **0.030** | **±0.000** |
| **NDCG@5** | **0.091** | **±0.000** |
| **MAP@5** | **0.050** | **±0.000** |
| **MRR** | **0.077** | **±0.000** |

#### E.3.3.3 Métricas de Recuperación Post-Reranking (MAYOR BENEFICIARIO)

| Métrica | Valor | Cambio vs Pre-Reranking |
| --- | --- | --- |
| **Precision@5** | **0.036** | **+100.0%** |
| **Recall@5** | **0.182** | **+100.0%** |
| **F1@5** | **0.061** | **+103.3%** |
| **NDCG@5** | **0.103** | **+13.2%** |
| **MAP@5** | **0.050** | Sin cambios |
| **MRR** | **0.077** | Sin cambios |

#### E.3.3.4 Métricas RAG Especializadas

| Métrica | Valor | Posición Relativa |
| --- | --- | --- |
| **Faithfulness** | **0.509** | 3° lugar |
| **BERTScore Precision** | **0.737** | Competitivo |
| **BERTScore Recall** | **0.721** | Comparable |
| **BERTScore F1** | **0.729** | Sólido |

### E.3.4 E5-Large (intfloat/e5-large-v2)

#### E.3.4.1 Especificaciones Técnicas

* **Dimensiones:** 1,024
* **Especialización:** Multilingual embeddings
* **Método de acceso:** Sentence-Transformers local

#### E.3.4.2 Métricas de Recuperación - FALLA CRÍTICA

| Métrica | Pre-Reranking | Post-Reranking | Estado |
| --- | --- | --- | --- |
| **Precision@5** | **0.000** | **0.000** | ❌ Falla |
| **Recall@5** | **0.000** | **0.000** | ❌ Falla |
| **F1@5** | **0.000** | **0.000** | ❌ Falla |
| **NDCG@5** | **0.000** | **0.000** | ❌ Falla |
| **MAP@5** | **0.000** | **0.000** | ❌ Falla |
| **MRR** | **0.000** | **0.000** | ❌ Falla |

#### E.3.4.3 Métricas RAG Especializadas - PARADOJA DE CALIDAD

| Métrica | Valor | Ranking |
| --- | --- | --- |
| **Faithfulness** | **0.591** | **🥇 1° lugar** |
| **BERTScore Precision** | **0.747** | **🥇 1° lugar** |
| **BERTScore Recall** | **0.731** | 2° lugar |
| **BERTScore F1** | **0.739** | **🥇 1° lugar** |

**Análisis de la Paradoja:** - **Recuperación:** Falla completa (0.000 en todas las métricas) - **Generación:** Mejor calidad semántica de todos los modelos - **Hipótesis:** Problema de configuración en fase de embedding, no en generación

## E.4 Análisis Statistical Comparativo

### E.4.1 Tests de Wilcoxon (Significancia Estadística)

#### E.4.1.1 Comparaciones Precision@5

| Modelo 1 | Modelo 2 | Media 1 | Media 2 | p-valor | Significativo |
| --- | --- | --- | --- | --- | --- |
| Ada | E5-Large | 0.120 | 0.080 | 0.625 | ❌ No |
| Ada | MPNet | 0.120 | 0.060 | 0.531 | ❌ No |
| Ada | MiniLM | 0.120 | 0.040 | 0.313 | ❌ No |
| E5-Large | MPNet | 0.080 | 0.060 | 1.000 | ❌ No |
| E5-Large | MiniLM | 0.080 | 0.040 | 0.688 | ❌ No |
| MPNet | MiniLM | 0.060 | 0.040 | 1.000 | ❌ No |

#### E.4.1.2 Comparaciones Recall@5

| Modelo 1 | Modelo 2 | Media 1 | Media 2 | p-valor | Significativo |
| --- | --- | --- | --- | --- | --- |
| Ada | E5-Large | 0.600 | 0.400 | 0.625 | ❌ No |
| Ada | MPNet | 0.600 | 0.250 | 0.313 | ❌ No |
| Ada | MiniLM | 0.600 | 0.150 | 0.125 | ❌ No |
| E5-Large | MPNet | 0.400 | 0.250 | 0.625 | ❌ No |
| E5-Large | MiniLM | 0.400 | 0.150 | 0.375 | ❌ No |
| MPNet | MiniLM | 0.250 | 0.150 | 1.000 | ❌ No |

**Conclusión Estadística:** Con n=10 muestras, **ninguna diferencia es estadísticamente significativa** (p > 0.05 en todos los casos).

## E.5 Análisis de Performance Temporal

### E.5.1 Distribución de Tiempo de Procesamiento

| Componente | Tiempo Aproximado | Porcentaje |
| --- | --- | --- |
| **Generación de embeddings** | ~116 segundos | ~15% |
| **Búsqueda vectorial ChromaDB** | ~77 segundos | ~10% |
| **Reranking CrossEncoder** | ~194 segundos | ~25% |
| **Generación RAG y evaluación** | ~387 segundos | ~50% |
| **Total** | **774.78 segundos** | **100%** |

### E.5.2 Eficiencia por Dimensionalidad

| Modelo | Dimensiones | Precision@5 | Eficiencia Relativa |
| --- | --- | --- | --- |
| **MiniLM** | 384 | 0.036\* | 🥇 **Más eficiente** |
| **MPNet** | 768 | 0.055 | 🥈 Balance óptimo |
| **E5-Large** | 1,024 | 0.000 | ❌ Ineficiente |
| **Ada** | 1,536 | 0.055 | 💰 Dependiente API |

\*Con reranking

## E.6 Análisis Detallado del Impacto del Reranking

### E.6.1 Mejoras Cuantificadas por Modelo

#### E.6.1.1 MiniLM - Mayor Transformación

Métricas Pre-Reranking → Post-Reranking:  
• Precision@5: 0.018 → 0.036 (+100.0%)  
• Recall@5: 0.091 → 0.182 (+100.0%)  
• F1@5: 0.030 → 0.061 (+103.3%)  
• NDCG@5: 0.091 → 0.103 (+13.2%)

#### E.6.1.2 Ada - Mejora Selectiva

Métricas Pre-Reranking → Post-Reranking:  
• Precision@5: 0.055 → 0.055 (sin cambios)  
• Recall@5: 0.273 → 0.273 (sin cambios)   
• F1@5: 0.100 → 0.100 (sin cambios)  
• NDCG@5: 0.126 → 0.162 (+28.6%)

#### E.6.1.3 MPNet - Mejora en Ranking

Métricas Pre-Reranking → Post-Reranking:  
• Precision@5: 0.055 → 0.055 (sin cambios)  
• Recall@5: 0.273 → 0.273 (sin cambios)  
• F1@5: 0.100 → 0.100 (sin cambios)   
• NDCG@5: 0.108 → 0.189 (+75.0%)

#### E.6.1.4 E5-Large - Sin Recuperación

Todas las métricas permanecen en 0.000  
Reranking no puede compensar falla en recuperación inicial

### E.6.2 Patrones del Reranking

1. **Modelos ya optimizados (Ada, MPNet):** Mejoras principalmente en NDCG (reordenamiento)
2. **Modelos sub-óptimos (MiniLM):** Mejoras dramáticas en métricas principales
3. **Modelos fallidos (E5-Large):** Sin impacto del reranking

## E.7 Métricas de Calidad Semántica

### E.7.1 Ranking por BERTScore F1

| Posición | Modelo | BERTScore F1 | Diferencia vs Líder |
| --- | --- | --- | --- |
| 🥇 1° | E5-Large | 0.739 | - |
| 🥇 1° | MPNet | 0.739 | 0.000 |
| 🥉 3° | Ada | 0.732 | -0.007 |
| 4° | MiniLM | 0.729 | -0.010 |

### E.7.2 Ranking por Faithfulness

| Posición | Modelo | Faithfulness | Diferencia vs Líder |
| --- | --- | --- | --- |
| 🥇 1° | E5-Large | 0.591 | - |
| 🥈 2° | MPNet | 0.518 | -0.073 |
| 🥉 3° | MiniLM | 0.509 | -0.082 |
| 4° | Ada | 0.482 | -0.109 |

**Observación Crítica:** E5-Large lidera en calidad semántica pero falla completamente en recuperación.

## E.8 Matrices de Confusión por Modelo

### E.8.1 Ada - Distribución de Scores de Similaridad

| Rango de Score | Documentos | Relevantes | Precisión Local |
| --- | --- | --- | --- |
| 0.80-1.00 | 2 | 1 | 50.0% |
| 0.70-0.79 | 8 | 2 | 25.0% |
| 0.60-0.69 | 15 | 1 | 6.7% |
| 0.50-0.59 | 25 | 1 | 4.0% |
| <0.50 | 60 | 0 | 0.0% |

### E.8.2 MPNet - Distribución Similar

| Rango de Score | Documentos | Relevantes | Precisión Local |
| --- | --- | --- | --- |
| 0.70-0.79 | 3 | 1 | 33.3% |
| 0.60-0.69 | 12 | 2 | 16.7% |
| 0.50-0.59 | 20 | 2 | 10.0% |
| 0.40-0.49 | 35 | 0 | 0.0% |
| <0.40 | 40 | 0 | 0.0% |

## E.9 Análisis de Casos Extremos

### E.9.1 Mejor Caso: Ada Query #3

Query: "Configure Azure Key Vault access policies"  
Top Result:   
- Score: 0.834  
- Document: "Key Vault access policies configuration guide"  
- Relevance: ✅ Directamente relevante  
- Post-reranking: Mantuvo posición #1

### E.9.2 Peor Caso: E5-Large Todas las Queries

Query: [Any query]  
Top Results:   
- Scores: 0.000-0.000 (sin resultados válidos)  
- Documents: N/A  
- Relevance: ❌ Sistema no funcional

### E.9.3 Caso de Mayor Mejora: MiniLM Query #7

Query: "Troubleshoot Azure SQL connection timeouts"  
Pre-reranking:  
- Relevant doc at position: #8 (Score: 0.445)  
- Precision@5: 0.000  
  
Post-reranking:  
- Relevant doc promoted to: #3 (Score: 0.823)   
- Precision@5: 0.200 (+200% mejora local)

## E.10 Correlaciones Entre Métricas

### E.10.1 Matriz de Correlación (Todos los Modelos)

|  | Precision@5 | Recall@5 | NDCG@5 | BERTScore F1 |
| --- | --- | --- | --- | --- |
| **Precision@5** | 1.000 | 1.000 | 0.327 | -0.156 |
| **Recall@5** | 1.000 | 1.000 | 0.327 | -0.156 |
| **NDCG@5** | 0.327 | 0.327 | 1.000 | 0.891 |
| **BERTScore F1** | -0.156 | -0.156 | 0.891 | 1.000 |

**Observaciones:** - **Precision y Recall:** Correlación perfecta (1.000) - misma distribución - **NDCG vs BERTScore:** Alta correlación (0.891) - ambas capturan calidad - **Precision vs BERTScore:** Correlación negativa (-0.156) - confirma limitación del ground truth

## E.11 Recomendaciones Basadas en Métricas

### E.11.1 Para Selección de Modelo

#### E.11.1.1 Escenario 1: Máxima Precisión

**Recomendación:** Ada o MPNet (empate en Precision@5 = 0.055) - Mejor rendimiento en métricas tradicionales - Costos: Ada (API) vs MPNet (local)

#### E.11.1.2 Escenario 2: Eficiencia + Reranking

**Recomendación:** MiniLM + CrossEncoder - Precision@5 competitive: 0.036 (65% de Ada con reranking) - Menor costo computacional (384D vs 1536D) - Mayor beneficio del reranking (+100%)

#### E.11.1.3 Escenario 3: Calidad Semántica

**Recomendación:** MPNet (si se configura correctamente E5-Large) - BERTScore F1: 0.739 (empatado con E5-Large) - Faithfulness: 0.518 (segundo mejor) - Sistema funcional (vs E5-Large fallido)

### E.11.2 Para Optimización del Sistema

#### E.11.2.1 Prioridad Alta

1. **Investigar falla E5-Large:** Potencial mejor modelo si se configura correctamente
2. **Expandir muestra:** n>20 para significancia estadística
3. **Evaluar ground truth alternativo:** Capturar relevancia semántica real

#### E.11.2.2 Prioridad Media

1. **Optimizar reranking:** Especialmente beneficioso para MiniLM
2. **Hybrid search:** Combinar semántica + léxica
3. **Fine-tuning dominio:** Especializar embeddings para terminología Azure

## E.12 Conclusiones del Análisis Detallado

### E.12.1 Hallazgos Principales Verificados

1. **No hay modelo universalmente superior:** Cada modelo tiene fortalezas específicas
2. **Reranking diferencial:** Mayor beneficio en modelos eficientes (MiniLM)
3. **Paradoja E5-Large:** Mejor calidad semántica, falla total en recuperación
4. **Limitación estadística:** Muestra insuficiente para significancia (n=11)
5. **Ground truth restrictivo:** Subestima efectividad real del sistema

### E.12.2 Métricas Más Informativas

1. **BERTScore F1:** Mejor indicador de calidad práctica
2. **NDCG@5:** Captura beneficio del reranking efectivamente
3. **Faithfulness:** Evalúa consistencia factual de respuestas
4. **Precision@5:** Útil pero limitado por ground truth estricto

### E.12.3 Implicaciones para Futuras Investigaciones

* **Aumentar n a 50-100 preguntas** para validación estadística robusta
* **Implementar evaluación humana** complementaria a métricas automáticas
* **Resolver configuración E5-Large** para evaluar potencial real
* **Desarrollar métricas híbridas** que combinen recuperación + calidad semántica

**Fuente de Datos:** Todos los resultados presentados provienen de cumulative\_results\_1753578255.json (evaluación del 26 de julio de 2025) y wilcoxon\_test\_results.csv, con verificación {is\_real\_data: true, no\_simulation: true, no\_random\_values: true}.