# 7. RESULTADOS Y ANÁLISIS (ACTUALIZADO)

## 7.2 Configuración Experimental

La evaluación experimental se ejecutó el 2 de agosto de 2025, procesando 1,000 preguntas de prueba distribuidas entre 4 modelos de embedding diferentes. La configuración experimental verificada incluye:

**Parámetros de Evaluación:**

• Preguntas evaluadas: 1,000 por modelo (91x más que la evaluación inicial)

• Modelos comparados: 4 (Ada, MPNet, MiniLM, E5-Large)

• Método de reranking: CrossEncoder (ms-marco-MiniLM-L-6-v2) con normalización Min-Max

• Top-k: 15 documentos por consulta

• Duración total: 28,216 segundos (7.8 horas)

• Framework de evaluación: RAGAS completo con API de OpenAI

**Corpus de Evaluación:**

• Documentos indexados: 187,031 chunks técnicos

• Dimensiones por modelo: Ada (1,536D), E5-Large (1,024D), MPNet (768D), MiniLM (384D)

• Ground truth: 2,067 pares pregunta-documento validados

### 7.2.2 Ada (OpenAI text-embedding-ada-002)

#### 7.2.2.1 Métricas de Recuperación

El modelo Ada mantuvo el mejor rendimiento general entre todos los modelos evaluados en la fase de recuperación inicial:

**Rendimiento ANTES del CrossEncoder Reranking:**

• Precision@5: 0.097 (±0.021)

• Recall@5: 0.399 (±0.078)

• F1@5: 0.152 (±0.029)

• NDCG@5: 0.228 (±0.045)

• MAP@5: 0.263 (±0.052)

• MRR: 0.217 (±0.043)

**Rendimiento DESPUÉS del CrossEncoder Reranking:**

• Precision@5: 0.079 (-18.3%)

• Recall@5: 0.324 (-18.7%)

• F1@5: 0.124 (-18.4%)

• NDCG@5: 0.206 (-9.6%)

• MAP@5: 0.199 (-24.3%)

• MRR: 0.197 (-9.5%)

Con el dataset ampliado, se observa un patrón inesperado: el reranking tuvo un impacto negativo en las métricas de Ada, sugiriendo que las representaciones iniciales del modelo ya capturan óptimamente la relevancia semántica.

### 7.2.3 MPNet (multi-qa-mpnet-base-dot-v1)

#### 7.2.3.1 Métricas de Recuperación

MPNet demostró rendimiento sólido, confirmando su especialización en tareas de pregunta-respuesta:

**Rendimiento ANTES del CrossEncoder Reranking:**

• Precision@5: 0.074 (±0.018)

• Recall@5: 0.292 (±0.065)

• F1@5: 0.112 (±0.024)

• NDCG@5: 0.199 (±0.041)

• MAP@5: 0.218 (±0.046)

• MRR: 0.185 (±0.038)

**Rendimiento DESPUÉS del CrossEncoder Reranking:**

• Precision@5: 0.070 (-5.6%)

• Recall@5: 0.280 (-4.2%)

• F1@5: 0.108 (-3.6%)

• NDCG@5: 0.196 (-1.5%)

• MAP@5: 0.210 (-3.7%)

• MRR: 0.185 (sin cambios)

Similar a Ada, MPNet mostró un impacto mínimo o negativo del reranking, validando la calidad de sus embeddings especializados para Q&A.

### 7.2.4 MiniLM (all-MiniLM-L6-v2)

#### 7.2.4.1 Métricas de Recuperación

MiniLM continúa siendo el modelo que más se beneficia del reranking:

**Rendimiento ANTES del CrossEncoder Reranking:**

• Precision@5: 0.053 (±0.013)

• Recall@5: 0.201 (±0.048)

• F1@5: 0.081 (±0.018)

• NDCG@5: 0.148 (±0.032)

• MAP@5: 0.160 (±0.035)

• MRR: 0.144 (±0.031)

**Rendimiento DESPUÉS del CrossEncoder Reranking:**

• Precision@5: 0.059 (+11.8%)

• Recall@5: 0.226 (+12.3%)

• F1@5: 0.091 (+12.3%)

• NDCG@5: 0.162 (+9.3%)

• MAP@5: 0.170 (+6.3%)

• MRR: 0.156 (+8.4%)

Los resultados confirman que MiniLM, a pesar de su menor dimensionalidad, alcanza rendimiento competitivo cuando se combina con reranking neural.

### 7.2.5 E5-Large (intfloat/e5-large-v2)

#### 7.2.5.1 Métricas de Recuperación - Resolución del Problema

E5-Large ahora muestra métricas válidas, resolviendo la falla crítica observada en la evaluación inicial:

**Rendimiento ANTES del CrossEncoder Reranking:**

• Precision@5: 0.060 (±0.015)

• Recall@5: 0.239 (±0.055)

• F1@5: 0.092 (±0.020)

• NDCG@5: 0.169 (±0.036)

• MAP@5: 0.183 (±0.040)

• MRR: 0.161 (±0.034)

**Rendimiento DESPUÉS del CrossEncoder Reranking:**

• Precision@5: 0.065 (+7.6%)

• Recall@5: 0.256 (+7.1%)

• F1@5: 0.100 (+8.7%)

• NDCG@5: 0.166 (-1.5%)

• MAP@5: 0.186 (+1.6%)

• MRR: 0.156 (-3.3%)

E5-Large muestra mejoras mixtas con el reranking, con ganancias en precisión y recall pero ligeras pérdidas en métricas de ranking.

## 7.3 Análisis Comparativo (Actualizado)

### 7.3.1 Métricas de Precisión

#### 7.3.1.1 Ranking de Modelos por Precision@5 (1000 preguntas)

**Ranking ANTES del Reranking:**

1. **Ada:** 0.097 (liderazgo claro)

2. **MPNet:** 0.074 (-23.7% vs Ada)

3. **E5-Large:** 0.060 (-38.1% vs Ada)

4. **MiniLM:** 0.053 (-45.4% vs Ada)

**Ranking DESPUÉS del Reranking:**

1. **Ada:** 0.079 (mantiene liderazgo pero reducido)

2. **MPNet:** 0.070 (-11.4% vs Ada)

3. **E5-Large:** 0.065 (-17.7% vs Ada)

4. **MiniLM:** 0.059 (-25.3% vs Ada)

### 7.3.2 Significancia Estadística

Con 1000 preguntas por modelo, las diferencias observadas son estadísticamente más robustas:

• Ada vs MiniLM: Diferencia significativa (p < 0.001)

• Ada vs E5-Large: Diferencia significativa (p < 0.001)

• Ada vs MPNet: Diferencia significativa (p < 0.05)

• MPNet vs MiniLM: Diferencia significativa (p < 0.01)

### 7.3.3 Impacto del CrossEncoder (Actualizado)

#### 7.3.3.1 Análisis por Modelo

**Patrón Emergente con Dataset Ampliado:**

1. **Modelos de alta calidad (Ada, MPNet):** Impacto negativo del reranking

2. **Modelos eficientes (MiniLM, E5-Large):** Beneficio positivo del reranking

Este patrón sugiere que el CrossEncoder es más efectivo cuando compensa deficiencias en embeddings iniciales, pero puede introducir ruido cuando los embeddings ya son de alta calidad.

## 7.6 Discusión de Resultados (Actualización)

### 7.6.1 Implicaciones de la Evaluación Ampliada

1. **Confiabilidad Estadística:** Con 91x más datos, las métricas son significativamente más confiables y representativas del rendimiento real.

2. **Jerarquía Clara de Modelos:** Ada > MPNet > E5-Large > MiniLM se confirma consistentemente a través de múltiples métricas.

3. **Reranking Contextual:** El impacto del CrossEncoder no es universalmente positivo; depende de la calidad inicial de los embeddings.

4. **E5-Large Funcional:** La resolución del problema confirma la importancia de la configuración apropiada más que deficiencias inherentes del modelo.

### 7.6.2 Recomendaciones Actualizadas

1. **Para máxima precisión:** Ada sin reranking

2. **Para balance costo-efectividad:** MPNet sin reranking

3. **Para restricciones de recursos:** MiniLM con reranking obligatorio

4. **Para investigación:** E5-Large requiere evaluación adicional post-configuración