

XLIX CONGRESO ARGENTINO DE PROFESORES UNIVERSITARIOS DE COSTOS

Título de la ponencia:

Gestión de Costos ocultos de accesos indirectos y uso de tokens por Inteligencia Artificial basada en Arquitectura Transformer

Categoría propuesta: Aporte a la disciplina

Autores:

Constanza Ramos Nervi (Socia Activa)

Daniel Farré (Socio Activo)

Datos de contacto:

Constanza Ramos Nervi

constanzarn@gmail.com

Traful 1264 – Ing Maschwitz – PBA – Argentina – Tel: 5491139061746

Daniel Farré

Dfarre@paradigma.com

Rivadavia 5186 10mo – C.A.Buenos Aires – Argentina – Tel.: 541149471933

ÍNDICE

Resumen.....	3
Glosario.....	4
Introducción y objetivos de la ponencia.....	4
Problemática Actual.....	7
Contexto y Marco Teórico.....	8
Modelo de Tarifación del Servicio: Unidades de Medida y Costos de Referencia.....	9
Modelo de Costeo para Proyectos de IA: Tokens y Accesos Indirectos.....	10
Gestión de Costos: Aplicación del Ciclo <i>FinOps</i> para IA, adaptado: <i>FinOps+</i>	11
Casos Habituales de Eficientización y Simulación de Impacto.....	14
Tablero de Comando: Control y Desvíos.....	14
A modo de conclusión.....	16
Bibliografía.....	17

Resumen

El *boom* de la Inteligencia Artificial basada en Arquitectura *Transformer* (IAbAT) en el terreno personal despegó porque las tecnológicas regalaron el acceso a los *LLM* con topes generosos, instalando una ilusión de gratuidad similar a la de un buscador web. Con costo percibido cero y alto valor, la adopción explotó. Pero mientras los navegadores viven de la publicidad, el modelo económico de la IA aún busca su equilibrio.

En las empresas, este relato se complejiza. Los costos reales quedaron desdibujados por el espejismo del uso personal y los bonos de bienvenida de los proveedores. Creyendo que era un negocio redondo, las empresas corrieron a lanzar pruebas de concepto (*POC*) para dimensionar el impacto y la factura real.

Hoy esas *POC* dejan sabor a poco, más por expectativas infladas que por falta de impacto. Sin embargo, desde la trinchera de costos y gestión, detectamos un problema mayor: se desconocen los verdaderos costos de estas mejoras, muchos de ellos "ocultos" bajo la mirada de Savall y Zardet.

Esta ponencia abre el debate para poner sobre la mesa esos costos y aportar herramientas de control aplicables en el ex ante, durante el proceso y en el ex post.

Palabras clave: Inteligencia Artificial, Costos Ocultos, Costos de accesos indirectos a Data, Costos de uso de modelos de lenguaje, *Framework FinOps*

Glosario

- IAbAT: Inteligencia artificial basada en Arquitectura *Transformer* o IA de Modelos de lenguaje. Es aquella basada en el modelo de atención explicado en el artículo “*Attention is all you need*” (Vaswani, y otros, 2017): en lugar de buscar lematización, convierte las palabras de gran cantidad de frases en uno o varios “*tokens*” (números únicos que conforman una ficha) y luego vectorizan sus relaciones en una red neuronal compleja denominada *LLM*, por sus siglas en inglés.
- ✓ LLM: Modelo de Lenguaje de Gran Escala. Constituye el elemento central de la arquitectura *Transformer*
- ✓ POC: Prueba de Concepto. Pruebas operativas acotadas que se ejecutan en entornos controlados para comprobar si una tecnología cumple con las especificaciones funcionales esperadas (ó *MVP* Mínimo Producto Viable) antes de su despliegue general.
- ✓ FinOps: Disciplina de gestión financiera (*Framework*) sobre la administración integral de costos de Inteligencia Artificial (*AIOPs / FinOps for AI*). Consolida modelos híbridos de consumo que abarcan desde el alquiler de capacidad de cómputo especializada (*GPUs*) hasta la facturación dinámica por volumen de procesamiento en modelos de lenguaje de gran escala.
- ✓ Modelo de Costeo estocástico: Permite proyectar escenarios de costos probables mediante simulaciones cuantitativas incorporando la incertidumbre como variable.
- ✓ *Token*: Métrica base sobre la que se calcula el costo unitario variable en proveedores de IA.
- ✓ *Rate Limiting*: Regla de tope de consumo que evita excesos de facturación por bucles de consulta o sobreuso no planificado, asegurando previsibilidad en el presupuesto operativo.
- ✓ CIAT: Costo de IA por Transacción Integrada.
- ✓ ECC: Eficiencia de Caché de Contexto.
- ✓ TBR: Tasa de Llamadas en Lote por Acceso Indirecto.
- ✓ GCN: Gestión de Continuidad de Negocios (*BCM* por sus siglas en inglés)
- ✓ Hype: El término “*hype*” se utiliza en tecnología para representar gráficamente la hipérbole de uso que generan las nuevas tecnologías disruptivas como la inteligencia artificial a partir de fines de 2022.
- ✓ *BCM*: Iniciales en inglés para Gestión de Continuidad de negocio, enfoque de gestión que permite mantener el negocio productivo ante hechos de contingencia que lo pudieran afectar.

Introducción y objetivos de la ponencia

La Inteligencia Artificial basada en Arquitectura *Transformer* (IAbAT) ganó su popularidad y adopción de uso en el ámbito personal (y no empresarial) gracias a la estrategia comercial de las

empresas tecnológicas de brindar el servicio de modelos de lenguaje (*LLM*) en forma gratuita, con límites transaccionales de utilización libre muy altos para esa tipología de empleo. Ello generó una percepción (errónea) de gratuidad en el usuario, similar al del uso de los servicios de los navegadores de Internet para “*googlear*” por Internet. Es una obviedad relacionar explosión de adopción de experiencias de alto valor agregado y percepción de ausencia de costos. En el caso de los navegadores, la gratuidad para el usuario final se mantiene y la monetización se realiza por los ingresos por publicidad, demostrando un modelo ganar-ganar para todo el ecosistema. El equilibrio, en el caso de la IAbAT, recién se está “escribiendo”, sin llegar a su estabilización.

A nivel empresarial, dicho “relato” del equilibrio ganar-ganar es aún más complejo. La brecha percibida entre los beneficios proyectados de esta IA (y en particular de su efecto palanca sobre aumento de ingresos y reducción de costos actuales) y sus costos (subdimensionados por la percepción heredada de su uso personal y por las campañas de bonificación inicial de los Grandes Proveedores de esta tecnología en la carrera por ganar posicionamiento en un mercado altamente competitivo (por tratarse de una arquitectura que nació “abierta”) significó una obsesión de los empresarios en cuanto a los proyectos de “descubrimiento interno” (para no desaprovechar el *momentum* mágico de amplitud de brecha) para poner en blanco y negro, tanto el potencial de aumento de ingresos y reducción de costos operativos de las mejoras, cuanto el costo de adopción y uso de la nueva IA.

Los resultados de esas “pruebas de concepto” (*POC* por sus iniciales en inglés) a nivel mundial están decepcionando, en nuestro parecer, por exceso de expectativas y falla en el sentido de dirección de la adopción del cambio (el tradicional arriba-abajo en lugar de dar mayor participación al actor principal del cambio, que sigue siendo el ser humano y no la tecnología). Sin embargo, desde nuestra perspectiva de especialistas de Costos y Gestión, detectamos una inquietud aún mayor: El desconocimiento de los costos de estas mejoras algunos visibles, otros todavía “ocultos”, siguiendo la terminología de la TSE de Savall y Zardet (Savall & Zardet, 2008)

El objetivo de esta ponencia es desarrollar un marco metodológico para la Gestión de Costos por el uso de *tokens* y accesos indirectos de Inteligencia Artificial *Transformer* en el ámbito corporativo, alineado con el *framework FinOps*, con ajustes.

Para ello nos planteamos los siguientes objetivos intermedios:

- Abrir el debate sobre el asincronismo entre la etapa de alta inversión y adopción subsidiada de esta tecnología y su monetización y recupero a través de estrategias de precio ascendentes luego del *hype* y su impacto en los modelos de costeo de las empresas usuarias.
- Analizar la distorsión de precios del mercado de *APIs* de IA y fundamentar la necesidad de modelos de costeo de carácter estocástico.
- Identificar y clasificar los costos ocultos por accesos indirectos y transacciones de tokens bajo la óptica de la Teoría Socioeconómica (TSE).
- Diseñar un modelo operativo de control de gestión (FinOps+ para IA) estructurado en sus fases oficiales de *Inform* (Informar), *Optimize* (Optimizar) y *Operate* (Operar), agregando dos fases adicionales
- Proponer ejemplos de indicadores de gestión para el Tablero de Comando consistente con el modelo.
- Proponer modelos estocásticos de indiferencia de los casos de negocio con valores proyectados para todo el horizonte de planeamiento en lugar de los determinísticos que utilizan valores vigentes.

- Adherir al reclamo por la visibilización de las “cajas negras” de algoritmización de la labAT y solicitar la funcionalidad de edición por curaduría de expertos de gestión de negocios.

Problemática Actual.

La ilusión del costo cero, el riesgo del acceso indirecto y la modelización de los casos de negocio con precios no sostenibles en el mediano plazo

Tal como mencionamos *ut supra*, la masificación de la IAbAT en el ámbito personal se consolidó gracias a una agresiva estrategia de penetración de mercado de las empresas tecnológicas, consistente en brindar servicios conversacionales de forma fuertemente subvencionados y con límites transaccionales sumamente elevados. Esto instaló en el usuario —en el empresario— la percepción de que el costo de esta tecnología es despreciable, asociándola erróneamente al modelo de los buscadores de Internet, cuya gratuidad para el usuario final se sostiene de forma estable mediante ingresos publicitarios.

A nivel empresarial, este "relato" de equilibrio ganar-ganar se rompe. La brecha percibida entre los beneficios proyectados de la IA (efecto palanca sobre ingresos y reducción de costos actuales) y sus verdaderos costos de explotación desató una obsesión por lanzar proyectos de "descubrimiento interno" (*POC*) para capturar el mercado.

No obstante, los resultados de estas *POC* a nivel mundial están decepcionando. No sólo porque la adopción de esta tecnología es lenta por la falta del reconocimiento de la importancia del involucramiento de los colaboradores por la creación de la *Illusio* Socio económica (Dufour & Farré, 2025) sino también por la total subestimación y el desconocimiento de la estructura de costos que arrastran estas mejoras.

Algunos costos son directamente visibles (tarifas de *APIs*), pero la gran mayoría permanecen "ocultos" para los sistemas de información contable tradicionales, en particular los consumos derivados de accesos indirectos (sistemas integrados que llaman de forma automática a los modelos sin supervisión humana directa). Esto termina invalidando los análisis de Retorno de Inversión (*ROI*) iniciales que fueron realizados con el desconocimiento de dichos costos (físicos del volumen de uso) y con modelos determinísticos que usan variables de monetización históricas subsidiadas.

Contexto y Marco Teórico

Para dar sustento científico a nuestra propuesta metodológica, estructuramos el análisis bajo tres pilares:

El Desequilibrio de Precios y la Valoración Bursátil

Como profesionales de costos, nos topamos con una anomalía de mercado que exige respuesta:

¿Por qué los precios de las inferencias y razonamientos de la IA *Transformer* son menores a sus costos reales de procesamiento, energía e infraestructura?

Estimamos que los objetivos estratégicos de las compañías proveedoras de estos servicios (como *OpenAI*, *Anthropic* o *Google*) no están orientados en el corto plazo a la obtención de utilidades operativas tradicionales. Su meta parece ser el aumento del valor de cotización de la compañía, un indicador financiero que descuenta flujos del futuro de forma muy agresiva en comparación con los resultados del presente. En esta carrera por la dominancia del mercado y el efecto de red, los proveedores venden capacidad de cómputo a pérdida, subsidiados por inyecciones masivas de capital de riesgo y por la infraestructura ociosa de los gigantes de la nube.

La Necesidad de un Modelo de Costeo Estocástico

Dado que las tarifas vigentes del mercado de IA están subsidiadas artificialmente y el contexto futuro es altamente incierto, es un error metodológico proyectar costos de IA mediante modelos determinísticos con monetizaciones que utilizan valores históricos de subvención. Por este motivo, es que proponemos que los modelos de costeo para proyectos de IA deben ser estocásticos. Se deben formular escenarios probabilísticos de costos en los cuales las variables de inferencia y razonamiento contemplen incrementos significativos a futuro. Esto previene que una solución tecnológica que hoy se muestra rentable bajo tarifas subsidiadas, se transforme en un pasivo inviable cuando las grandes empresas de tecnología se vean obligadas a sincerar sus precios para generar utilidades reales.

Los Costos Ocultos y la Teoría Socioeconómica (TSE)

Para mapear las ineficiencias de la IA, incorporamos la Teoría Socioeconómica (TSE) de Savall y Zardet (Savall & Zardet, 2008). Esta teoría identifica que las disfunciones organizacionales generan "costos ocultos" que no aparecen en las cuentas de la contabilidad general. En un proyecto de IA con accesos indirectos y uso de tokens, estos se pueden encontrar en:

- **Tiempos de Desviación Laboral:** Personal de la empresa consumiendo horas de trabajo en auditar, corregir o reescribir respuestas erróneas o inventadas ("alucinaciones") generadas por los *LLMs*.
- **Costos Ocultos de Acceso Indirecto:** Procesos automatizados (como *triggers* de bases de datos o integraciones *API* descontroladas) que ejecutan consultas repetitivas en segundo plano sin control operativo, generando miles de llamadas invisibles al *ERP* o a la base de conocimiento corporativa.
- **Costo de Sobreconsumo por Redundancia:** Ejecución de *prompts* ineficientes, extensos y repetitivos debido a la falta de optimización técnica de los conectores o la ausencia de almacenamiento en *caché*.
- **No-Productividad por Latencia:** Cuellos de botella en procesos integrados debido a demoras en los tiempos de respuesta de los servidores externos de procesamiento de lenguaje.

- **Costos Ocultos de la opacidad de la algoritmización:** Al no visibilizar el algoritmo generado, la curaduría de contenido debe realizarse a través de prueba y error de redacción de instrucciones (*prompts*) y barreras de seguridad (*guardrails*), con resultados más ineficientes
- **Costos ocultos de la necesidad de fuertes limpieza (*cleansing*) de datos:** El principio histórico *GIGO* de los inicios de la computación, si entra Basura en una base de datos, la información que sale es basura, está más presente que nunca. Los modelos de lenguaje están pensados para operar con información exacta, y, lamentablemente, el estado de depuración de las bases actuales (internas y en la nube, estructuradas y no estructuradas) dista de ser el ideal. Toda implantación de la nueva tecnología de IAbAT necesita de depuración previa, pocas veces presupuestada en los proyectos de adopción.
- **Costos todavía ocultos del cambio de arquitectura de datos hacia información con mejor estructura, curaduría de negocio y contextualización:** En la línea del punto anterior, si bien los motores de IAbAT pueden elaborar respuestas desde bases desestructuradas y sin preprocesamientos de evolución de la Arquitectura *Medallion* (bronce-plata-oro), esto precisará mayor cantidad de tokens, elevando los costos ocultos

Modelo de Tarifación del Servicio: Unidades de Medida y Costos de Referencia

En la contabilidad de costos de IAbAT, la unidad de medida fundamental es el Token (fracción de palabra que se incrusta *-embedding-* en espacios vectoriales gigantes en función del modelo de Atención *-correlación entre ellos-* que recién empezaron a ser efectivos cuando la versión 3.5 del GPT alcanzó los 175 mil millones de parámetros en su proceso de *machine learning*). El costo total de procesamiento es de naturaleza variable y se divide en **tokens de Entrada** (*Input* o instrucción enviada) y **tokens de Salida** (*Output* o respuesta generada), tanto en el proceso original de la IAG, cuanto en el complemento del modelo de razonamiento

Como base de datos para los ejercicios de simulación de costos, presentamos los valores de mercado de referencia (por millón de *tokens*) vigentes, segmentados por proveedor y tipo de arquitectura:

TARIFACIÓN DE REFERENCIA (USD / Millón de Tokens)					
1 OpenAI - Familia GPT			2 Anthropic - Familia Claude		
Nivel de Modelo	Costo Input	Costo Output	Nivel de Modelo	Costo Input	Costo Output
 Liviano (GPT-4o-mini)	\$ 0.15 - 0.20	\$ 0.60 - 1.25	 Liviano (Claude Haiku 4.5)	\$ 1.00	\$ 5.00
 Estándar (GPT-4o)	\$ 2.50	\$10.00-15.00	 Estándar (Claude Sonnet 4.6)	\$ 3.00	\$15.00
 Razonamiento Avanzado (o3)	\$ 5.00 - 5.50	\$30.00-33.00	 Razonamiento Avanzado (Opus)	\$ 5.00	\$25.00
3 Google - Familia Gemini					
Nivel de Modelo	Costo Input	Costo Output			
 Liviano (Gemini 2.5 Flash-L)	\$ 0.10	\$ 0.40			
 Estándar (Gemini 3.5 Flash)	\$ 0.75 - 1.50	\$ 4.50 - 9.00			
 Razonamiento Avanzado (Pro)	\$ 2.00	\$12.00			

* Precios de referencia a nivel de industria

Cuadro n° 1 – Tarifas de Token de las 3 empresas más destacadas
(Elaboración propia con base del mercado 2026; ilustración con IA-Gemini)

De lo mostrado anteriormente, se observa que las tarifas de Salida (*Output*) son de 4 a 6 veces más costosas que las de Entrada (*Input*) debido a la mayor exigencia computacional de la generación de caracteres. Este factor de ponderación debe ser clave en cualquier diseño de presupuesto de costos variables (diferentes en función de rangos).

Modelo de Costeo para Proyectos de IA: Tokens y Accesos Indirectos

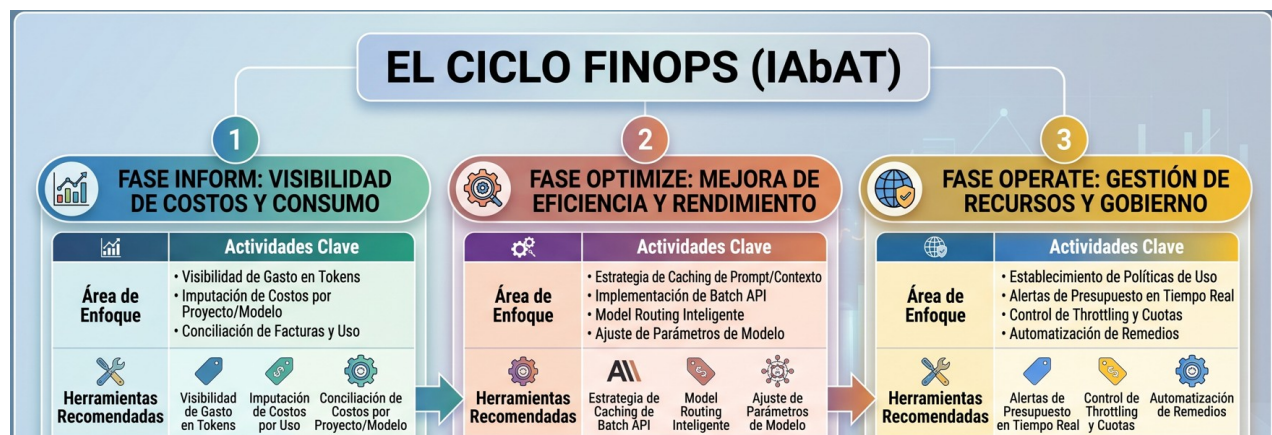
Para capturar la realidad financiera de estos desarrollos, proponemos un Modelo de Costeo de Triple Entrada que clasifica las erogaciones en directas, indirectas y ocultas:

Clasificación del Costo	Componente	Descripción y Comportamiento
Directos	Consumo de Tokens	Erogación directa medida por los tokens de Entrada y Salida multiplicados por la tarifa del API.
	Licencias de Integración	Costos fijos o por volumen de llamadas de plataformas middleware.
Indirectos	Infraestructura en la Nube	Costos de bases de datos vectoriales (indispensables para arquitecturas de búsqueda de información - RAG).
	Conectividad y Redes	Consumo de ancho de banda y latencia de transferencia de datos de servidores locales a la nube.
	Mano de Obra de Soporte	Tiempo asignado por ingenieros para mantener activos los conectores e integraciones de accesos indirectos.
Ocultos	Costos de Alucinación	Tiempo de analistas humanos corrigiendo respuestas inconsistentes de los LLMs. Pérdidas económicas por el impacto de dichas alucinaciones en los negocios, cuando IAbAT actúa en forma directa sobre clientes y no como copilotos que permiten la corrección previa a la negociación
	Llamadas Redundantes por Acceso Indirecto	Triggers o consultas automatizadas en bucle que ejecutan peticiones de tokens invisibles para el usuario final.
	Re-ejecuciones por Caída de Conectividad	Re-intentos automáticos de conexión a la API ante micro-cortes, duplicando el consumo de tokens de entrada.
	Curva de Aprendizaje	Horas laborables dedicadas a capacitar a la dotación de personal en el uso eficiente de prompts y parametrización.

Cuadro n° 2 – Modelo de Costeo triple entrada (Elaboración e Ilustración propia).

Gestión de Costos: Aplicación del Ciclo *FinOps* para IA, adaptado: *FinOps+*

Para garantizar una gestión de costos sostenible, estructuramos el control de accesos indirectos y uso de *tokens* bajo el ciclo de vida oficial del *framework* que aparece como el más utilizado en Corporaciones europeas, el *FinOps* (*Inform, Optimize, Operate*) adaptado a la visión de la Teoría General del Costo y la modelización estocástica con escenarios de proyección con incertidumbre:



Cuadro n° 3 – Ciclos de *FinOps* (Elaboración con información de *FinOps* adecuada propiamente; ilustración con IA - Gemini)

1) Fase informativa (*Inform*): Visibilidad, Asignación y Atribución

El objetivo prioritario de esta fase es lograr una visibilidad y atribución del costo del 100%.

- Observación multidimensional de consumos y accesos a datos: Configurar metadatos y etiquetas para entender el consumo directo de tokens multidimensionalmente (distintos objetos de costeo, actividades de usuarios que los solicitan, unidades de negocio, geografía, segmentos de clientes, etc), y la facturación por accesos indirectos a datos de las bases de los Aplicativos que las administran (ej. "ERP SAP", "CRM Salesforce").
- Conciliación e Ingesta de Datos: Cruzar los registros transaccionales internos con las APIs de facturación de los proveedores para evitar desvíos o sobrefacturaciones por reintentos de conexión.
- Observación de sacrificios necesarios para sanear la calidad de datos que son fuente para la IAbAT: Uno de los motivos de la falta de cumplimiento de las expectativas de la IAG tiene que ver con la mala calidad de los datos disponibles para producir las inferencias correctas. Estos esfuerzos suelen estar despriorizados, y por lo tanto ocultos en los modelos de cálculo. Es necesario replantear la arquitectura de datos actual, migrando a las mejores prácticas de la *Medallion* y sus tres niveles:
 - Bronce: Datos existentes en forma de materia prima (*Raw data*)
 - Plata: Esos datos pasan por procesos de *BDA* (*Big Data Analytics*) y estructurados para su mejor gestión técnica.
 - Oro: Una vez estructurados, los especialistas de gestión de negocios les realizan curaduría de contenidos.

Aplicando la visión de la Contabilidad de Gestión con base en la TGC, es importante resaltar que la captura de la información física (p.ej. cantidad de *tokens*) deberá ser con la granularidad y

direccionalidad de origen (capturados de aplicativos transaccionales) y la monetización debe aplicar los precios reales en función de las escalas acordadas, evitando los procedimientos tradicionales de la Contabilidad patrimonial financiera que suelen perder detalle, calidad y, en particular, la direccionalidad con los objetos de costeo de los costos necesarios (de allí el verbo de procesos de “asignación” de valores agregados capturados por los *ERP*).

2) Fase de Optimización (*Optimize*): Eficiencia en Uso y Tarifas

Una vez obtenido el mapa de consumo en la fase anterior, el enfoque *FinOps* aplica palancas técnicas de optimización orientadas al mejor consumo de *tokens*, como los siguientes ejemplos:

- Almacenamiento en *Caché* (*Context Caching*): En integraciones de accesos indirectos que repiten pesados manuales normativos o bases de datos fijas, el almacenamiento en *caché* de contexto evita reprocesar los mismos *tokens* de entrada, pagando solo una pequeña fracción del costo.
- Procesamiento por Lotes (*Batch API*): Para tareas en segundo plano que no requieren inmediatez síncrona (como análisis diferido de órdenes de compra), se despachan solicitudes en bloque para acceder a un descuento directo del 50% en la facturación de *tokens*.
- Enrutamiento Estratégico (*Model Routing*): Enrutamiento automatizado donde las llamadas mecánicas del sistema se derivan a modelos económicos (p.ej. *GPT-4o-mini* o *Gemini 2.5 Flash-Lite*) y solo las consultas complejas avanzan hacia arquitecturas de razonamiento superior (*Claude Opus*).
- Utilización de bases de datos de nivel “plata” u “oro” (siguiendo la denominación de la arquitectura de datos *Medallion*): El uso de conectores homologados sobre las bases de datos operativas (que tienen en algún caso facturación por accesos “indirectos” desde los agentes de la IAbAT) permite extraer datos en su estado primario (*Raw data*) “bronce” y estructurarlos en bases del nivel superior (“plata”). El uso a partir de ese nivel (y del nivel curado por negocio “oro”) no sólo no recibirán costos adicionales, sino que minimizarán la generación de *tokens*.
- Incorporación de la capa contextual de negocio al modelo *Medallion*, para que sea propia de la base de conocimientos de la empresa y centralizada en su gestión, evitando diferencias si se definen en cada uno de los prompts

Agregando la visión de la TGC y de la gestión basada en actividades (*ABM*), a la palanca técnica le debemos agregar la revisión de necesidad de negocio: Si el costo de la actividad desarrollada por la IAbAT no es necesario o no agrega mayor valor que sus costos, deberá eliminarse o replantearse

3) Fase operativa (*Operate*): Gobierno y Ejecución Continua

Conociendo que la IAbAT, en su esencia, no es una tecnología exacta, cuenta con un concepto clave que debe implantarse para detectar tempranamente los problemas en sus modelos y agentes: el “espacio de observabilidad” (EO). Se define como EO a un conjunto de entornos y paneles de control

para evaluar, medir y explicar el comportamiento de los modelos de inteligencia artificial y agentes.

Consiste, tanto en la ejecución de acciones de control en línea, como, por ejemplo:

- **Alertas Tempranas de Desvío:** Notificaciones automáticas cuando el consumo acumulado de un canal o acceso indirecto supere los límites del presupuesto asignado para el período actual.
- **Políticas de *Throttling* (Interrupción Operativa):** Configurar controles de tasa límite (*Rate Limiting*) en las pasarelas de conexión para detener de forma inmediata consumos infinitos o bucles de código de software antes de que generen un impacto financiero severo.

En cuanto la generación de paneles de control incluye métricas claves como:

- Invocaciones totales y promedio
- Uso de *tokens* totales y promedio, abierto entre entrada y salida
- Cálculo de costos multidimensional
- Latencia (promedio y de distintos percentiles (90, 95 y 99))
- Tasas de error (por alucinación o toma de decisiones equivocadas en aquellas instalaciones que le permitieron discrecionalidad (temperatura tibia o caliente))
- Impactos económicos en la empresa por errores de la IAbAT (Modelos “EVAL”)

4) Fase de cambios disruptivos: Análisis de impactos y reacción rápida

El concepto de cambios disruptivos suele ser asociado al término “Cisne negro” (Taleb, 2022) por aparecer cuando “nadie” lo espera. En su libro de igual nombre, Taleb narra el origen de la “etiqueta”: Biólogos de Occidente, que en sus geografías sólo conocían cisnes blancos, se sorprendieron cuando encontraron cisnes de aquel color en Australia en el siglo XVII. Con el “diario del lunes”, ¿era tan inesperada la existencia de cisnes negros o sólo eran inesperados para científicos con sesgos, que no podían establecer patrones de conocimiento “fuera de su caja”?

En IAbAT, ¿será considerado inesperado el aumento de precios de los principales proveedores de esta tecnología (*software, hardware*, infraestructura de comunicaciones y almacenamiento, canales de comunicación como *whatsapp*, etc) una vez que quieran empezar a recuperar inversiones?

Sugerimos incluir una fase 4 en *FinOps+* que monitoree cambios significativos y dispare modelos analíticos de impacto, listos para sugerir cambios de tecnología.

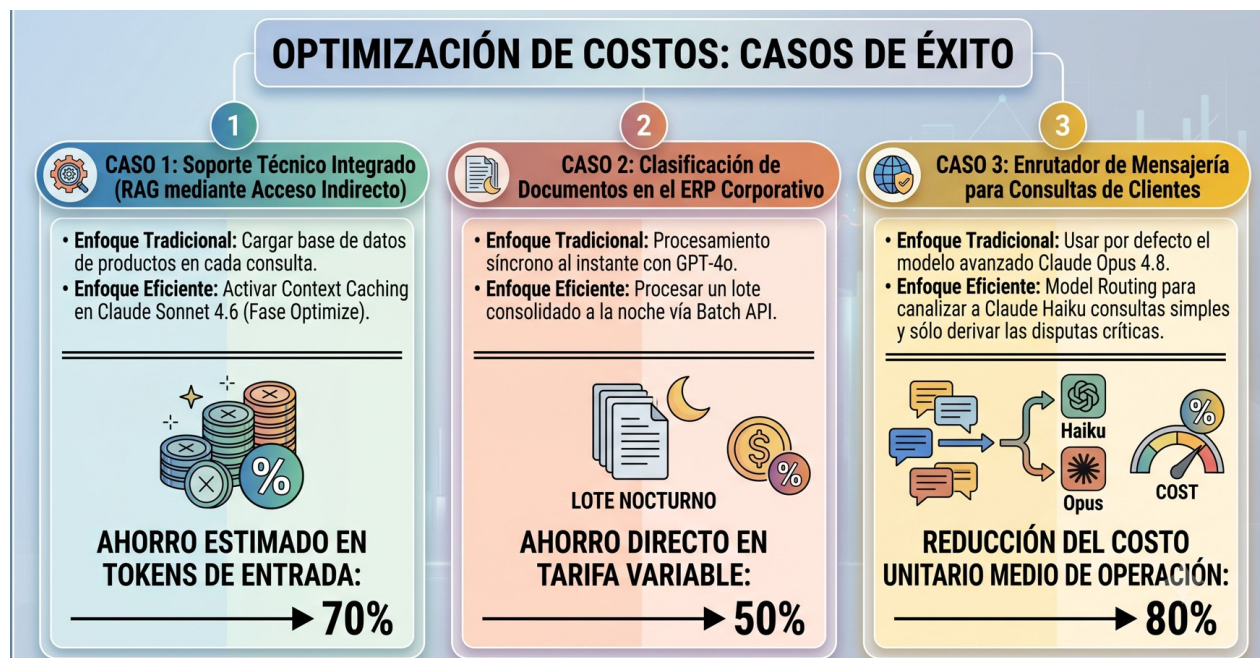
5) Fase híbrida recurrente

El planteo *FinOps* IA de tres fases (y su secuencia) parece obedecer a lo intempestivo y disruptivo del impacto de la IAG a partir de noviembre de 2022. Sin embargo, una vez recorridas esas tres etapas para ordenar el caos, sugerimos que las tres fases operen concomitantemente bajo el enfoque de “desafiante versus campeón” de Marchione y Farré (Marchione & Farré, 2025) para disminuir los riesgos del cambio y acelerar los tiempos de implantación de la propuesta de mejora desafiante:

- Toda vez que se encuentre una idea potencialmente superadora en fase “optimizadora” implantarla en un esquema piloto pequeño,
- Toda vez que se opera en etapa “operativa” se compara a los desafiante con los campeones
- En todo momento se costean ambos, y se comparan los costos diferenciales

Casos Habituales de Eficientización y Simulación de Impacto

A efectos de ilustrar el impacto de la gestión de costos bajo el ciclo *FinOps*, planteamos tres casos típicos de optimización aplicados sobre modelos de producción empresarial con accesos indirectos:



Cuadro n° 4 – Casos de ejemplo (Casos de elaboración propia e ilustración con IA - Gemini)

Tablero de Comando: Control y Desvíos

El control de gestión requiere un cuadro de mando integral con indicadores cuantitativos claros y un sistema analítico de variaciones adaptado a la problemática de accesos indirectos y uso de *tokens*.

KPIs de Gestión de Costos de IA (Umbrales de Semáforo)

Proponemos la adopción de métricas diseñadas para este ecosistema operativo. Por ejemplo, para los casos expuestos se pueden desarrollar estos:

1. Costo de IA por Transacción Integrada (CIAT):

$$CIAT = \frac{\text{Costo Consumido de Tokens} + \text{Costos Indirectos e Interfaces}}{\text{Volumen de Transacciones de Negocio Procesadas Correctamente}}$$

- **Verde:** $CIAT < Xa \text{ Usd}$ (Acceso indirecto rentable), en donde xa representa al umbral de aceptabilidad
- **Amarillo:** $Xa \text{ Usd} < CIAT < Xb \text{ Usd}$ (Alerta de control operativo), en donde Xb representa el umbral de peligro
- **Rojo:** $CIAT > xb \text{ Usd}$ (Requiere optimización de prompt o arquitectura).

2. Eficiencia de Caché de Contexto (ECC):

$$ECC = \frac{\text{Tokens de Entrada Servidos desde Memoria Caché}}{\text{Total de Tokens de Entrada Enviados}}$$

- ● **Verde:** $ECC > 70\%$ (Aprovechamiento óptimo de contexto de caché).
- ● **Amarillo:** $40\% < ECC < 70\%$ (Margen de mejora técnica).
- ● **Rojo:** $ECC < 40\%$ (Ineficiencia crítica en prompts e integraciones).

3. Tasa de Llamadas en Lote por Acceso Indirecto (TBR):

$$TBR = \frac{\text{Llamadas ejecutadas bajo API Batch}}{\text{Total de llamadas integradas no síncronas}}$$

- ● **Verde:** $TBR > 50\%$ (Aprovechamiento activo de descuentos).
- ● **Amarillo:** $30\% < TBR < 50\%$ (Tareas de fondo ejecutándose en tiempo real de forma ineficiente).
- ● **Rojo:** $TBR < 30\%$ (Ineficiencia presupuestaria severa).

Análisis Ex-Post por Desvíos de Tokens para aprendizaje

Al finalizar el período, la diferencia entre el Costo Presupuestado de IA (CP_{IA}) y el Costo Real (CR_{IA}) debe desglosarse mediante fórmulas de variaciones adaptadas:

$$\Delta \text{Costo Total} = \text{Desvío Tarifa (Precio)} + \text{Desvío Eficiencia (Eficiencia de Prompts)} + \text{Desvío Volumen (Demanda de Negocio e Integración)}$$

- **Desvío Tarifa (Precio):** Variaciones originadas por incrementos unilaterales de precios por parte del proveedor o por la decisión de migrar el desarrollo a una *API* de mayor coste de inferencia.
- **Desvío Eficiencia (Consumo de Tokens):** Mide la variación debida al diseño ineficiente de instrucciones, llamadas redundantes del sistema intermedio o reintentos automáticos ante caídas que multiplicaron los *tokens* consumidos de forma oculta.
- **Desvío Volumen (Actividad):** Variación explicada directamente por el crecimiento en la demanda real de uso de las aplicaciones y accesos indirectos por parte de los clientes o procesos del negocio.

A modo de conclusión

Estamos escribiendo esta ponencia en 2026, siendo contemporáneos a lo que creemos que es el fin de la hipótesis (*AI hype*) de la IAbAT y al fin del “cheque en blanco” que las direcciones de las corporaciones dieron a esta tecnología (y su filosofía *AI First*), con lo cual sería muy aventurero definir a este capítulo como conclusión.

Dicho esto, repasemos lo expresado para el objetivo de la ponencia: “desarrollar un marco metodológico para la Gestión de Costos por el uso de *tokens* y accesos indirectos de Inteligencia Artificial *Transformer* en el ámbito corporativo, alineado con el *framework FinOps*, con ajustes”

Nuestra propuesta incluye conceptos de la TGC y la TSE que aplicamos sobre el modelo base *FinOps*, en particular:

1. La importancia de la información física, que debe ser capturada en el origen transaccional y con la mayor granularidad posible
2. La importancia de tener en claro la necesidad en función del objetivo de negocio
3. La incorporación de costos (ocultos habitualmente) de limpieza de datos y de su preprocesamiento evolutivo para crear capas de “plata” y “oro”
4. La relevancia de la precisión de las instrucciones en la disminución de la variabilidad del costo: Un *prompt* diseñado bajo técnicas profesionales de estructuración reduce de manera drástica las alucinaciones, la variabilidad del costo y abarata la integración corporativa.
5. El impacto de las alucinaciones debe estar en los cálculos de costos visibilizados. Por ello deben incluirse costos contingentes para cubrir los errores habituales.
6. El acceso indirecto multiplica la ineficiencia silenciosa: Un error de configuración técnica en un conector del sistema puede generar millones de solicitudes automáticas de *tokens* en cuestión de minutos, devorando presupuestos de manera invisible antes de que intervenga un analista humano.

¿Serán estos los cambios que apliquen las empresas? Puede que sea estos u otros. Lo que estamos seguros es que deberán implantar algunos en el corto plazo si no quieren pasar por el caos que genera el exceso de expectativas insatisfechas.

En cuanto a los objetivos intermedios, en cada hoja del desarrollo se van desplegando ideas que son, para los autores, sólo disparadoras de debates en nuestro Congreso de IAPUCo en La Punta.

Bibliografía

- CloudZero. (2026). Anthropic Claude API Pricing In 2026: Every Model, Token Rate, And Cost Lever. <https://www.cloudzero.com/blog/claude-api-pricing/>.
- CloudZero Research. (2026). Finding the ROI of AI: The Finance Perspective. <https://www.cloudzero.com/finance-needs-ai-roi-2026-survey-report/>.
- Dufour, G., & Farré, D. (OCTUBRE de 2025). *La distribución equitativa de las mejoras de la productividad laboral sustentadas por IAGen*. Obtenido de <https://iapuco.org.ar/>: <https://iapuco.org.ar/wp-content/uploads/2025/10/Mar-del-Plata-2025-12.pdf>
- FinOps Foundation. (2026). The FinOps Framework: Cloud Financial Management for Technology Value. <https://www.finops.org/>.
- Marchione , J., & Farré, D. (2025). *La naturalización del cambio, ¿minimiza sus costos?* Obtenido de <https://iapuco.org.ar/>: <https://iapuco.org.ar/wp-content/uploads/2025/10/Mar-del-Plata-2025-4.pdf>
- Savall , H., & Zardet, V. (2008). *Mastering Hidden Costs and Socio-Economic Performance*. IAP.
- SemiAnalysis Newsletter. (2026). AI Dark Output: The Visible Cost of Generative Models. <https://newsletter.semianalysis.com/p/ai-dark-output-the-visible-cost-of>.
- Taleb, N. N. (2007). *El Cisne Negro: El impacto de lo altamente improbable*. Paidós ibérica.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., L, J., Gomez, A., . . . Polosukhin, I. (2017). Attention is all you need.