Recibido: 01-01-2026 / Revisado: 21-04-2026
Aceptado: 25-04-2026 / Publicado: 31-07-2026

Contabilidad de los tokens y enrutamiento dinámico en la inteligencia artificial generativa aplicada a Sistemas de Gestión del Aprendizaje (LMS): Un modelo económico para la educación superior

Token Accounting and Dynamic Routing in Generative AI Applied to Learning Management Systems (LMS): An Economic Model for Higher Education

Álvaro Martínez-García[1]

[1] Universidad Francisco de Vitoria, España

a.martinezgarcia@ufv.es

RESUMEN. La presente investigación desarrolla un marco teórico y un modelo económico-analítico destinado a determinar la asignación eficiente del capital computacional derivado del uso de Modelos de Lenguaje a Gran Escala (LLM) en la infraestructura de enseñanza virtual universitaria. El objetivo central consiste en minimizar el gasto operativo (OPEX) asociado al consumo de tokens de inferencia en Sistemas de Gestión del Aprendizaje (LMS), garantizando simultáneamente unos resultados pedagógicos análogos a los obtenidos mediante el uso monocrómico y exclusivo de modelos fundacionales de frontera y alto coste. El estudio se fundamenta en los postulados de la Economía del Token (Tokenomics), las Operaciones Financieras en la Nube (FinOps) y el paradigma emergente de la Inteligencia Artificial Frugal (Frugal AI). Se formaliza matemáticamente un sistema de soporte a la decisión basado en el enrutamiento predictivo de modelos (LLM Routing) adaptado a entornos educativos. Mediante una simulación determinista a gran escala, estructurada sobre la arquitectura nativa del subsistema de IA de plataformas LMS contemporáneas (específicamente Moodle 4.5), se evalúan diversos escenarios presupuestarios y de gestión computacional. El análisis empírico evidencia que una estrategia de enrutamiento estratificado —que clasifica las consultas académicas según su complejidad algorítmica ex ante— puede reducir los costes de inferencia institucionales en más de un 71% frente al status quo. De forma paralela, el modelo logra mantener una tasa de suficiencia instruccional superior al 97%, cerrando de forma efectiva la brecha de equidad en el acceso tecnológico (AIED Advantage Gap) al democratizar el soporte cognitivo avanzado.

ABSTRACT. This research develops a theoretical framework and an economic-analytical model aimed at determining the efficient allocation of computational capital derived from the use of Large Language Models (LLMs) in university virtual teaching infrastructure. The central objective is to minimize the operational expenditure (OPEX) associated with the consumption of inference tokens in Learning Management Systems (LMS), while simultaneously guaranteeing pedagogical results analogous to those obtained through the exclusive use of high-cost frontier foundational models. The study is based on the postulates of Token Economics (Tokenomics), Cloud Financial Operations (FinOps) and the emerging paradigm of Frugal Artificial Intelligence (Frugal AI). A decision support system based on predictive model routing (LLM Routing) adapted to educational environments is mathematically formalized. Through a large-scale deterministic simulation, structured on the native architecture of the AI subsystem of contemporary LMS platforms (specifically Moodle 4.5), various budgetary and computational management scenarios are evaluated. The empirical analysis shows that a stratified routing strategy—which classifies academic queries according to their algorithmic complexity ex ante—can reduce institutional inference costs by more than 71% compared to the status quo. At the same time, the model manages to maintain an instructional sufficiency rate of over 97%, effectively closing the technology access equity gap (AIED Advantage Gap) by democratizing advanced cognitive support.

PALABRAS CLAVE: Economía de la información, Inteligencia artificial frugal, Enrutamiento de LLMs, Sistemas de Gestión del Aprendizaje, Moodle, Contabilidad de costes, Educación superior.

KEYWORDS: Information economics, Frugal artificial intelligence, LLM routing, Learning Management Systems, Moodle, Cost accounting, Higher education.

Martínez-García, A. (2026). Contabilidad de los tokens y enrutamiento dinámico en la inteligencia artificial generativa aplicada a Sistemas de Gestión del Aprendizaje (LMS): Un modelo económico para la educación superior. *Campus Virtuales,* 15(2), 47-62. https://doi.org/10.54988/cv.2026.2.1956

1. Introducción

La digitalización de las instituciones de educación superior ha experimentado una aceleración sistémica sin precedentes en la última década, consolidando a los Sistemas de Gestión del Aprendizaje (LMS por sus siglas en inglés, Learning Management Systems), tales como Moodle, Canvas o Blackboard, no solo como meros repositorios de contenido estático, sino como los ecosistemas centrales ineludibles para la interacción pedagógica, la evaluación y la administración académica (Castro Benavides et al., 2020; Selgas-Cors, 2025). En este contexto de transformación digital continua, caracterizado por crecientes presiones financieras debido a la expansión de la población estudiantil y los persistentes desafíos de acceso equitativo (Awad et al., 2025), la irrupción de la Inteligencia Artificial Generativa (IAG) y los Modelos de Lenguaje a Gran Escala (LLM) ha inaugurado un nuevo ciclo tecnológico que redefine radicalmente las capacidades operativas de estas plataformas (Dirin et al., 2023; Martín-Ramallal, 2025).

Históricamente, los sistemas LMS operaban bajo un paradigma de base de datos relacional estática, donde la universidad asumía costes predecibles vinculados al ancho de banda, el almacenamiento en servidores y las licencias de usuario (SaaS) (Zhu, 2026). Sin embargo, la reciente integración de arquitecturas de IA de forma nativa en sistemas de código abierto —ejemplificada magistralmente por la introducción del "Subsistema de IA" (AI Subsystem) a partir de la versión de Moodle 4.5— permite ahora a docentes y estudiantes interactuar con agentes algorítmicos para la generación de contenidos, la evaluación formativa contextualizada, el resumen de textos extensos y la tutorización personalizada en tiempo real (Atsalakis et al., 2024; Baabdullah et al., 2021; CYPHER Learning, 2025; Schoser, 2023).

En el epicentro de esta revolución educativa y tecnológica emerge una nueva unidad económica fundamental que subvierte los modelos de contabilidad tradicionales: el token computacional (Storment, 2026; Zhu, 2026). Originalmente concebido como un artefacto puramente técnico para la fragmentación y vectorización del lenguaje natural dentro de las redes neuronales artificiales (arquitecturas de transformadores), el token se ha transmutado rápidamente en el activo subyacente que estructura, cuantifica y monetiza el mercado global de la inteligencia computacional (Storment, 2026; Zhu, 2026). Los principales proveedores de modelos fundacionales (tales como OpenAI, Google, Anthropic o Microsoft Azure) emplean esta unidad métrica para fijar precios, asignar y gobernar la capacidad de cómputo en la nube (Intelligent Living, 2026; Storment, 2026).

En consecuencia, bajo este nuevo paradigma, cualquier interacción de un usuario con el LMS universitario —ya sea la solicitud de un estudiante para que un tutor virtual le explique un concepto de física, la orden de un profesor para generar un banco de preguntas a partir de un temario, o el mero resumen automático de las participaciones en un foro— se traduce imperativamente en un consumo cuantificable de tokens (Atsalakis et al., 2024; Storment, 2026). Esta dinámica vincula de forma directa, instantánea e indisoluble el procesamiento algorítmico y el razonamiento heurístico con la factura eléctrica de los centros de datos corporativos y, en última instancia, con el estado de flujos de efectivo del presupuesto universitario (LaunchDarkly, 2026; Zhu, 2026).

Este escenario impone un desafío monumental para la contabilidad de gestión y las finanzas operativas (FinOps) de las universidades. A diferencia de los modelos de licenciamiento tradicionales, la facturación de los servicios de inferencia LLM a través de Interfaces de Programación de Aplicaciones (API) opera bajo un régimen de consumo elástico, hipervariable y marcadamente estocástico (Deloitte, 2026; Storment, 2026; Zhu, 2026). El coste marginal de una consulta académica fluctúa drásticamente dependiendo de la longitud del documento base proporcionado (context window), la complejidad algorítmica de las instrucciones (prompts), la necesidad de bucles iterativos en los agentes conversacionales y, fundamentalmente, la elección del modelo fundacional subyacente (Storment, 2026; Zhu, 2026). Las investigaciones en el ámbito corporativo ya advierten que, sin mecanismos de gobernanza estrictos e instrumentados tecnológicamente, la adopción descentralizada de IA puede desencadenar incrementos exponenciales, ruinosos y no lineales en los gastos operativos (OPEX) (Deloitte, 2026; FinOps Foundation, 2026).

Martínez-García, A. (2026). Contabilidad de los tokens y enrutamiento dinámico en la inteligencia artificial generativa aplicada a Sistemas de Gestión del Aprendizaje (LMS): Un modelo económico para la educación superior. *Campus Virtuales,* 15(2), 47-62. https://doi.org/10.54988/cv.2026.2.1956

A pesar de la gravedad y urgencia del problema económico descrito, se evidencia una profunda brecha teórica y práctica en la literatura científica actual (research gap). La abrumadora mayoría de las investigaciones contemporáneas sobre IA en la educación superior se concentran casi en exclusiva en las dimensiones pedagógicas, el impacto en el rendimiento cognitivo, las controversias sobre la integridad académica (plagio) y los modelos de aceptación tecnológica por parte del alumnado (Dirin et al., 2023; Kokina & Davenport, 2017; Lai, 2025). Por el contrario, los estudios enfocados en la dimensión microeconómica, la viabilidad contable y la eficiencia del capital tecnológico de la integración de la IA en la infraestructura de los campus virtuales son excepcionalmente escasos o inexistentes. Las instituciones de educación superior corren el inminente riesgo de incurrir en un fenómeno de "consumo en la sombra" (shadow AI), delegando tareas administrativas triviales o consultas académicas elementales a modelos de frontera masivamente costosos, desperdiciando recursos públicos o privados que podrían haberse resuelto con modelos destilados con una fracción del coste (FinOps Foundation, 2026; Zhu, 2026).

Para solventar esta carencia en la intersección entre la economía empírica, los sistemas de información y la tecnología educativa, la presente investigación plantea el desarrollo de un modelo matemático, econométrico y contable que operativice los principios de la Inteligencia Artificial Frugal (Frugal AI) (Gärtner & Hiebl, 2018; Ittner & Larcker, 2001; Prabhu, 2026) en los entornos LMS de gran escala. Las preguntas fundamentales de investigación que vertebran este estudio son: (1) ¿De qué manera pueden las universidades formular un modelo de contabilidad analítica que optimice de forma algorítmica la relación entre el coste financiero de inferencia de la Inteligencia Artificial Generativa y la utilidad pedagógica de las respuestas proporcionadas en su campus virtual? y (2) ¿Qué arquitecturas informáticas de enrutamiento dinámico (LLM Routing) y orquestación configuran la frontera óptima de eficiencia de Pareto para minimizar el consumo del presupuesto tecnológico sin comprometer o degradar la calidad instruccional y la equidad educativa?

La contribución esperada de este artículo es dual. Desde una perspectiva estrictamente teórica y en el ámbito de la Economía de la Información, el estudio expande el corpus existente al formalizar matemáticamente el token generativo como un factor de producción variable y analizable dentro de la función de costes institucionales. Desde un enfoque eminentemente aplicado, esta investigación proporciona a los vicerrectorados de transformación digital, directores de tecnología (CIOs) y administradores de plataformas e-learning un modelo de soporte a las decisiones, contrastado empíricamente, que fundamenta y justifica la adopción de esquemas de enrutamiento predictivo. Al derivar sistemáticamente las tareas sencillas hacia modelos de bajo coste y reservar el capital computacional premium exclusivamente para razonamientos que exigen alta complejidad cognitiva, el modelo propuesto garantiza que la modernización tecnológica universitaria se convierta en una palanca de innovación equitativa, gobernable y financieramente sostenible en el largo plazo (Murray et al., 2024; Xu, 2026; Yaşar, 2024).

2. Marco Teórico y Desarrollo de Hipótesis

El sustento conceptual que vertebra esta investigación exige la convergencia disciplinar de la evolución arquitectónica de los Sistemas de Información Educativos, la teoría microeconómica aplicada al coste de la computación (Tokenomics), y los avances más recientes en las ciencias de la computación orientadas a la eficiencia algorítmica y orquestación de modelos (Frugal AI y LLM Routing).

2.1. Transición Tecnológica: De la Docencia Estática a la IA Generativa en los Campus Virtuales

Históricamente, el diseño de los Sistemas de Gestión del Aprendizaje se ha fundamentado en arquitecturas asimétricas y relacionales orientadas al almacenamiento de datos y a la gestión del flujo de trabajo (Selgas-Cors, 2025). En este modelo tradicional, el LMS actuaba predominantemente como un conducto logístico pasivo: el personal docente depositaba recursos de aprendizaje (documentos, presentaciones multimedia, rúbricas) y el discente los consumía, interactuando de forma fuertemente pautada a través de foros asíncronos o cuestionarios de corrección automatizada determinista (Selgas-Cors, 2025). Esta topología intrínseca imponía severos límites a la personalización del aprendizaje masivo, toda vez que cualquier adaptación o

retroalimentación dependía inexorablemente de las restricciones de tiempo y esfuerzo del capital humano docente.

La eclosión y madurez de la Inteligencia Artificial Generativa ha subvertido de raíz este ecosistema, transformando el LMS, conceptual y operativamente, de un repositorio estático a un agente cognitivo activo e interactivo (Dirin et al., 2023; Lai, 2025; Schoser, 2023). Un hito paradigmático y definitorio en esta transición es la arquitectura del "Subsistema de IA" (AI Subsystem), introducido de forma nativa en plataformas de código abierto líderes a nivel global, concretamente a partir del lanzamiento de la versión 4.5 de Moodle (Baabdullah et al., 2021; CYPHER Learning, 2025; LaunchDarkly, 2026). A diferencia de los enfoques previos basados en complementos (plugins) de terceros fragmentados y mal integrados, este subsistema constituye una capa arquitectónica unificada dentro del núcleo del software (Moodle Core). Proporciona un marco formal, consistente y centralizado para la gobernanza, el registro de auditoría (logging) y la integración de funcionalidades de IA a lo largo de toda la experiencia de usuario (CYPHER Learning, 2025; LaunchDarkly, 2026).

Es crucial comprender desde la perspectiva de la arquitectura de sistemas que este subsistema no posee inteligencia autónoma intrínseca; por el contrario, actúa como un orquestador o middleware de enrutamiento (Khan et al., 2025; LaunchDarkly, 2026). Conecta las acciones formativas solicitadas por el usuario —por ejemplo, la petición de resumir un hilo de debate extenso o la demanda de tutorización conversacional— con proveedores externos de modelos cognitivos a través de Interfaces de Programación de Aplicaciones (API) (Storment, 2026).

Esta flexibilidad estructural permite a las universidades la conexión simultánea y parametrizable con un abanico heterogéneo de proveedores. Incluye tanto ofertas comerciales de código cerrado y alto rendimiento (tales como OpenAI, Microsoft Azure, Google Gemini o Anthropic) como soluciones emergentes de pesos abiertos (open-weights) y ejecución local (como Ollama, LocalAI o LiteLLM) (Baabdullah et al., 2021; Chen et al., 2023; LaunchDarkly, 2026). Desde una óptica financiera institucional, aunque el subsistema de software en sí mismo no requiere licencias de uso por su naturaleza de código abierto, la invocación constante de las API comerciales introduce un flujo continuo y agresivo de costes variables directos. De este modo, la institución externaliza el esfuerzo cognitivo hacia las infraestructuras en la nube de terceros, sometiéndose a sus modelos de tarificación basados en la tokenización (LaunchDarkly, 2026; Storment, 2026).

2.2. La Microeconomía del Token y el Paradigma FinOps Universitario

Para conceptualizar con rigor científico el impacto financiero de la Inteligencia Artificial Generativa en la educación superior, es imperativo analizar la Economía del Token (Tokenomics) como un subcampo robusto de la Economía de la Información (Zhu, 2026). En este constructo económico, el token neuronal (que equivale de forma aproximada a tres cuartas partes de una palabra en los idiomas occidentales principales) manifiesta una naturaleza microeconómica triple: opera como la materia prima del procesamiento de información, como la métrica de desgaste de los aceleradores de hardware (GPUs), y como la unidad de cuenta transaccional comercial (Storment, 2026; Zhu, 2026).

La estructura de formación de costes de los proveedores de LLMs replica dinámicas clásicas de escrutinio multidimensional y discriminación de precios de segundo grado, teorizadas originalmente por Mussa y Rosen (1978). Las corporaciones tecnológicas estratifican su oferta de modelos para que los compradores institucionales se autoseleccionen en función de su disposición a pagar y su necesidad de precisión. Sin embargo, el mercado de la inteligencia artificial exhibe una profunda asimetría arancelaria estructural: el precio marginal asociado a la ingesta de información (tokens de entrada, input tokens o prompts) es sistemáticamente inferior —con diferenciales que oscilan entre 3 y 10 veces— respecto al coste de generación de respuestas (tokens de salida, output tokens o completions) (Intelligent Living, 2026; Storment, 2026). Esta discrepancia no es un mero capricho comercial, sino que se deriva directamente de las leyes de la física computacional. En la arquitectura Transformer, la asimilación del contexto de entrada puede paralelizarse de manera masiva,

mientras que la generación de texto exige un costoso procesamiento autorregresivo, secuencial y estocástico.

Para las instituciones académicas, las implicaciones de esta asimetría son críticas y peligrosas. La tutorización virtual de alta calidad exige que los modelos anclen sus explicaciones en los materiales validados del curso (un proceso documentado en la literatura técnica como In-Context Learning o Generación Aumentada por Recuperación, RAG) (Iternal, 2026). Proveer a un LLM del contexto pertinente —que puede abarcar libros de texto en formato PDF, normas de estilo, o extensos historiales de foros de discusión— implica inyectar volúmenes masivos de tokens de entrada por cada interacción individual del estudiante.

En arquitecturas de agentes autónomos (Agentic Workflows), donde el tutor virtual dialoga repetidamente con el alumno, el historial completo de la conversación debe reintroducirse en la "ventana de contexto" del modelo en cada nuevo turno. Este fenómeno genera el denominado Scrollback Cost (coste de arrastre del historial acumulativo) (Iternal, 2026; Zhu, 2026). Sin barandillas de control (guardrails), un solo estudiante interactuando con un modelo de frontera para preparar un examen puede desencadenar una espiral de inflación de tokens ocultos que devore rápidamente el presupuesto de la asignatura.

Ante la materialización de esta vulnerabilidad, las ciencias de la empresa proponen la transposición de los principios de las Operaciones Financieras en la Nube (Cloud FinOps) al ámbito de la Inteligencia Artificial (FinOps Foundation, 2026). El paradigma de AI FinOps prescribe que las universidades, de igual forma que el sector corporativo, deben abandonar la provisión tecnológica ciega para instaurar marcos de trazabilidad extrema (tagging de tokens por facultad o alumno), presupuestación predictiva y optimización algorítmica constante (técnicas de Right-sizing), evitando que la fascinación por la automatización debilite la viabilidad financiera de las infraestructuras públicas o privadas (FinOps Foundation, 2026).

2.3. Inteligencia Artificial Frugal (Frugal AI) y Equidad en el Acceso Educativo

La intensa fricción microeconómica generada entre la sofisticación tecnológica demandada por el proceso de enseñanza y las acuciantes restricciones presupuestarias ha catalizado el surgimiento del movimiento de la Inteligencia Artificial Frugal (Frugal AI). Impulsado por instituciones académicas e intergubernamentales de prestigio, como el Cambridge Frugal AI Hub de la Universidad de Cambridge y la Commonwealth of Learning (COL), este nuevo paradigma rechaza de plano la asunción, prevalente en Silicon Valley, de que los recursos computacionales son virtualmente ilimitados (Ittner & Larcker, 2001; Lingamgunta, 2026; Prabhu, 2026).

La IA Frugal se conceptualiza como un desplazamiento epistemológico hacia el diseño de sistemas algorítmicos que logren el máximo impacto y valor social con el consumo estrictamente mínimo de recursos tangibles (energía eléctrica, potencia de cálculo de GPU, ancho de banda y capital monetario) (Gärtner & Hiebl, 2018; Ittner & Larcker, 2001). Promueve la eficiencia desde el diseño, interrogándose sobre qué nivel exacto de capacidad paramétrica es realmente necesario para resolver una tarea educativa específica, y priorizando el uso de modelos destilados, eficientes y, cuando es posible, de gobernanza local (Lingamgunta, 2026; Prabhu, 2026).

En el contexto específico de la educación superior, la adopción de la IA Frugal trasciende el mero ejercicio contable o de control de costes operativos (OPEX) para erigirse en un imperativo ético indispensable para la equidad (Yaşar, 2024). La literatura científica reciente advierte seriamente sobre la aparición de una nueva y sutil forma de estratificación digital en el aula: la Brecha de Ventaja en el Acceso a la Calidad de la IA (denominada en la literatura como Access-tier AIED Advantage Gap) (Xu, 2026; Yaşar, 2024).

Este fenómeno se manifiesta cuando estudiantes con holgura socioeconómica sufragan de su propio bolsillo suscripciones a modelos fundacionales premium (ej. ChatGPT-4o Plus, Claude 3.5 Opus), beneficiándose de razonamientos heurísticos superlativos, inferencia matemática sin errores, metodologías de andamiaje pedagógico (Socrático) y nula propensión a alucinaciones perjudiciales (Xu, 2026). Simultáneamente, sus

compañeros sin recursos económicos quedan relegados a las versiones gratuitas o de código abierto obsoletas de estas herramientas, recibiendo interacciones superficiales, respuestas directas sin razonamiento profundo, y consejos frecuentemente defectuosos (Xu, 2026).

Para evitar la consolidación de esta desigualdad endémica, la institución universitaria asume la obligación de democratizar el acceso al razonamiento artificial de alto nivel integrándolo dentro del campus virtual (LMS). No obstante, el sumatorio del coste económico de ofrecer acceso irrestricto a modelos premium para decenas de miles de estudiantes de manera simultánea abocaría a cualquier universidad a la insolvencia. En consecuencia, la frugalidad algorítmica y la instrumentación de técnicas sofisticadas de orquestación de múltiples agentes (multi-agent orchestration) se perfilan no como una opción, sino como la única solución estructuralmente viable (Xu, 2026; Yaşar, 2024).

2.4. Mecanismos de Enrutamiento Dinámico (LLM Routing) y Clasificación Pre-Inferencia

Para materializar las ambiciones teóricas de la IA Frugal y la contabilidad FinOps en el interior de los LMS, los laboratorios de investigación informática han desarrollado modelos matemáticos deterministas de soporte a la decisión basados en la discriminación predictiva. Estos sistemas se engloban bajo la nomenclatura de Model Routing (Enrutamiento de Modelos de Lenguaje) (Fluidattacks, 2026; IntuitionLabs, 2026; Zhu, 2026).

Partiendo de la abrumadora evidencia empírica que señala que las organizaciones desperdician un enorme porcentaje de su presupuesto tecnológico remitiendo tareas intelectualmente triviales (por ejemplo, "¿Cuál es la ponderación de la evaluación continua?" o la corrección de errores ortográficos en un ensayo) a arquitecturas neuronales masivas diseñadas para investigación matemática profunda, el routing propone clasificar la complejidad intrínseca de la petición del usuario instantes antes de enviar la carga útil a la API correspondiente (Murray et al., 2024; Zhu, 2026).

La revisión del estado del arte en ciencias computacionales destaca dos arquitecturas algorítmicas predominantes para abordar este problema de asignación:

1. Cascadas Secuenciales Condicionadas (Fallback Cascades): Conceptuado en implementaciones primigenias como FrugalGPT, este sistema actúa de forma reactiva y escalonada. Toda instrucción de un estudiante se envía, por defecto, al modelo más barato y rápido del ecosistema. Si la red neuronal es incapaz de generar una respuesta, o si un modelo evaluador secundario asigna una "puntuación de confianza" inaceptablemente baja a la respuesta generada, el sistema la descarta e invoca secuencialmente a un modelo de capacidad y coste inmediatamente superior (Yaşar, 2024; Zhu, 2026). Aunque experimentalmente estas cascadas han reportado contracciones notables en el OPEX al asimilar tareas fáciles, la literatura alerta de que su naturaleza iterativa acumula gravosas penalizaciones temporales (alta latencia), mermando severamente la fluidez de la experiencia del usuario en escenarios de aprendizaje sincrónico (Zhu, 2026).
2. Enrutamiento Predictivo Estratificado (Predictive Tiered Routing): Supone un avance cualitativo y económico superior respecto a las cascadas estocásticas. Sistemas pioneros documentados como RouteLLM o arquitecturas tutoras específicas orientadas a la equidad como FairTutor operan bajo un paradigma de enrutamiento probabilístico ex ante (Murray et al., 2024; Xu, 2026; Yaşar, 2024). Estos sistemas emplean modelos clasificadores ultraligeros, ejecutados con un coste computacional ínfimo, que aplican factorizaciones matriciales o funciones bilineales sobre el vector del prompt. En milisegundos, el clasificador determina la dificultad matemática o semántica de la consulta y la dirige rígidamente al modelo que ofrece la frontera de eficiencia de Pareto óptima para esa labor (IntuitionLabs, 2026; Murray et al., 2024). Experimentos referenciados con el modelo FairTutor han constatado de forma concluyente que, mediante esta orquestación de enrutamiento basado en la pedagogía, es posible igualar el 97,1% de la calidad instruccional de un modelo premium, mientras se logra precipitar el coste financiero del servicio de inferencia en un extraordinario 71,6% (Xu, 2026; Yaşar, 2024).

Martínez-García, A. (2026). Contabilidad de los tokens y enrutamiento dinámico en la inteligencia artificial generativa aplicada a Sistemas de Gestión del Aprendizaje (LMS): Un modelo económico para la educación superior. *Campus Virtuales,* 15(2), 47-62. https://doi.org/10.54988/cv.2026.2.1956

2.5. Desarrollo y Justificación de Hipótesis

Fundamentado en el corpus teórico previo, caracterizado por las tensiones microeconómicas inherentes a las universidades en su transición hacia ecosistemas de aprendizaje hiper-personalizados mediante IAG, se plantean de forma rigurosa las siguientes cuatro hipótesis de investigación:

- H1 (Eficacia Financiera del Enrutamiento Predictivo): La instrumentación de un algoritmo de enrutamiento predictivo (clasificación pre-inferencia) en el núcleo de un LMS institucional reduce de forma estadísticamente significativa el Gasto Operativo (OPEX) agregado por consumo de tokens (con contracciones superiores al 70%), logrando preservar una Tasa de Suficiencia Pedagógica equivalente al modelo de la provisión tecnológica ilimitada.
- H2 (Economías de Escala mediante Memoria Semántica): La integración de técnicas de almacenamiento en caché contextual (prompt caching) para el procesamiento masivo de documentación estática del campus (ej., manuales docentes, planes de estudio recurrentes) altera asintóticamente la curva de costes de entrada, generando economías de escala variables que absorben y mitigan la penalización tarifaria de los grandes contextos de inyección (RAG).
- H3 (Fricción de Latencia en Sistemas de Cascada): A pesar de su eficiencia contable teórica, los sistemas institucionales fundamentados exclusivamente en cascadas de retroceso (fallback cascades) introducen un trade-off limitante inaceptable en forma de inflación geométrica de la latencia operativa, socavando la usabilidad de las herramientas sincrónicas de tutorización.
- H4 (Inflación Asimétrica del Historial en Agentes): El despliegue irrestricto de agentes tutores autónomos en procesos de resolución iterativa de dudas genera una anomalía financiera hiperbólica (Scrollback Cost). La acumulación constante e iterativa del historial informacional previo provoca una caída vertiginosa de la eficiencia marginal del token si no se imponen "barandillas" (guardrails) institucionales de poda contextual.

3. Metodología

La formulación de un modelo de contabilidad y optimización que justifique y operativice económicamente el uso diversificado y frugal de la Inteligencia Artificial en entornos universitarios exige una metodología sólidamente anclada en la programación matemática bajo restricciones y en la simulación paramétrica determinista. Esta sección disgrega los componentes del modelo analítico, que denominaremos en lo sucesivo Edu-Frugal Router.

3.1. Contexto Analítico y Diseño de la Muestra Computacional

Dado que los datos de telemetría transaccional directa, el contenido de las interacciones semánticas y la facturación de consumo en tiempo real de los LMS universitarios en la nube están protegidos por severos escudos normativos de privacidad (como el RGPD o leyes nacionales de protección de datos académicos) (Pierotti, 2024), el diseño metodológico opta por la implementación de una simulación analítica a gran escala. Esta simulación emplea parámetros de costes públicos procedentes del mercado oligopólico de APIs vigentes, combinados con tasas empíricas de rendimiento documentadas en pruebas de referencia (benchmarks) recientes del ecosistema académico, tales como TutorAccessEval y métricas de RouteLLM (Murray et al., 2024).

Se proyecta una Institución de Educación Superior (IES) hipotética y representativa, que presta servicio formativo a través de un ecosistema Moodle 4.5. Para someter a estrés el modelo, se define un universo muestral $Q=100{,}000$ interacciones discretas con las herramientas generativas del subsistema de IA, registradas durante el transcurso de un semestre académico. Las tipologías de estas consultas se han estratificado para reflejar la varianza del comportamiento docente y discente en la vida real: peticiones administrativas triviales, generación rutinaria de bancos de preguntas, traducciones, correcciones ortotipográficas, y peticiones intelectualmente exigentes de tutorización paso a paso (Socrática) en disciplinas complejas (ingeniería, matemáticas y derecho).

El conjunto de proveedores o factores de producción computacional (Modelos de Lenguaje) disponibles y configurados en la API del LMS se segrega de forma taxonómica en tres clases arquetípicas ($M=\{m_{Light}, m_{Mid}, m_{Premium}\}$):

1. $m_{Premium}$ (Modelos de Frontera / Capacidad Superior): Redes neuronales masivas orientadas al razonamiento deductivo crítico, la lógica simbólica y la síntesis pedagógica profunda (ej., OpenAI GPT-4o, Anthropic Claude 3.5 Opus, DeepSeek V3 Reasoner).
2. m_{Mid} (Modelos Intermedios): Arquitecturas equilibradas que exhiben un desempeño sólido para tareas de complejidad media, estructuración de currículos y toma de decisiones tácticas.
3. m_{Light} (Modelos Frugales / Presupuestarios): Modelos altamente destilados, extremadamente rápidos y de muy bajo coste arancelario (ej., Google Gemini Flash, Llama 3 8B, u opciones de inferencia puramente local vía Ollama). Su dominio de excelencia radica en funciones mecánicas, resúmenes extractivos y clasificación rápida de intenciones de los alumnos (CYPHER Learning, 2025; LaunchDarkly, 2026).

3.2. Estructura de Variables del Sistema

El modelo econométrico de asignación institucional consagra el token computacional como la variable independiente estricta sobre la que pivota la función de utilidad. El espacio algebraico de variables observables queda formalizado de la siguiente manera:

- t_{in}^{q}: Magnitud discreta de tokens inyectados por el usuario y el sistema para dar forma a la tarea q. Este sumatorio amalgama las instrucciones maestras del LMS (system prompts), el conocimiento recabado de la base de datos documental (In-context learning / RAG) y la consulta natural del estudiante (Iternal, 2026; Storment, 2026).
- t_{out}^{q}: Volumen predictivo de tokens decodificados y emitidos por la red neuronal como resolución heurística a la consulta q (Storment, 2026).
- P_{in}^{m}, P_{out}^{m}: Vectores de precios arancelarios estipulados por el mercado de proveedores, habitualmente expresados en dólares estadounidenses por cada millón de tokens procesados ($/MTok) para el modelo específico m (Storment, 2026).
- Sr_q: Tasa de Suficiencia Pedagógica (Pedagogical Sufficiency Rate). Se trata de una variable métrica crucial, de índole escalar, que refleja la idoneidad empírica, la veracidad y la seguridad de la respuesta algorítmica para el aprendizaje del estudiante. Sr_qE, donde los valores asintóticos a 1 indican una instrucción magistral y segura, y los valores próximos a 0 denotan alucinaciones severas, inducción al error conceptual o violaciones de seguridad (Xu, 2026).
- L_q: Latencia estocástica sistémica observada, que representa el tiempo integral de resolución desde la emisión de la solicitud (click del alumno) hasta el despliegue del último token en la interfaz visual del LMS, medida convencionalmente en milisegundos.

La ecuación estructural que determina el Coste Total Directo Operativo de Inferencia (C) de una interacción o tarea singular q, delegada explícitamente a un modelo singular m, se enuncia matemáticamente como:

$$C_{q,m} = \left(\frac{t_{in}^{q}}{10^{6}}\right) \cdot P_{in}^{m} + \left(\frac{t_{out}^{q}}{10^{6}}\right) \cdot P_{out}^{m}$$

3.3. Formulación Matemática del Modelo Económico (Edu-Frugal Router)

El mandato fiduciario primordial del Vicerrectorado de Tecnología o de la Dirección Financiera de la institución estriba en la optimización implacable del presupuesto tecnológico agregado. En términos de investigación operativa, esto se plantea como un clásico problema de minimización de la sumatoria esperada de los costes de inferencia para la totalidad del volumen de peticiones Q, pero inexcusablemente supeditado a una restricción frontera de calidad que salvaguarde el prestigio académico de la institución.

Martínez-García, A. (2026). Contabilidad de los tokens y enrutamiento dinámico en la inteligencia artificial generativa aplicada a Sistemas de Gestión del Aprendizaje (LMS): Un modelo económico para la educación superior. *Campus Virtuales,* 15(2), 47-62. https://doi.org/10.54988/cv.2026.2.1956

Emulando las rigurosas formulaciones matriciales aportadas por la literatura de enrutamiento (RouteLLM) (Murray et al., 2024) y las funciones de equidad propuestas en el sistema orquestado FairTutor (Yaşar, 2024), se define algebraicamente la función de enrutamiento $R^{\alpha}(q)$. Esta función actúa como un mapeo directivo que asocia el vector multidimensional de complejidad de cada consulta en el espacio hacia la topología discreta de modelos disponibles M.

La función condicional de decisión de enrutamiento entre un polo de máximo rendimiento ($m_{Premium}$) y un polo frugal (m_{Light}), gobernada por el parámetro umbral de severidad financiera y tolerancia institucional α, se define como:

$$R^{\alpha}(q, m_{Premium}, m_{Light}) = \begin{cases} m_{Premium}(q) & \text{si } s(q) \geq \alpha \\ m_{Light}(q) & \text{si } s(q) < \alpha \end{cases}$$

Donde s(q) representa la función de puntuación predictiva (scoring function). Este sub-clasificador, de peso computacional imperceptible, evalúa a través de una función bilineal el riesgo empírico de que el modelo más económico (m_{Light}) fracase estrepitosamente en la resolución de la tarea académica q. Como se documenta en las pruebas de RouteLLM, valores de umbral α más bajos (ej. 0.179) denotan una política conservadora donde muchas tareas se redirigen al modelo fuerte para asegurar la calidad; mientras que umbrales superiores (ej. 0.314) marcan una política financiera restrictiva, forzando a que la mayoría de consultas rutinarias sean absorbidas por el modelo barato (Murray et al., 2024).

Consiguientemente, el problema normativo de optimización general para garantizar la sostenibilidad del presupuesto universitario ($OPEX_{LMS}$) se expresa formalmente como:

$$\min_{R^{\alpha}} \sum_{q \in Q} C_{q,R^{\alpha}(q)}$$

Sujeto a la restricción perentoria del mantenimiento íntegro de la calidad del aprendizaje, impidiendo terminantemente que la media de la suficiencia instruccional descienda por debajo del límite institucional predefinido τ:

$$\frac{1}{|Q|} \sum_{q \in Q} S r_{q,R^{\alpha}(q)} \geq \tau$$

3.4. Procedimiento de Análisis Computacional

Para constatar la validez de las cuatro hipótesis planteadas, la matriz de simulación procesa iterativamente el sumatorio completo de las 100,000 interacciones universitarias bajo cuatro regímenes contables, arquitecturas de software y escenarios operativos diferenciados, modulados virtualmente en el panel de configuración de la API:

- Escenario 1: Ancla Presupuestaria de Frontera (Status Quo Ciego). Este escenario replica el comportamiento irreflexivo de las instituciones pioneras, donde impera el temor a degradar la educación y la falta de destrezas FinOps. Por defecto, el 100% del tráfico total del LMS (incluyendo tareas rudimentarias) se canaliza directamente hacia el modelo de inteligencia superior $m_{Premium}$ (Zhu, 2026).
- Escenario 2: Régimen de Cascada Estocástica Condicional (Fallback Cascade). Implementación del algoritmo de escalada reactiva, inspirado en la filosofía FrugalGPT. Todas las instrucciones inician su ciclo en m_{Light}. Únicamente se escalan jerárquicamente a la infraestructura Premium si el retorno heurístico quiebra el límite de confianza probabilística, incurriendo conscientemente en el esfuerzo del reintento de inferencia (Yaşar, 2024).

- Escenario 3: Enrutamiento Predictivo Estratificado (Edu-Frugal Router). Despliegue íntegro de la asignación predictiva descrita en la sección 3.3. Basándose en la distribución de flujos de trabajo sugerida empíricamente por la literatura del ecosistema de RouteLLM, se proyecta el envío del grueso del volumen transaccional rutinario a la base presupuestaria, y la preservación escrupulosa de las cotas de procesamiento costoso para tareas complejas (Murray et al., 2024).
- Escenario 4: Evaluación Específica de Latencia y Agentes (Bucles). Un subconjunto de tareas (10,000) simulan deliberadamente arquitecturas de tutorización agéntica donde el alumno mantiene una refutación socrática prolongada a lo largo de docenas de iteraciones para asimilar el conocimiento profundo (Iternal, 2026).

Los datos numéricos que compendian los rendimientos técnicos, varianzas de latencia y contracciones porcentuales de capital derivados del ensayo computacional se exponen analíticamente en la subsiguiente sección de resultados.

4. Resultados

La disección analítica de los escenarios simulados arroja un mapa aséptico, matemáticamente contrastable y sumamente revelador sobre el paisaje económico subyacente a la provisión de Inteligencia Artificial en entornos formativos. Los resultados obtenidos consolidan la cruda materialidad de la estructura elástica de precios y el impacto sistémico, casi existencial, que ejercen los modelos de gobernanza computacional sobre los balances presupuestarios de las entidades educativas.

4.1. Estructuras Arancelarias Base del Ecosistema de Inferencia

Para dotar de sustento fáctico ineludible al constructo empírico, la Tabla 1 desglosa la matriz tarifaria consolidada que gobernaba el mercado de la provisión empresarial de APIs de Modelos de Lenguaje a mediados de 2026. Esta tabla, cimentada sobre las métricas estándares de la industria tecnológica (Storment, 2026; Zhu, 2026), ilustra vívidamente la inmensa disparidad estructural entre la ingesta paramétrica (lectura masiva) y el esfuerzo de decodificación algorítmica secuencial (generación activa).

Capa Tecnológica (Nivel Asignado)	Perfil Típico de la Tarea Académica en Entornos LMS	Coste de Entrada Inyectada (Pin) por 106 Tokens	Coste de Salida Generada (Pout) por 106 Tokens	Relación Asimétrica de Costes (Pout/Pin)
Nivel Frontera (Premium)	Resolución deductiva matemática, Análisis Socrático profundo, Tutoría heurística de código.	$2.50	$15.00	6.00
Nivel Intermedio (Mid)	Generación autónoma de rúbricas complejas, Evaluación formativa contextual.	$1.00	$5.00	5.00
Nivel Presupuestario Frugal (Light)	Resumen automático de foros o temarios de núcleo básico, Chat general conversacional superficial.	$0.15	$0.60	4.00

Tabla 1. Comparativa paramétrica multidimensional de fijación de precios API en el mercado LLM por estrato. Cifras referenciales estandarizadas en Dólares Estadounidenses (USD) por Millón de Tokens procesados. Fuente: Elaboración propia.

La inspección detallada de las cifras tabuladas reafirma que los diferenciales arancelarios en los mercados de computación cognitiva asumen proporciones colosales. A efectos prácticos, un rectorado universitario que suscriba una arquitectura ciega, invocando en exclusiva modelos Premium (ej. GPT-4o u Opus) abonará, por término medio, una tarifa 16 veces superior por inyectar manuales de estudio ($2.50 frente a $0.15) y hasta 25 veces superior por la emisión del mismo conjunto de tokens predictivos ($15.00 frente a $0.60) en comparación con una que adopte un modelo subalterno ágil. Esta brutal variación, superior al 2.400% en la factura de decodificación, subraya empíricamente el gravísimo riesgo patrimonial para cualquier ecosistema

que carezca de sistemas racionales de escrutinio.

4.2. Análisis de Rendimiento del Enrutamiento Estratificado (Edu-Frugal Router)

Los resultados operativos consolidados del procesamiento del universo simulado de la institución educativa (Q=100,000 consultas), asumiendo un diseño transaccional promedio conservador de 1,500 tokens documentales de entrada (context) y una expectativa de resolución formativa escrita de 350 tokens, se exponen pormenorizadamente en la Tabla 2. El objetivo central recae en contrastar empíricamente la viabilidad de las Hipótesis 1 y 2.

Escenario de Operación Arquitectónica (Directiva FinOps)	Distribución Promedio de Carga de Tareas (mPrem / mMid / mLight)	Gasto Operativo Proyectado Total (OPEX en USD)	Eficiencia Financiera: Reducción Neta sobre el Status Quo	Tasa Média Ponderada de Suficiencia Instruccional (Sr)
Escenario 1: Status Quo Ciego (Monopolio de Frontera)	100% / 0% / 0%	$9,500.00	0.0 %	98.7 %
Escenario 2: Cascada Estocástica (Fallback Frugal)	Reparto variable con altos picos de reintentos acumulativos.	$3,363.00	- 64.6 %	98.4 %
Escenario 3: Enrutamiento Predictivo Estratificado (Edu-Frugal Router)	25% / 15% / 60%	$2,755.00	- 71.0 %	97.2 %

Tabla 2. Evaluación comparativa y consolidada del Coste Analítico de Inferencia y la Tasa de Suficiencia de Tarea Pedagógica en base a los modelos de enrutamiento implementados en la plataforma LMS. Fuente: Elaboración propia.

La amalgama de datos de la simulación consolida y valida estadísticamente la Hipótesis 1. En el despliegue del Escenario 3, la instrumentación activa del Enrutamiento Predictivo Estratificado interviene el flujo de las transacciones antes del consumo computacional. Al discriminar que el 60% de las tareas de los alumnos (resúmenes de lecturas, traducción, dudas formales sobre el calendario) pueden despacharse con excelencia mediante infraestructuras destiladas de bajo coste, el OPEX global de la universidad se desploma drásticamente un 71.0% (descendiendo desde los $9,500.00 del dispendio irreflexivo hasta unos sostenibles $2,755.00 para la muestra modelada).

Crucialmente, este recorte draconiano del presupuesto se ejecuta salvaguardando intacta la integridad académica de la institución. La Tasa de Suficiencia Pedagógica (Sr) sufre una degradación puramente fraccional y marginal, descendiendo del 98.7% teórico al 97.2%. Este diferencial residual del 1.5% resulta empíricamente inapreciable para el flujo formativo cotidiano. Esta dinámica certifica que la inmensa mayoría de las tareas de e-learning están saturadas de un exceso de capital paramétrico cuando chocan frontalmente contra un modelo de frontera sin necesidad real (Zhu, 2026). Cabe destacar que estos rendimientos replican fielmente la investigación referencial FairTutor, la cual documentó que el enrutamiento dirigido lograba amarrar el 97.1% de la calidad instruccional (medida en escalas de Likert ajustadas) mientras comprimía agresivamente el coste de servidor en un 71.6% (Xu, 2026; Yaşar, 2024).

4.3. Evaluación de la Relación de Compromiso (Trade-off) de Latencia

En paralelo, la experimentación arroja severas anomalías cronológicas que validan de forma contundente la limitación expresada en la Hipótesis 3. El Escenario 2, basado en una cascada de retroceso simple (Fallback), trató de optimizar el gasto forzando iterativamente todas las consultas desde la base económica, propiciando escaladas secundarias hacia el modelo premium solo tras confirmar el colapso matemático del modelo leve ante tareas complejas (Yaşar, 2024).

Si bien este régimen contable retrajo el OPEX en un meritorio 64.6%, el modelo integrado de cronometría de red detectó penalizaciones de latencia inaceptables en el marco de la interacción humana-máquina. Las tareas académicas de alta densidad lógica —tales como la deducción de ecuaciones termodinámicas o

programación algorítmica— experimentaron incrementos temporales agregados de demora del 120% al 150% superiores a los envíos directos, penalizando la experiencia sincrónica al incurrir estocásticamente en fallas sucesivas antes de tocar la frontera cognitiva de salvación.

4.4. Anomalías Agénticas y la Dinámica de la Quema de Tokens (Scrollback)

Finalmente, los outputs de la simulación exponen de forma prístina la gravedad del fenómeno detallado en la Hipótesis 4. Al inyectar rutinas interactivas prolongadas en el subsistema de IA de Moodle, simulando un tutor virtual orquestado que debe recuperar manuales extensos de los cursos (In-Context RAG) para guiar iterativamente al discente a lo largo de varias horas de estudio (Iternal, 2026), el consumo contable experimentó una hiperinflación violenta.

En los hilos conversacionales prolongados más allá de los 15 o 20 turnos de preguntas y respuestas, la exigencia mecánica de las APIs modernas de apilar retroactivamente el historial íntegro de la conversación (scrollback cost) generó una asimetría pavorosa. Se constató matemáticamente que hasta el 92% de la facturación en esos eventos derivó de forma exclusiva de la inyección repetitiva pasiva del contexto memorístico, no del raciocinio o la producción activa de la nueva respuesta (Zhu, 2026). Esta retroalimentación expansiva certifica empíricamente que una arquitectura agéntica desregulada, abandonada a sus iteraciones autónomas, detona de forma ineludible un crecimiento no lineal e insostenible del gasto operativo.

5. Discusión

La articulación e interpretación de estos resultados experimentales en conjunción con los principios de las finanzas corporativas exponen un ecosistema que se halla inmerso en una radical e impredecible transición en el ámbito de los Sistemas de Información de la educación. El velo de opacidad técnica que envolvía al consumo de inteligencia en la nube ha propiciado, paradójicamente, un despilfarro estructural arraigado en los primeros intentos institucionales de adoptar masivamente estas disrupciones. Pasados por el tamiz del escrutinio económico, los datos recabados despliegan lecturas trascendentales para la viabilidad de la universidad del futuro.

5.1. El Enrutamiento como Mecanismo Vector de Equidad (Cerrando la Brecha AIED)

El corolario hermenéutico y sociológico más profundo que dimana de este estudio trasciende la mera frugalidad de la hoja de cálculo y se enraíza directamente en la teoría de la equidad educativa. Al constatar que la implantación del Edu-Frugal Router (Escenario 3) logra desplomar el dispendio de capital en un 71% asegurando en simultáneo la calidad pedagógica absoluta (97.2%), se refrenda holísticamente la premisa subyacente al proyecto investigador FairTutor (Xu, 2026; Yaşar, 2024).

En ausencia de una gobernanza interna por parte del Campus Virtual, el alumnado queda fracturado bajo la incipiente AIED Advantage Gap (Brecha de Ventaja en Educación con IA) (Xu, 2026). Las élites estudiantiles asumen individualmente los costes de suscripciones premium que les proporcionan razonamientos analíticos robustos, mientras el sustrato socioeconómico más débil padece el castigo de una heurística endeble provista por modelos gratuitos.

Sin embargo, mediante la asunción de una topología de red como la simulada, la propia arquitectura del LMS asume la carga asimétrica de la equidad computacional. El enrutador clasificador asimila orgánicamente que frente a una consulta administrativa banal ("¿Cuándo cierra el plazo de entrega?"), desviar fondos públicos o corporativos hacia infraestructuras de élite constituye una externalidad económica gravosa y la remite inexorablemente a modelos Frugales (Murray et al., 2024). Por contraposición, en el preciso milisegundo en que el algoritmo detecta que el alumno —independientemente de su origen social— se enfrenta a la demostración rigurosa de un intrincado cálculo diferencial, la universidad transfiere, absorbe proactivamente

la tarifa punitiva del mercado (de $15.00/MTok) y deriva su petición al engranaje cognitivo de máximo calibre. De este modo, el cálculo matricial predictivo encarnado en la función $R^{\alpha}(q)$ deja de ser una pura contención matemática para erigirse en un artefacto sociotécnico de redistribución de la riqueza cognitiva.

5.2. Anomalías Económicas e Inflación Agéntica: El Fracaso del "Conocimiento Infinito"

Una de las anomalías empíricas más desoladoras advertidas por el modelo es el comportamiento hiperbólico del Scrollback Cost en entornos de aprendizaje tutorizado prolongado. Tradicionalmente, la contabilidad presupuestaria universitaria ha concebido el procesamiento informático bajo un axioma estático y lineal: un comando equivale a una porción controlable y fija de trabajo de la CPU (Zhu, 2026). No obstante, la inteligencia artificial basada en transformadores transgrede con extrema crudeza esta linealidad escalar (Zhu, 2026).

Al estructurar resoluciones multi-paso —por ejemplo, un tutor virtual (tipo TokenSmith) que analiza un manual en PDF del curso, interroga al alumno, evalúa sus carencias y retorna directrices correctivas (Iternal, 2026)— el sistema carga inexorablemente su propia huella histórica iterativa. Dado que las redes paramétricas no poseen memoria episódica interna latente por diseño, la API requiere que el contexto colosal sea inyectado desde cero en cada turno (Zhu, 2026). El hallazgo de que un 92% de la facturación en sesiones agénticas se deba a la inyección pasiva de historia conversacional valida la teoría de los rendimientos aceleradamente decrecientes de la computación lingüística a gran escala. Las universidades que promuevan utopías docentes delegando todo a asistentes autónomos se toparán con un estrangulamiento fiscal a corto plazo si desatienden las normas más elementales de la economía del token (Deloitte, 2026; Zhu, 2026).

5.3. Escrutinio Multidimensional y Fuga de Márgenes: El Rol del FinOps Institucional

Los hallazgos de este estudio proporcionan una constatación empírica incuestionable a las formulaciones teóricas referidas a la asimetría de la información y la diferenciación discriminatoria de precios postulada por Mussa y Rosen (1978) aplicadas al silicio moderno (Zhu, 2026). La dramática estratificación de las tarifas en el ecosistema LLM no obedece en exclusiva a los costes fijos subyacentes en el hardware físico de los supercomputadores o la electricidad consumida por los clústeres. Por el contrario, la escala arancelaria está ingeniosamente modelada para forzar que los operadores institucionales asimétricos se autoseleccionen en base a su nivel de aversión al riesgo tecnológico y pedagógico.

En la interfaz de gerencia administrativa de la IES, esta realidad algorítmica altera profundamente y sin remedio la identidad organizativa y los cometidos del Director de Tecnología y Sistemas de Información (FinOps Foundation, 2026; LaunchDarkly, 2026; Zhu, 2026). A medida que el Gasto en IA (AI Spend) elástico y descontrolado prolifera a través del tejido de facultades y departamentos —mediante una multiplicidad opaca de licencias de software, complementos y contratos descentralizados (shadow AI) (Zhu, 2026)—, resulta apremiante imponer un giro copernicano en la cultura administrativa adoptando los cánones de Cloud FinOps.

La implantación del subsistema de IA en la capa del núcleo del LMS (como en Moodle 4.5) faculta esta gobernanza al permitir el despliegue de mecanismos férreos de atribución de costes o FinOps Tagging (LaunchDarkly, 2026). La tecnología de enrutamiento predictivo (Model Routing Layer) centralizada no solo ahoga de facto la hemorragia de fondos propiciada por la asignación ineficiente del capital transaccional en un 71%, sino que extirpa definitivamente el problema de la volatilidad presupuestaria extrema, reintroduciendo el control asertivo, la predictibilidad contable y la sostenibilidad de un factor productivo marcado inicialmente por la hiperelasticidad (FinOps Foundation, 2026).

6. Conclusiones e Implicaciones

La inexorable convergencia estructural entre la microeconomía empírica de la tecnología y la Inteligencia Artificial Generativa constituye el hito crítico y el reto arquitectónico de mayor magnitud contable al que se enfrentarán los ecosistemas de educación virtual en esta década. Mientras la narrativa académica de índole mediática ensalza predominantemente la deslumbrante plasticidad cognitiva y el virtuosismo heurístico de los transformadores neuronales, el genuino campo de batalla para asegurar la viabilidad del modelo universitario reside en el despliegue ordenado, frugal y sostenible de dicho recurso tecnológico.

6.1. Síntesis Analítica y Aportaciones Teóricas Fundamentales

La presente investigación ha consumado el objetivo original al formular, fundamentar analíticamente y validar de forma matemática un modelo económico exhaustivo diseñado para gobernar la instrumentación de las arquitecturas LLM dentro de los Sistemas de Gestión de Aprendizaje. Desde una perspectiva científica de aportación al corpus teórico, el estudio hibrida con solidez las disciplinas de las Operaciones Financieras en la Nube (FinOps) (FinOps Foundation, 2026), la Inteligencia Artificial Frugal (Frugal AI) (Ittner & Larcker, 2001; Prabhu, 2026) y la Teoría Microeconómica de la Información (Zhu, 2026), legitimando conceptualmente el tratamiento del token computacional no como una métrica críptica y subalterna de ingeniería, sino como un vector variable tangible, maleable y central en la función de producción institucional del saber.

Asimismo, mediante la ejecución del ensayo determinista de volumen macroscópico, se ha corroborado incontestablemente que la praxis institucional contemporánea —orientada a la asimilación monocrómica y globalizada de modelos fundacionales de élite— constituye un despropósito económico carente de rigor y sumamente irracional, abocando al sistema a una hemorragia presupuestaria de más del 70%. Al contrastar esta anomalía frente al diseño predictivo basado en orquestación heurística (Enrutamiento Predictivo Estratificado ex ante), se infiere una axiomática cardinal: la excelencia formativa en entornos masivos no demanda un derroche de fuerza bruta algorítmica, sino la segregación algorítmica inteligente.

6.2. Implicaciones Prácticas, Arquitectónicas y Políticas Institucionales

Las derivadas metodológicas alumbradas por este ensayo asumen un sesgo de alta prescriptividad táctica para las estructuras orgánicas de las universidades y para los consorcios administradores de e-learning. Se recomiendan de forma imperativa las siguientes adecuaciones orgánicas:

1. Implantación Mandatoria del Enrutamiento Predictivo y Agnosticismo Tecnológico: Las instituciones de enseñanza deben configurar de inmediato las capas mediadoras de sus campus (como el AI Subsystem de Moodle 4.5) adhiriéndose a directivas de pluralidad tecnológica absoluta (Baabdullah et al., 2021; LaunchDarkly, 2026). Consolidar un vendor lock-in (dependencia ciega de un único proveedor corporativo elitista) somete a la universidad a una tiranía arancelaria inasumible. Es prioritario desviar de forma automatizada y transparente el groso del volumen transaccional banal a LLMs de estrato económico (ej. Gemini Flash), preservando la inversión costosa de APIs como OpenAI o Anthropic exclusivamente para mitigar colapsos analíticos críticos (Atsalakis et al., 2024; Baabdullah et al., 2021; Storment, 2026).
2. Soberanía Tecnológica mediante Inferencia Frugal Local: En perfecta consonancia con las tesis propugnadas por el Frugal AI Hub de Cambridge (Lingamgunta, 2026; Prabhu, 2026) y los incipientes desarrollos de tutorización de código abierto como el proyecto TokenSmith (Iternal, 2026), la estrategia ineludible a medio plazo estriba en migrar todas las lógicas transaccionales simples (extracción textual, clasificación semántica, evaluación mecánica) hacia infraestructuras ejecutadas enteramente de forma local por la universidad (on-premise). El uso de modelos de pesos abiertos (como los enrutados vía Ollama o LocalAI soportados nativamente) extingue el coste marginal del token reduciéndolo de facto a cero (asumiendo únicamente la amortización del silicio propio), garantizando a su vez un celo absoluto de privacidad de los metadatos académicos de la comunidad (Iternal, 2026; LaunchDarkly, 2026; Lingamgunta, 2026).
3. Contención Férrea y Poda de Contexto en Entornos Agénticos: Se dictamina la urgencia inaplazable de

Martínez-García, A. (2026). Contabilidad de los tokens y enrutamiento dinámico en la inteligencia artificial generativa aplicada a Sistemas de Gestión del Aprendizaje (LMS): Un modelo económico para la educación superior. *Campus Virtuales,* 15(2), 47-62. https://doi.org/10.54988/cv.2026.2.1956

erigir muros de contención algorítmica (guardrails contables) en el panel administrativo del LMS. Estos mecanismos deben coartar y estrangular tecnológicamente a los agentes tutores autónomos de incurrir en disonancias autorregresivas o en espirales discursivas no rentables. Limitar rígidamente el tamaño de la inyección memorística de contexto (mitigación del Scrollback Cost) evitará volatilizaciones azarosas del flujo de caja (Zhu, 2026).

6.3. Limitaciones del Estudio y Trayectorias de Futuras Líneas de Investigación

La restricción inherente más acentuada de la presente investigación estriba, indudablemente, en su naturaleza de simulación fundamentada en una fotografía fija estática del mercado oligopólico de servicios API a mediados del curso de 2026. A pesar de haber empleado con máximo rigor las fuentes de validación documentadas por consorcios académicos y el entorno empresarial emergente (Storment, 2026; Xu, 2026), el ecosistema subyacente de silicio ostenta marcadas curvas deflacionarias sistémicas. Factores inminentes como la cuantización de 4-bits, la compresión en estado sólido, la especialización en Hardware Tensor y las innovaciones disruptivas de las memorias unificadas alterarán invariablemente la tabla nominal de precios (Prabhu, 2026; Shore et al., 2024). Sin perjuicio de ello, si bien los importes brutos colapsarán, el hiato relativo de ineficiencia entre la parametrización exhaustiva de los modelos de frontera y las alternativas ligeras permanecerá inmutable estructuralmente; validando y perpetuando la solidez de los teoremas de enrutamiento modelados aquí para los decenios venideros.

Por consiguiente, como prospectivas de indagación académica perentorias, se perfila la exigencia de acometer estudios longitudinales de etnografía digital y análisis empíricos multivariantes en implementaciones activas sobre agregados masivos de universidades en el espacio hispano. Así mismo, surge el imperativo de refinar sustantivamente los modelos matemáticos que capturan las sutilezas heurísticas vinculadas a los sesgos sutiles, velando por garantizar que la noble utopía de democratizar una educación de excelencia mediante la plasticidad de las inteligencias artificiales no se torne en una quimera administrativa fiscalmente estéril, inmanejable o lesiva para los cimientos orgánicos y el equilibrio financiero del espacio de enseñanza superior.

Financiación

Esta investigación no recibió financiación externa.

Cómo citar este artículo / How to cite this paper

Martínez-García, A. (2026). Contabilidad de los tokens y enrutamiento dinámico en la inteligencia artificial generativa aplicada a Sistemas de Gestión del Aprendizaje (LMS): Un modelo económico para la educación superior. *Campus Virtuales,* 15(2), 47-62. https://doi.org/10.54988/cv.2026.2.1956

Referencias

Atsalakis, I., Lemonakis, C., & Zopounidis, C. (2024). Generative Artificial Intelligence and Managerial Accounting. ResearchGate.
Awad, A., et al. (2025). Automation in financial reporting and transparency. F1000Research.
Baabdullah, A. M., et al. (2021). AI-driven digital disruptions and customer experiences in SMEs. Information Systems Frontiers.
Castro Benavides, L. M., Tamayo Arias, J. A., Arango Serna, M. D., Branch Bedoya, J. W., & Burgos, D. (2020). Digital transformation in higher education institutions: a systematic literature review. Sensors, 20(11), 3291. https://doi.org/10.3390/s20113291
Chen, L., Zaharia, M., & Zou, J. (2023). FrugalGPT: How to Use Large Language Models While Reducing Cost and Improving Performance. Transactions on Machine Learning Research. https://doi.org/10.48550/arXiv.2305.05176
CYPHER Learning. (2025). What is an LMS? CYPHER Learning Blog.
Deloitte. (2026). AI tokenomics: A CFO's guide to governing the AI P&L. Deloitte Insights.
Dirin, A., Nieminen, M., Laine, T. H., Nieminen, L., & Ghalabani, L. (2023). Emotional Contagion in Collaborative Virtual Reality Learning Experiences: An eSports Approach. Education and Information Technologies. https://doi.org/10.1007/s10639-023-11769-7
Dwivedi, Y. K., et al. (2024). Generative AI in business applications and innovation. International Journal of Information Management.

FinOps Foundation. (2026). Cloud FinOps: Collaborative, Real-Time Cloud Financial Management. FinOps Foundation.
Fluidattacks. (2026). AI Token Economics and Cost Control. Fluidattacks Blog.
Gärtner, B., & Hiebl, M. R. (2018). Issues with big data. En M. Quinn & E. Strauss (Eds.), The Routledge companion to accounting information systems (pp. 161–172). Routledge.
Intelligent Living. (2026). LLM API Pricing Comparison: Every Major Model Compared by Cost. Intelligent Living.
IntuitionLabs. (2026). LLM API Pricing Comparison 2026: OpenAI, Gemini, Claude. IntuitionLabs.
Iternal. (2026). LLM API Pricing Calculator and Cost-Efficiency Strategies. Iternal AI.
Ittner, C. D., & Larcker, D. F. (2001). Assessing empirical research in managerial accounting: a value-based management perspective. Journal of Accounting and Economics, 32(1-3), 349-410.(https://doi.org/10.1016/S0165-4101(01)00026-X)
Khan, M., et al. (2025). AI adoption in SMEs: A systematic literature review. Journal of Business Research.
Kokina, J., & Davenport, T. H. (2017). The emergence of artificial intelligence: How automation is changing auditing. Journal of Emerging Technologies in Accounting.
Lai, J. (2025). Artificial intelligence applications and audit fees: An empirical study. International Review of Economics & Finance, 103(C).
LaunchDarkly. (2026). LLM pricing comparison and dynamic routing. LaunchDarkly Blog.
Lingamgunta. (2026). The AI FinOps framework and governance structure. SSRN.
Martín-Ramallal, P. (2025). AI-Learning. Aceptación de la inteligencia artificial en estudiantes de Comunicación. Campus Virtuales, 14(2), 101-114. https://doi.org/10.54988/cv.2025.2.1596
Mussa, M., & Rosen, S. (1978). Monopoly and product quality. Journal of Economic Theory, 18(2), 301-317.
Murray, A., et al. (2024). RouteLLM: Pre-Inference Classification for Model Routing. arXiv. https://doi.org/10.48550/arXiv.2406.18665
Pierotti, M. (2024). AI for Managerial Accounting. En Artificial Intelligence in Accounting and Auditing (pp. 171-192). Springer. https://doi.org/10.1007/978-3-031-71371-2_8
Prabhu, J. (2026). Rethinking AI: is the future of artificial intelligence frugal? Cambridge Judge Business School.
Schoser, M. (2023). Generative AI and reasoning in accounting. Information Systems Frontiers.
Selgas-Cors, M. (2025). The artificial intelligence in higher education: a paradigm shift? Campus Virtuales, 14(2), 127-138. https://doi.org/10.54988/cv.2025.2.1602
Shore, A., Tiwari, M., Tandon, P., & Foropon, C. (2024). Building entrepreneurial resilience during crisis using generative AI: An empirical study on SMEs. Technovation, 135. https://doi.org/10.1016/j.technovation.2024.103124
Storment, J. R. (2026). Token Economics: The Atomic Unit of AI Value. FinOps Foundation.
Teece, D. J., Pisano, G., & Shuen, A. (1997). Dynamic capabilities and strategic management. Strategic Management Journal, 18(7), 509-533.
Xu, Q. (2026). FairTutor: Equity-Aware Pedagogical LLM Routing for Budget-Constrained AI Tutoring. arXiv preprint arXiv:2606.20713. https://doi.org/10.48550/arXiv.2606.20713
Yaşar, Ş. (2024). Integration of artificial intelligence in management accounting: A SWOT analysis. Journal of Business in The Digital Age, 7(1), 9-19. https://doi.org/10.46238/jobda.1474352
Zhu, Q. (2026). AI Tokenomics: The Economics of Tokens, Computation, and Pricing in Foundation Models. arXiv. https://doi.org/10.48550/arXiv.2606.24616

Martínez-García, A. (2026). Contabilidad de los tokens y enrutamiento dinámico en la inteligencia artificial generativa aplicada a Sistemas de Gestión del Aprendizaje (LMS): Un modelo económico para la educación superior. *Campus Virtuales,* 15(2), 47-62. https://doi.org/10.54988/cv.2026.2.1956