RECÁLCULO EJEMPLO 3

Análisis Comparativo de Costos e Infraestructura

Noviembre 2025 vs Mayo 2026 | Modelos: Claude · Gemini · Copilot

Análisis realizado por: Claude (Anthropic / Sonnet 4)	**Fecha del recálculo:** Mayo 2026
Referencia original: *alfonsoorozcoaguilar.com/2025/11/21*	***Referencia hardware:*** *alfonsoorozcoaguilar.com/2025/11/22*

1. Contexto del Ejemplo 3

En noviembre de 2025, se evaluó la viabilidad de automatizar con Inteligencia Artificial un proceso de revisión documental en una paraestatal mexicana. El proceso involucraba verificar expedientes ciudadanos de 32 estados, con formatos diferentes, para emitir una decisión binaria (Sí/No) que determinaba si el trámite continuaba.

Parámetro	Descripción
Documentos por expediente	7 documentos (el primero con 50 datos / 3 columnas)
Formatos distintos posibles	35 formatos diferentes por estado
Variables de análisis	60 variables; resultado: 45 posibles por solicitante
Decisión requerida	Binaria: Sí / No (paso 2 del proceso)
Volumen estimado	900 – 1,200 expedientes / mes
Personal actual	3 analistas humanos (~25 funciones cada uno)
Función automatizable	Solo 1 de las ~25 funciones por analista
Costo actual (3 analistas)	~700,000 MXN / año (~32,400 USD)

El error lógico central de las propuestas originales: Gemini y Copilot proponían capturar 150 campos por expediente (que las analistas no capturaban antes) para simplificar el paso 20 del proceso. Esto le añadía trabajo al recurso humano mientras el costo del servidor y el desarrollo del paso 20 «mágicamente desaparecía». Copilot estimó el costo como «cercano a cero» por reutilizar infraestructura existente. Gemini dijo «cero» sin que se le preguntara. Claude en noviembre fue el único en documentar el ROI negativo.

2. Evolución de Precios: Noviembre 2025 → Mayo 2026

Los precios de tokens e infraestructura cloud han cambiado entre la evaluación original y la fecha de este recálculo. La siguiente tabla muestra las variaciones relevantes:

Componente	Nov 2025	May 2026	Variación	Impacto
Claude Opus (input/output) $/1M tokens	$15 / $75	$15 / $75	Sin cambio	Neutro
Claude Sonnet (input/output) $/1M tokens	$3 / $15	$3 / $15	Sin cambio	Neutro
GPT-4o (Azure OpenAI) $/1M tokens	$2.50 / $10	$2.50 / $10	Sin cambio	Neutro
GPT-4o-mini $/1M tokens	$0.15 / $0.60	$0.15 / $0.60	Sin cambio	Neutro
Azure VM D4s v3 (16 GB RAM, 4 vCPU)	~$140/mes	~$145/mes	+3.5%	Leve alza
Azure Document Intelligence OCR (por 1,000 páginas)	$10 USD	$10 USD	Sin cambio	Neutro
Azure Standard SSD (por GB/mes)	$0.10	$0.10	Sin cambio	Neutro
MXN / USD (tipo de cambio promedio referencial)	~$17.50	~$19.80	+13%	Alza costos MXN
GPU K80 (Azure NC6)	Disponible (obsoleta)	DESCONTINUADA por Azure	Eliminada	Copilot hoy la sigue sugiriendo

Nota: La depreciación del peso mexicano (+13% en tipo de cambio) es el cambio más significativo para el contexto de la paraestatal, ya que eleva el costo en MXN de cualquier servicio denominado en dólares, aunque los precios en USD sean estables.

3. Recálculo de Costos de Tokens (Mayo 2026)

Conclusión técnica clave: No se requiere fine-tuning. RAG (Retrieval-Augmented Generation) con prompt engineering es suficiente para manejar los 35 formatos. Esta fue la corrección más importante aportada por Gemini en mayo 2026, y corrige un error del análisis de Claude de noviembre 2025 que sí asumía fine-tuning.

Concepto	Nov 2025 (Claude Opus)	May 2026 (GPT-4o-mini / Sonnet)	Diferencia
Volumen mensual base	1,050 docs × 2,500 tok	1,050 docs × 3,000 tok	+20% tokens base
Multiplicadores aplicados	Re-proc 2x, RAG 2x, Validación 1.5x = 6x	Re-proc 2x, RAG 2x, Validación 1.5x = 6x	Igual
Tokens mensuales totales	11,812,500	~12,600,000	+6.7%
Costo tokens (modelo Opus/caro)	$354 USD/mes ($177 in + $177 out)	$354 USD/mes (sin cambio en precios)	Igual
Costo tokens (GPT-4o-mini)	N/A estimado	~$0.45 USD/mes (optimista, sin OCR)	Gemini hoy: correcto pero incompleto
OCR (Azure Doc. Intelligence) 7 docs × 1,050 personas	No calculado	$73.50 USD/mes (7 docs × $10/1000 págs)	Costo real omitido en nov 2025

Costo tokens + OCR (realista)	**~$354 USD/mes**	**~$427 USD/mes (tokens Opus + OCR)**	**+20.6% vs nov 2025**

Nota crítica: El cálculo de GPT-4o-mini ($0.45/mes) de Gemini es técnicamente correcto solo para tokens de texto puro, pero omite OCR, RAG, orquestación y los 6 documentos restantes del expediente. El costo real de tokens + OCR no baja de $400 USD/mes bajo supuestos realistas de volumen y re-procesamiento.

4. Recálculo de Infraestructura en Azure (Mayo 2026)

Se comparan tres escenarios de infraestructura: el estimado original de noviembre 2025, el escenario mínimo recalculado para mayo 2026, y la propuesta optimista de Gemini hoy (serverless + VM pequeña). Los tres superan el costo base que los modelos originales estimaron como «cero» o «cercano a cero».

Componente	Nov 2025 (Estimado original)	May 2026 Mínimo realista	May 2026 Gemini optimista	May 2026 Claude conservador
VM principal (tipo y RAM)	D4s v3 16 GB RAM $140/mes	D4s v3 16 GB RAM $145/mes	D2as v5 8 GB RAM $80-110/mes + Azure Functions	D4s v3 16 GB RAM $145/mes
Disco SSD premium	$30/mes (200 GB)	$30/mes (200 GB)	$10-20/mes (50 GB)	$30/mes (200 GB)
Base de datos RAG (vector DB)	$150-250/mes	$150-250/mes	$50/mes (micro-servicio)	$150-250/mes
API Gateway / orquestación	$50-100/mes	$50-100/mes	Azure Functions $0.01/ejecución (~$10-30/mes)	$50-100/mes
Monitoreo, backups, seguridad	$80-120/mes	$80-130/mes	$20-40/mes	$80-130/mes
OCR (Azure Doc. Intelligence)	No calculado	$73.50/mes (7 docs × 1,050)	$70/mes	$73.50/mes
SUBTOTAL INFRAESTRUCTURA	**$500-950/mes**	**$528-748/mes**	**$230-290/mes (escenario mínimo)**	**$528-748/mes**
Tokens de IA (ejecución)	$354/mes (Opus)	$427/mes (Opus + OCR)	$60-100/mes (modelo optimizado)	$427/mes
TOTAL OPERATIVO (solo un documento)	**$850-1,300/mes**	**$955-1,175/mes**	**$290-390/mes (optimista)**	**$955-1,175/mes**
TOTAL OPERATIVO ANUAL	**$10,200-15,600**	**$11,460-14,100**	**$3,480-4,680**	**$11,460-14,100**

Incluso en el escenario más optimista de Gemini ($290-390 USD/mes), el costo sigue siendo entre 10x y 13x mayor que los $30 USD/mes que originalmente se propusieron. El escenario optimista es también el menos probado en producción con estos volúmenes.

5. Análisis de ROI — ¿Es Negativo en Todos los Casos?

La pregunta central: ¿confirman todos los escenarios analizados que el ROI es negativo? La respuesta es **sí, en todos los casos**, incluyendo el escenario más optimista de Gemini. La razón estructural es la misma independientemente del costo de tokens o VM:

Razón estructural del ROI negativo (invariante al escenario):

La IA solo automatiza 1 de las ~25 funciones de cada analista. La analista ("la dama Margarita") sigue siendo indispensable para las otras 24. Por lo tanto, el costo de personal NO se reduce — se SUMA el costo de la IA. Esto hace matemáticamente imposible un ROI positivo bajo cualquier configuración de infraestructura mientras el personal no pueda eliminarse.

Escenario	Costo IA Año 1 (USD)	Costo personal manual (USD/año)	ROI (USD)	¿Positivo?
Noviembre 2025 (Claude original)	$104,200–163,600	$35,000–45,000	–$59,200 a –$118,600	NO ✗
Mayo 2026 Conservador (Claude)	$115,000–175,000 (tipo cambio +13%)	$32,400–45,000 (3 analistas)	–$70,000 a –$142,600	NO ✗
Mayo 2026 Optimista (Gemini)	$45,000–75,000 (desarrollo mínimo + ops optimistas)	$32,400–45,000	–$12,600 a –$42,600	NO ✗
Escenario teórico 'cero' (nov 2025 Gemini/Copilot)	$0 (propuesta original)	$32,400–45,000 (sin reducción real)	–$32,400 a –$45,000 (personal no se elimina)	NO ✗ (lógica inválida)

VEREDICTO: El ROI es NEGATIVO en el 100% de los escenarios analizados. Ni siquiera el escenario más optimista de Gemini (que asume desarrollo a costo mínimo y modelo pequeño) logra un ROI positivo, porque el ahorro en trabajo humano es cero mientras las analistas sigan siendo necesarias para sus otras 24 funciones.

6. Comparativa de Modelos: Noviembre 2025 vs Mayo 2026

Criterio	Claude Nov 25	Claude May 26	Copilot Nov 25	Copilot May 26	Gemini Nov 25	Gemini May 26
Costo cero = FALSO	✓	✓	✗	✓	✗	✓
Fine-tuning: NO necesario	✗*	✗*	✗	✗	✗	✓
OCR como costo real	✗	✗	✗	✗	✗	✓
Paradoja Margarita (+trabajo)	✓	✓	✗	~	✗	✓
ROI negativo explícito	✓	✓	✗	✗	✗	✗
VM/hardware realista	✓	✓	✗	~(K80 obsoleta)	✗	✓ serverless
Las otras 24 funciones	✓	✓	✗	✗	✗	✗
Pregunta sobre paso 20	✗	✗	✗	✗	✗	✓
GPU K80 descontinuada	N/A	N/A	N/A	ERROR	N/A	N/A

(*) Claude asumió fine-tuning en noviembre 2025 y en la consulta de mayo 2026 — error técnico corregido por Gemini. ~ = parcialmente correcto.

7. Conclusiones

1. El ROI es negativo en el 100% de los escenarios.

No existe ninguna configuración de infraestructura, modelo o proveedor que genere un ROI positivo mientras el personal humano sea indispensable para las otras 24 funciones. La reducción de costo de personal es cero porque Margarita sigue trabajando.

2. Los precios de tokens no cambiaron en 6 meses.

Tanto Claude Opus como GPT-4o mantienen sus precios desde noviembre 2025. El único cambio significativo en costos es el tipo de cambio MXN/USD (+13%), que encarece cualquier servicio cloud denominado en dólares para la paraestatal.

3. El costo de OCR era el más omitido.

Ningún modelo en noviembre 2025 calculó el costo de Azure Document Intelligence ($10/1,000 páginas). Con 7 documentos por expediente y 1,050 expedientes/mes, esto agrega ~$73.50 USD/mes al costo operativo real — que tampoco existía antes.

4. Gemini mejoró más. Copilot mejoró en forma, no en fondo.

Gemini pasó de 'cero' a la respuesta técnicamente más madura, identificando OCR, descartando fine-tuning y recomendando serverless. Copilot mejoró el tono pero sigue sin calcular ROI y recomienda hardware obsoleto (GPU K80 descontinuada).

5. El argumento del servidor compartido sigue siendo inválido.

Un servidor que ya usa su RAM para otras aplicaciones no puede absorber 12-16 GB adicionales de carga de IA sin degradar los servicios existentes o requerir upgrade. Incluso el escenario serverless de Gemini tiene costos reales no nulos.

6. El principio fundamental no cambió.

Si el costo de la solución excede el costo del error humano que pretende evitar, la solución es inviable. En noviembre 2025 esto era cierto. En mayo 2026, con el peso más depreciado y los precios de cloud estables, sigue siendo igual de cierto.

---

*Análisis realizado por **Claude** (Anthropic · Sonnet 4.6) | Mayo 2026*
Basado en documentos originales de noviembre 2025 publicados en alfonsoorozcoaguilar.com
Este documento es un recálculo comparativo, no una auditoría formal.