💰 Análisis de Costos - APIs para Jurisprudencia Jujuy

📊 Resumen Ejecutivo

Costo estimado para procesar TODO el repositorio inicial de Jujuy:

- **Una sola vez (setup inicial):** USD $150 - $300
- **Costo mensual de mantenimiento:** USD $20 - $50/mes

🔍 Estimación del Volumen de Datos

Supuestos Base

```
- Fallos totales estimados en Jujuy: ~5,000 - 10,000 fallos históricos
- Promedio de palabras por fallo: 2,000 - 5,000 palabras
- Nuevos fallos por mes: ~100 - 200 fallos
```

Para los cálculos usaremos:

- **Volumen inicial:** 7,500 fallos (punto medio)
- **Palabras por fallo:** 3,000 palabras promedio
- **Nuevos fallos/mes:** 150 fallos

💵 Costos de Claude API (Análisis y Etiquetado)

Pricing Claude Sonnet 4.5

```
Modelo: claude-sonnet-4-5-20250929
- Input: $3.00 por millón de tokens
- Output: $15.00 por millón de tokens
```

Cálculo por Fallo Individual

Tokens por fallo:

- Input (texto del fallo): ~4,000 tokens (3,000 palabras × 1.33)
- Input (prompt del sistema): ~800 tokens
- **Total Input:** ~4,800 tokens por fallo
- Output (análisis JSON): ~500 tokens
- **Total Output:** ~500 tokens por fallo

Costo por fallo:

```
Input:  4,800 tokens × $3.00 / 1,000,000 = $0.0144
Output:   500 tokens × $15.00 / 1,000,000 = $0.0075
──────────────────────────────────────────────────
TOTAL POR FALLO: $0.0219 (~2.2 centavos)
```

Costos Totales - Claude API

Concepto	Cantidad	Costo Unitario	Total
Setup Inicial	7,500 fallos	$0.0219	**$164.25**
Mantenimiento Mensual	150 fallos	$0.0219	**$3.29/mes**
Año 1 completo	Setup + 12 meses	-	**$203.73**

🔢 Costos de OpenAI Embeddings

Pricing OpenAI Embeddings

```
Modelo: text-embedding-3-small (recomendado)
- Costo: $0.020 por millón de tokens
- Dimensiones: 1536 (ideal para búsqueda semántica)
```

Alternativa más económica:

```
Modelo: text-embedding-3-large
- Costo: $0.130 por millón de tokens
- Dimensiones: 3072 (mayor precisión)
```

Cálculo de Embeddings

Por cada fallo generamos 2 embeddings:

1. Embedding del texto completo del fallo
2. Embedding del resumen generado por Claude

Tokens por embedding:

- Texto completo: ~4,000 tokens
- Resumen: ~200 tokens
- **Total:** 4,200 tokens por fallo

Costo por fallo (text-embedding-3-small):

4,200 tokens × $0.020 / 1,000,000 = $0.000084 (~0.008 centavos)

Costos Totales - OpenAI Embeddings

Concepto	Cantidad	Costo Unitario	Total
Setup Inicial	7,500 fallos	$0.000084	**$0.63**
Mantenimiento Mensual	150 fallos	$0.000084	**$0.01/mes**
Año 1 completo	Setup + 12 meses	-	**$0.75**

📈 Tabla de Costos Consolidada

Año 1 - Costo Total por Componente

Componente	Setup Inicial	Mensual	Año 1 Total
Claude API (análisis)	$164.25	$3.29	$203.
OpenAI Embeddings	$0.63	$0.01	$0.75
Buffer 20% (errores, re-procesamiento)	$32.98	$0.66	$40.9
———			
TOTAL	**$197.86**	**$3.96**	**$245.**

💡 Optimizaciones para Reducir Costos

1. Procesamiento Selectivo

Ahorro potencial: 30-40%

```python
# No procesar TODOS los fallos de golpe
# Priorizar por:
# - Fallos más recientes (últimos 2 años)
# - Fallos más consultados
# - Materias más demandadas (Laboral, Civil, Penal)

# Ejemplo:
Procesar inicialmente: 3,000 fallos → Costo: $65.70
Resto bajo demanda: 4,500 fallos → $98.55 (cuando sea necesario)
```

2. Cacheo Inteligente

Ahorro potencial: 15-20%

```python
# Cache de análisis similares
# Si un fallo es muy similar a otro ya procesado,
# reutilizar el análisis y solo ajustar detalles específicos

# Implementación:
if similitud_con_fallo_existente > 0.95:
    clonar_analisis_y_ajustar()  # Costo: $0.003 vs $0.0219
```

3. Modelos Alternativos

Ahorro potencial: 60-70%

```
Opción A: Claude Haiku 4.5
- Input: $0.40 / millón tokens (vs $3.00 Sonnet)
- Output: $2.00 / millón tokens (vs $15.00 Sonnet)
- Costo por fallo: $0.0026 (~87% más barato)
- Trade-off: Menor calidad de análisis

Opción B: Mix de modelos
- Haiku para etiquetado simple → 70% de fallos
- Sonnet para casos complejos → 30% de fallos
- Ahorro promedio: 60%
```

4. Batch Processing

Ahorro potencial: 50%

Claude API ofrece descuentos para procesamiento batch (no urgente):

Claude API Batch:
- Input: $1.50 / millón tokens (50% descuento)
- Output: $7.50 / millón tokens (50% descuento)
- Procesamiento en 24 horas
- Ideal para setup inicial

Nuevo costo con batch:

Setup inicial: $164.25 → $82.13 (¡50% menos!)

🎯 Costos por Escenario de Uso

Escenario 1: MVP Mínimo (Solo casos recientes)

- Fallos a procesar: 1,500 (últimos 2 años)
- Setup inicial: $32.85 (Claude) + $0.13 (OpenAI) = $32.98
- Mensual: $3.30
- Ideal para: Validar el concepto

Escenario 2: Cobertura Media (Casos relevantes)

- Fallos a procesar: 4,000 (casos importantes + recientes)
- Setup inicial: $87.60 (Claude) + $0.34 (OpenAI) = $87.94
- Mensual: $3.30
- Ideal para: Lanzamiento público

Escenario 3: Cobertura Completa (Todo el repositorio)

- Fallos a procesar: 7,500 (histórico completo)
- Setup inicial: $164.25 (Claude) + $0.63 (OpenAI) = $164.88
- Mensual: $3.30
- Ideal para: Producto maduro

Escenario 4: Ultra Optimizado (Mix de modelos)

- 70% con Haiku, 30% con Sonnet
- Setup inicial: $52.00
- Mensual: $1.50
- Ideal para: Presupuesto ajustado

📊 Comparación con Alternativas

Opción A: Contratar Equipo Manual

- 2 pasantes de derecho: $800/mes c/u = $1,600/mes
- Tiempo para etiquetar 7,500 fallos: 6-8 meses
- Costo total: $9,600 - $12,800
- Consistencia: Variable

Opción B: Solución con IA (Nuestra propuesta)

- Costo total primer año: $245.38
- Tiempo de procesamiento: 2-3 días
- Costo total: $245.38
- Consistencia: 100%
- Ahorro: 97.5% vs equipo manual

💳 Costos de Infraestructura Adicional

Base de Datos (PostgreSQL)

Opción 1: Railway.app
- Plan Hobby: $5/mes (500MB)
- Plan Pro: $20/mes (8GB) ← Recomendado

Hosting Backend

Opción 1: Render.com
- Plan Free: $0/mes (limitado)
- Plan Starter: $7/mes ← Recomendado para MVP

Hosting Frontend

Opción 1: Vercel
- Plan Free: $0/mes (suficiente para MVP)
- Plan Pro: $20/mes (para producción)

Monitoreo y Logs

Opción 1: Sentry (errores)
- Plan Free: $0/mes (5k events)

Opción 2: PostHog (analytics)
- Plan Free: $0/mes (1M eventos)

Total infraestructura mensual: $20-50/mes

📅 Proyección de Costos a 3 Años

Año	Setup	API Mensual	Infra Mensual
Año 1	$197.86	$3.96	$27
Año 2	$0	$4.75	$35
Año 3	$0	$5.70	$35
———			
Total 3 años			

Promedio mensual 3 años: $41.33/mes

🎁 Modelos Gratuitos como Alternativa

Si el presupuesto es MUY limitado:

Opción 1: Llama 3.1 (405B) via Groq

- Costo: GRATIS (con límites)
- Rate limit: 30 req/min
- Calidad: 80-85% vs Claude
- Tiempo de procesamiento: 2-3x más lento

Opción 2: Mistral Large via HuggingFace

- Costo: GRATIS o $0.50/millón tokens
- Calidad: 75-80% vs Claude
- Setup más complejo

Opción 3: Embeddings Open Source

- Modelo: sentence-transformers/all-MiniLM-L6-v2
- Costo: GRATIS (ejecutar localmente)
- Calidad: 70-75% vs OpenAI
- Requiere GPU ($10-20/mes en Lambda Labs)

🏆 Recomendación Final

Para MVP (Primeros 3 meses)

Configuración Recomendada:

```yaml
Setup Inicial:
  - Procesamiento: Batch API de Claude (50% descuento)
  - Fallos iniciales: 3,000 (casos recientes + importantes)
  - Modelo: Claude Sonnet 4.5
  - Embeddings: OpenAI text-embedding-3-small
  - Costo: $49.28 (una sola vez)

Operación Mensual:
  - Fallos nuevos: 150/mes
  - API costs: $3.96/mes
  - Infraestructura: $27/mes
  - Total mensual: $30.96/mes

Primer trimestre completo: $142.16
```

Si el presupuesto es limitado

Configuración Ultra-Budget:

```yaml
Setup Inicial:
  - Modelo: Claude Haiku 4.5 (batch)
  - Fallos: 1,500 (solo últimos 18 meses)
  - Embeddings: Open source (sentence-transformers)
  - Costo: $7.80 (una sola vez)

Operación Mensual:
  - API: $0.39/mes
  - Infraestructura: $12/mes (Railway hobby + Render free)
  - Total mensual: $12.39/mes

Primer trimestre: $45.00
```

📞 Contacto para Descuentos

Anthropic y OpenAI ofrecen créditos para proyectos educativos/legales:

- Anthropic: claude-partnerships@anthropic.com
- OpenAI: https://openai.com/api/pricing (créditos para startups)

Potencialmente puedes obtener:

- $500-1,000 en créditos gratuitos
- Descuentos de 20-50% para proyectos de bien público
- Soporte técnico prioritario

🎯 Conclusión

ROI del Proyecto

Inversión Año 1: ~$522 **Valor generado:**

- Ahorras 97% vs equipo manual ($12,000)
- Producto escalable a otras provincias
- Potencial de ingresos: $300-600/mes (con solo 30-60 suscriptores Pro)

Break-even: Mes 2-3 con modelo de suscripción

El costo de las APIs NO es un bloqueante para este proyecto. ✅

📚 Recursos para Monitoreo de Costos

```bash
# Dashboard de costos en tiempo real
- Claude: https://console.anthropic.com/settings/billing
- OpenAI: https://platform.openai.com/usage

# Alerts para evitar sorpresas
- Configurar alertas en $50, $100, $200
- Límites de tasa (rate limits)
- Monitoreo diario con scripts
```

Última actualización: Enero 2026 **Precios sujetos a cambios:** Verificar en sitios oficiales