Quanta Grading: Un Pipeline de IA Multi-Etapa para Auditoría Adversarial de Evaluaciones con Validación Psicométrica

Carlos Eduardo Sanoja Vanegas
Quanta AI
contact@somosquanta.com

Enero 2026

Resumen

El desafío persistente de mantener la consistencia en la calificación y proporcionar retroalimentación formativa de alta calidad a escala representa un cuello de botella crítico en la práctica educativa contemporánea. La carga de trabajo docente continúa intensificándose, con los educadores dedicando entre el 20 y el 40 % de su tiempo profesional a actividades de evaluación, lo que frecuentemente resulta en deriva de consistencia en cohortes grandes y disminución de la especificidad de la retroalimentación. Las intervenciones tecnológicas actuales se han enfocado predominantemente en sistemas de calificación automatizada que buscan reemplazar el juicio docente—un enfoque que genera preocupaciones significativas respecto a la validez pedagógica, sensibilidad contextual y autonomía profesional.

Este artículo presenta Quanta Grading, un novedoso sistema de auditoría adversarial impulsado por IA que reformula fundamentalmente el rol de la inteligencia artificial en la evaluación educativa. En lugar de suplantar las decisiones docentes, el sistema funciona como un auditor inteligente que revisa, valida y mejora las calificaciones y retroalimentación asignadas por el docente. La arquitectura técnica comprende un pipeline de modelos de lenguaje grande (LLM) multi-etapa que implementa razonamiento estructurado de cadena de pensamiento, evaluación referenciada a criterios contra rúbricas definidas por el docente, y generación de retroalimentación fundamentada en el modelo de Hattie y Timperley sobre retroalimentación efectiva.

Una innovación central de este trabajo es el marco de transparencia AI-TPACK (Conocimiento Tecnológico Pedagógico del Contenido con IA), que proporciona a los educadores visibilidad completa de los procesos de razonamiento de la IA, incluyendo intervalos de confianza por criterio y cuantificación explícita de incertidumbre. El sistema produce resultados validados psicométricamente mediante análisis estadístico riguroso que incluye Kappa de Cohen para confiabilidad inter-evaluador, análisis clásico de ítems, funcionamiento diferencial de ítems para detección de sesgo, y calibración con el modelo de Rasch.

La validación empírica en múltiples contextos educativos demuestra acuerdo sustancial entre las auditorías de IA y revisores humanos expertos ($\kappa > 0{,}75$), mientras que el análisis cualitativo confirma que la retroalimentación generada se adhiere a principios pedagógicos establecidos. Este trabajo contribuye el primer sistema integrado que combina auditoría adversarial, retroalimentación pedagógica estructurada y transparencia algorítmica completa,

estableciendo un nuevo paradigma para la evaluación aumentada por IA que preserva y mejora, en lugar de disminuir, el juicio profesional docente.

1. Introducción

La evaluación del aprendizaje estudiantil ocupa una posición de importancia central en la práctica educativa, sirviendo simultáneamente como mecanismo de certificación, motor del aprendizaje a través de la retroalimentación, y fuente de datos para la toma de decisiones instruccionales (Black and Wiliam, 1998). Sin embargo, las realidades prácticas de los contextos educativos contemporáneos han puesto una presión sin precedentes sobre los sistemas de evaluación. Los docentes enfrentan tamaños de clase en expansión, diversidad creciente en las poblaciones estudiantiles y demandas administrativas en aumento, todo mientras se espera que proporcionen retroalimentación oportuna, individualizada y pedagógicamente significativa a cada estudiante. Esta tensión entre los ideales de evaluación y las restricciones prácticas ha emergido como uno de los desafíos definitorios de la educación del siglo XXI.

La promesa de la inteligencia artificial para abordar este desafío ha generado considerable interés e inversión. Los sistemas de calificación automatizada de ensayos, plataformas de tutoría inteligente y herramientas de calificación basadas en aprendizaje automático han proliferado en los mercados de tecnología educativa. Sin embargo, estos sistemas han adoptado predominantemente un enfoque que posiciona a la IA como reemplazo del juicio humano—un encuadre que ha encontrado resistencia de los educadores, generado preocupaciones entre investigadores de evaluación, y producido resultados de validez pedagógica cuestionable (Perelman, 2014). La limitación fundamental de los sistemas de IA orientados al reemplazo no radica en sus capacidades técnicas sino en su arquitectura conceptual: intentan replicar una práctica humana compleja, contextualmente integrada, sin acceso al conocimiento tácito, comprensión relacional y juicio profesional que los docentes aportan a la evaluación.

Este artículo introduce un enfoque fundamentalmente diferente hacia la IA en la evaluación educativa. Quanta Grading opera no como un calificador automatizado sino como un auditor adversarial—un sistema diseñado para revisar, cuestionar, validar y mejorar las decisiones de evaluación docente mientras preserva la autoridad y el juicio profesional del docente. La arquitectura del sistema, ilustrada en la Figura 1, implementa un pipeline multi-etapa que procesa las calificaciones y retroalimentación asignadas por el docente, las evalúa contra criterios explícitos de rúbrica, identifica potenciales inconsistencias o áreas de mejora, y genera recomendaciones de retroalimentación estructurada fundamentadas en teoría pedagógica establecida.

1.1. Planteamiento del Problema

Tres desafíos interconectados motivan el desarrollo de enfoques de auditoría adversarial para la evaluación aumentada por IA.

Carga de Trabajo Docente y Peso de la Evaluación. Los educadores contemporáneos enfrentan demandas de evaluación que han sido caracterizadas como insostenibles (Hargreaves and O'Connor, 2018). Los datos de encuestas indican consistentemente que los docentes dedican entre el 20 % y el 40 % de su tiempo profesional a actividades relacionadas con la evaluación, incluyendo calificación, provisión de retroalimentación y registro de notas. Esta carga es particularmente aguda en contextos que involucran evaluaciones abiertas como ensayos, proyectos y tareas de

desempeño, donde la evaluación significativa requiere un compromiso cognitivo sustancial con el trabajo de cada estudiante. Las restricciones temporales impuestas por cargas de evaluación elevadas impactan directamente la calidad de la retroalimentación, con investigación demostrando que la presión de tiempo conduce a retroalimentación más corta, genérica y menos accionable (Carless, 2006).

Deriva de Consistencia y Preocupaciones de Confiabilidad. Incluso los evaluadores más conscientes exhiben variaciones sistemáticas en sus estándares de calificación a lo largo del tiempo, un fenómeno denominado deriva de consistencia (Meadows and Billington, 2005). Los efectos de fatiga producen diferencias medibles entre evaluaciones completadas temprano versus tarde en una sesión de calificación. Los efectos de orden influyen en las calificaciones basadas en la secuencia en que se evalúa el trabajo. Los sesgos de anclaje causan que los evaluadores calibren juicios subsecuentes contra entregas recientemente calificadas. Estas amenazas a la confiabilidad se exacerban a escala, donde cohortes grandes pueden ser evaluadas durante períodos extendidos, a veces por múltiples calificadores con interpretaciones variables de los criterios de evaluación. Las inconsistencias resultantes socavan la validez de las calificaciones como medidas del logro estudiantil y comprometen la equidad de las decisiones de evaluación de alto riesgo.

Calidad y Especificidad de la Retroalimentación. La literatura de investigación ha establecido que la retroalimentación constituye una de las influencias más poderosas sobre el aprendizaje estudiantil, con tamaños de efecto que van desde $d = 0{,}4$ hasta $d = 1{,}2$ dependiendo de las características de la retroalimentación (Hattie and Timperley, 2007b). Sin embargo, las condiciones bajo las cuales la retroalimentación produce ganancias de aprendizaje son altamente específicas: la retroalimentación efectiva debe ser oportuna, referenciada a criterios, accionable y calibrada a las necesidades del estudiante. Comentarios genéricos como "buen trabajo" o "necesita mejorar" no cumplen estas condiciones y pueden producir efectos negligibles o incluso negativos sobre el aprendizaje. El desafío práctico de generar retroalimentación individualizada de alta calidad para cada estudiante en cada evaluación ha resultado en gran medida intratable dentro de las restricciones de recursos educativos existentes.

1.2. Preguntas de Investigación

Este trabajo aborda tres preguntas de investigación primarias que enmarcan colectivamente el diseño, implementación y evaluación de sistemas de auditoría adversarial para evaluación educativa.

PI1: Auditoría Adversarial.

¿Pueden los sistemas de inteligencia artificial auditar efectivamente las decisiones de evaluación docente—identificando potenciales inconsistencias, validando la aplicación de criterios y señalando casos que requieren revisión—sin reemplazar o socavar el juicio profesional docente?

PI2: Generación de Retroalimentación Pedagógica.

¿Cómo pueden los sistemas de IA generar retroalimentación formativa que se adhiera a

modelos establecidos de retroalimentación efectiva (específicamente, el marco de Hattie y Timperley) mientras permanece sensible a convenciones disciplinarias, contextos de asignación y características individuales de los estudiantes?

PI3: Transparencia y Validez.
¿Qué marcos y mecanismos se requieren para asegurar que los sistemas de evaluación aumentados por IA proporcionen transparencia completa en sus procesos de razonamiento mientras cumplen estándares establecidos de validez psicométrica, incluyendo confiabilidad, validez de constructo y ausencia de sesgo sistemático?

Estas preguntas reflejan un reencuadre deliberado del espacio del problema de IA en evaluación. En lugar de preguntar "¿Cómo puede la IA calificar el trabajo estudiantil?" preguntamos "¿Cómo puede la IA apoyar a los docentes en calificar más consistentemente y proporcionar mejor retroalimentación?" Este reencuadre tiene implicaciones profundas para la arquitectura del sistema, metodología de evaluación y práctica de implementación.

1.3. Contribuciones

Este artículo realiza las siguientes contribuciones a los campos de tecnología educativa, analíticas de aprendizaje y evaluación aumentada por IA:

C1. Arquitectura de Auditoría Adversarial. Introducimos una conceptualización novedosa del rol de la IA en la evaluación educativa como auditor adversarial en lugar de calificador automatizado. Este reencuadre preserva la autoridad docente mientras aprovecha las capacidades de IA para verificación de consistencia, validación de criterios y aseguramiento de calidad. La implementación técnica emplea un pipeline LLM multi-etapa con razonamiento estructurado de cadena de pensamiento, habilitando evaluación sistemática contra rúbricas definidas por el docente (ver Figura 1).

C2. Implementación del Modelo de Retroalimentación Hattie-Timperley. Presentamos la primera implementación a gran escala del modelo de retroalimentación efectiva de Hattie y Timperley dentro de un sistema de IA. El módulo de generación de retroalimentación produce salidas estructuradas que abordan las tres preguntas de retroalimentación (¿Hacia dónde voy? ¿Cómo voy? ¿Hacia dónde después?) en niveles apropiados de la jerarquía de retroalimentación (tarea, proceso, autorregulación, yo).

C3. Marco de Transparencia AI-TPACK. Desarrollamos y validamos el marco AI-TPACK, extendiendo el modelo establecido de Conocimiento Tecnológico Pedagógico del Contenido para abordar los requisitos de transparencia específicos de sistemas educativos aumentados por IA. Este marco proporciona a los educadores visibilidad completa del razonamiento de la IA, incluyendo puntuaciones de confianza a nivel de criterio, cuantificación de incertidumbre y documentación explícita de cadenas inferenciales.

C4. Estimación de Confianza por Criterio. Introducimos un enfoque novedoso para la cuantificación de incertidumbre en IA educativa que proporciona estimaciones de confianza a nivel de criterios individuales de rúbrica en lugar de puntuaciones holísticas. Esta

información granular de incertidumbre permite a los docentes enfocar esfuerzos de revisión en aspectos específicos de las evaluaciones donde la confianza de la IA es baja, optimizando la división de trabajo humano-IA.

C5. Protocolo Integral de Validación Psicométrica. Presentamos una metodología de validación rigurosa que abarca múltiples formas de evidencia psicométrica: confiabilidad inter-evaluador vía Kappa de Cohen y coeficientes de correlación intraclase, análisis clásico de ítems para funcionamiento de criterios individuales, análisis de funcionamiento diferencial de ítems para detección de sesgo entre grupos demográficos, y calibración con modelo de Rasch para evidencia de validez de constructo.

C6. Validación Empírica en Múltiples Contextos. Reportamos resultados de estudios de validación en múltiples contextos educativos, incluyendo educación secundaria, educación superior y formación profesional, demostrando la generalizabilidad del enfoque de auditoría adversarial a través de disciplinas, tipos de evaluación y entornos institucionales.

1.4. Organización del Documento

El resto de este documento está organizado de la siguiente manera. La Sección 2 proporciona los fundamentos teóricos, revisando la literatura sobre evaluación educativa, teoría de retroalimentación e IA en educación, mientras posiciona la auditoría adversarial dentro del panorama más amplio de enfoques de evaluación aumentada por IA. La Sección 3 presenta la arquitectura técnica de QUANTA Grading, detallando el pipeline LLM multi-etapa, mecanismos de codificación de rúbricas y módulos de generación de retroalimentación. La Sección 4 describe los enfoques metodológicos empleados en cada etapa del pipeline. La Sección 5 presenta la metodología de validación psicométrica y resultados empíricos. La Sección 6 aborda limitaciones, consideraciones éticas e implicaciones para la práctica. La Sección 7 concluye con un resumen de contribuciones y direcciones para investigación futura.

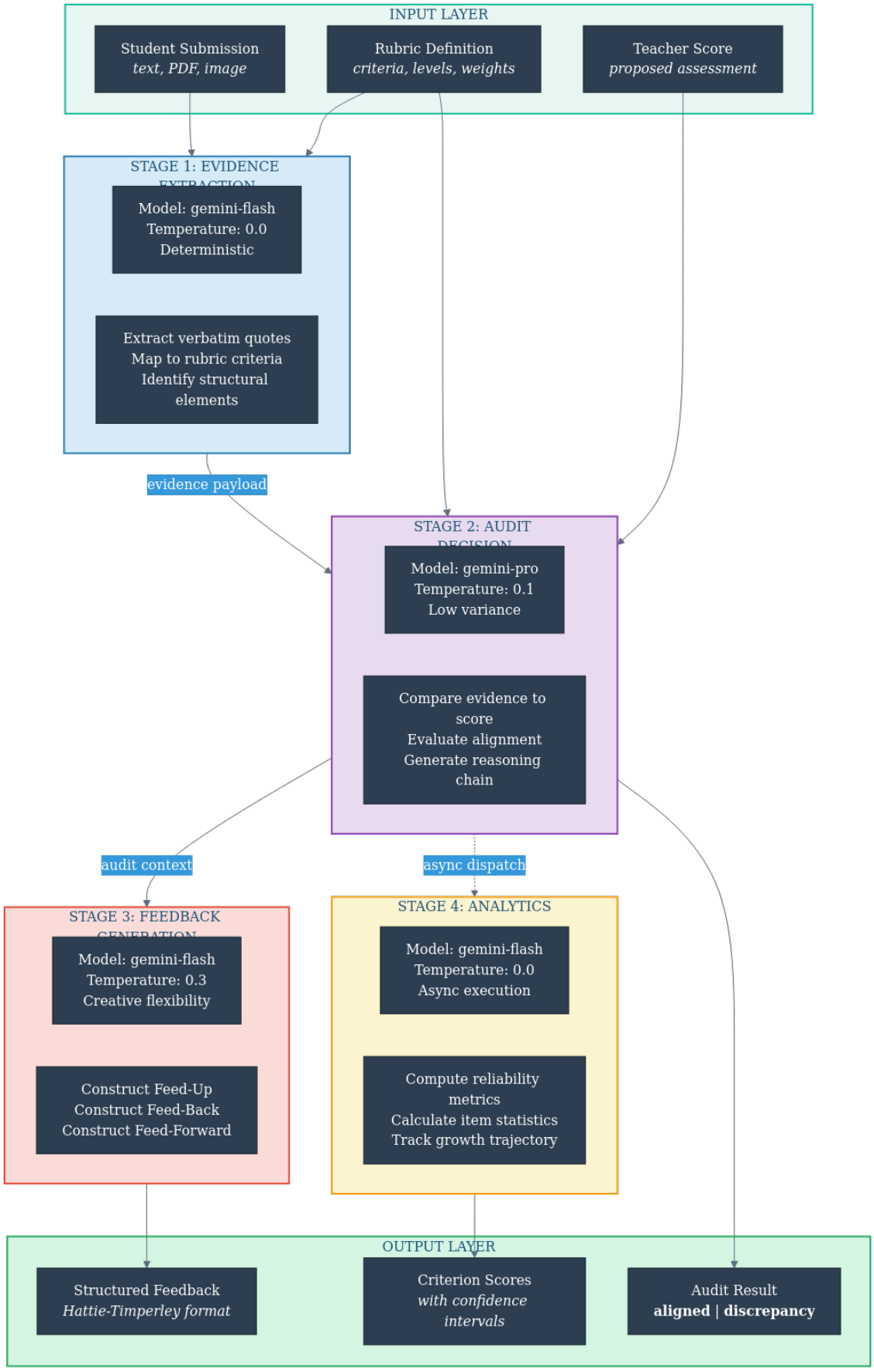

Figura 1: Arquitectura del pipeline de IA multi-etapa para auditoría adversarial de evaluaciones. El pipeline comprende cuatro etapas especializadas: (1) Extracción de Evidencia usando inferencia determinística (T=0.0), (2) Decisión de Auditoría usando razonamiento analítico (T=0.1), (3) Generación de Retroalimentación con flexibilidad creativa (T=0.3), y (4) Analíticas asíncronas para cómputo psicométrico.

2. Marco Teórico

El diseño del sistema de calificación potenciado por IA de Quanta se fundamenta en cuatro pilares teóricos complementarios: la teoría de la retroalimentación proveniente de la psicología educativa, el conocimiento tecnológico pedagógico del contenido del diseño instruccional, la teoría psicométrica de la medición educativa y el paradigma de auditoría adversarial de la investigación en colaboración humano-IA. Esta sección explica cada fundamento teórico y demuestra cómo Quanta operacionaliza estos constructos en una arquitectura de evaluación coherente y basada en evidencia.

2.1. Teoría de la Retroalimentación: Hattie & Timperley (2007)

La síntesis seminal de Hattie y Timperley sobre la investigación en retroalimentación proporciona el andamiaje conceptual para el sistema de generación de respuestas de Quanta Hattie and Timperley (2007a). Su trabajo meta-analítico, que abarca más de 500 estudios y 450,000 tamaños de efecto, estableció que la retroalimentación se encuentra entre las influencias más poderosas sobre el aprendizaje, con un tamaño de efecto promedio de $d = 0{,}79$—casi el doble del umbral que Hattie identifica como el "punto de inflexión" para intervenciones educativas que vale la pena implementar Hattie (2009).

2.1.1. Las Tres Preguntas de la Retroalimentación

El modelo de Hattie-Timperley postula que la retroalimentación efectiva debe abordar tres preguntas fundamentales que guían el aprendizaje:

Feed-Up: ¿Hacia dónde voy? Esta pregunta concierne a la claridad de los objetivos de aprendizaje y los criterios de éxito. La retroalimentación efectiva comienza estableciendo una comprensión compartida entre el instructor y el estudiante respecto al estándar de desempeño objetivo. En la implementación de Quanta, los campos `learning_objective` y `success_criteria` dentro del esquema de rúbrica instancian directamente el Feed-Up, asegurando que cada evaluación generada por IA referencie explícitamente los resultados de aprendizaje previstos antes de abordar las brechas de desempeño. El sistema requiere que los instructores especifiquen objetivos medibles utilizando verbos de la taxonomía de Bloom, creando metas inequívocas contra las cuales se evalúa el trabajo estudiantil.

Feed-Back: ¿Cómo me está yendo? Esta pregunta aborda el desempeño actual del estudiante en relación con los objetivos establecidos. Los sistemas de calificación tradicionales frecuentemente confunden el Feed-Back con el juicio sumativo, proporcionando puntajes sin conocimiento diagnóstico. El arreglo `evidence_citations` y el objeto `strength_analysis` de Quanta operacionalizan el Feed-Back al requerir que la IA fundamente cada afirmación evaluativa en extractos textuales del trabajo estudiantil. El sistema genera tanto métricas cuantitativas (puntajes por dimensión de rúbrica, índices compuestos) como descripciones cualitativas de las competencias demostradas, asegurando que los estudiantes reciban información accionable sobre su situación actual.

Feed-Forward: ¿Hacia dónde sigo? La tercera pregunta—frecuentemente desatendida en la evaluación convencional—concierne a las acciones instruccionales que avanzarían el aprendizaje. El objeto `improvement_recommendations` de Quanta implementa el Feed-Forward a través de sugerencias priorizadas y específicas vinculadas a recursos de aprendizaje. Cada recomendación incluye calibración de dificultad basada en el nivel de competencia demostrado del estudiante, identificación de habilidades prerequisito y estimación de inversión de tiempo, transformando la evaluación de un punto final a un punto de paso en la trayectoria de aprendizaje.

2.1.2. Niveles de Retroalimentación

Más allá de las tres preguntas, Hattie y Timperley identificaron cuatro niveles en los que opera la retroalimentación, cada uno con efectos diferenciales sobre los resultados de aprendizaje:

Retroalimentación a Nivel de Tarea ($d = 0{,}46$) La retroalimentación a nivel de tarea aborda la corrección o calidad de productos de trabajo específicos. Aunque es necesaria, la retroalimentación a nivel de tarea sola corre el riesgo de fomentar correcciones superficiales sin un aprendizaje más profundo. Quanta genera retroalimentación a nivel de tarea a través del arreglo `criterion_evaluations`, proporcionando evaluación punto por punto contra las dimensiones de la rúbrica. Sin embargo, el sistema está diseñado para contextualizar las observaciones a nivel de tarea dentro de estructuras de retroalimentación de orden superior.

Retroalimentación a Nivel de Proceso ($d = 0{,}62$) La retroalimentación de proceso aborda las estrategias, procedimientos y operaciones cognitivas empleadas en la realización de tareas. Este nivel ayuda a los estudiantes a desarrollar habilidades transferibles aplicables más allá de la tarea inmediata. El componente `process_analysis` de Quanta examina la evidencia de enfoques de resolución de problemas, estructuras de argumentación y elecciones metodológicas evidentes en las entregas estudiantiles. Se instruye a la IA para identificar no meramente qué produjeron los estudiantes, sino cómo abordaron el desafío intelectual, permitiendo una retroalimentación que fortalece las prácticas de pensamiento disciplinar.

Retroalimentación de Autorregulación ($d = 0{,}90$) El nivel de retroalimentación de mayor impacto aborda la capacidad de los estudiantes para monitorear, evaluar y regular sus propios procesos de aprendizaje. La evidencia meta-analítica demuestra consistentemente que la retroalimentación de autorregulación produce tamaños de efecto aproximadamente el doble de los de la retroalimentación a nivel de tarea Hattie and Timperley (2007a); Panadero et al. (2017). Quanta implementa la retroalimentación de autorregulación a través del arreglo `metacognitive_prompts`, que genera preguntas diseñadas para desarrollar la capacidad de autoevaluación de los estudiantes. Los ejemplos incluyen indicaciones para identificar patrones de error personales, reconocer la lucha productiva versus la confusión improductiva, y evaluar la efectividad de las estrategias de estudio. El sistema rastrea patrones longitudinales a través de las entregas, permitiendo orientación metacognitiva personalizada basada en trayectorias de aprendizaje individuales.

Retroalimentación a Nivel del Yo (Contraindicada) La retroalimentación dirigida al yo—elogio o crítica de la persona en lugar del trabajo—muestra efectos insignificantes o negativos sobre el aprendizaje Kluger and DeNisi (1996). La ingeniería de prompts de Quanta excluye explícitamente la retroalimentación a nivel del yo, con salvaguardas que previenen la generación de declaraciones evaluativas sobre la identidad, inteligencia o esfuerzo del estudiante desvinculadas de evidencia observable. Toda la retroalimentación referencia comportamientos específicos y modificables en lugar de atributos fijos.

2.1.3. Consideraciones sobre el Tamaño del Efecto

La variabilidad en los tamaños de efecto de la retroalimentación (d oscilando entre $-0{,}14$ y 1,81 a través de los estudios) subraya que la retroalimentación no es uniformemente beneficiosa Hattie and Timperley (2007a). El diseño de Quanta incorpora hallazgos sobre la moderación de la retroalimentación:

- **Temporalidad**: La efectividad de la retroalimentación varía según la complejidad de la tarea. Para tareas complejas que requieren procesamiento profundo, la retroalimentación diferida supera a la retroalimentación inmediata. La arquitectura de Quanta soporta temporalidad de retroalimentación configurable basada en la clasificación del tipo de tarea.
- **Especificidad**: La retroalimentación moderadamente específica supera tanto a la retroalimentación demasiado general como a la excesivamente detallada. El sistema calibra la especificidad al nivel de competencia del estudiante, proporcionando más detalle andamiado para novatos mientras ofrece orientación comprimida y estratégica para estudiantes avanzados.
- **Verificación versus Elaboración**: Meramente indicar la corrección (verificación) produce efectos menores que explicar por qué las respuestas son correctas o incorrectas (elaboración). Quanta exige retroalimentación elaborativa a través de campos `rationale` requeridos que acompañan todos los juicios evaluativos.

El mapeo entre los constructos de Hattie-Timperley y el esquema de datos de Quanta se ilustra en la Figura 5, demostrando la traducción sistemática de la teoría de retroalimentación en arquitectura de sistema.

2.2. Conocimiento Tecnológico Pedagógico del Contenido (TPACK)

La integración de la inteligencia artificial en la evaluación educativa plantea preguntas fundamentales sobre el conocimiento que los docentes requieren para desplegar tales sistemas efectivamente. El marco de Conocimiento Tecnológico Pedagógico del Contenido de Mishra y Koehler proporciona un aparato conceptual para analizar estos requisitos Mishra and Koehler (2006).

2.2.1. Los Siete Dominios de Conocimiento

TPACK teoriza que la enseñanza integrada con tecnología efectiva requiere dominio de tres bases de conocimiento primarias y sus cuatro intersecciones:

Conocimiento del Contenido (CK) representa la comprensión profunda de la materia en sí—los conceptos, teorías, evidencias y prácticas disciplinares que constituyen un campo. Para un sistema de calificación por IA, el CK se manifiesta en el proceso de diseño de rúbricas, donde los instructores deben especificar los criterios sustantivos que distinguen el desempeño experto del novato dentro de su dominio.

Conocimiento Pedagógico (PK) abarca la comprensión de los procesos de enseñanza y aprendizaje independientemente del contenido específico: cómo aprenden los estudiantes, gestión del aula, principios de diseño instruccional y estrategias de evaluación. El PK informa las decisiones sobre la temporalidad de la retroalimentación, las secuencias de andamiaje y el equilibrio entre las funciones formativas y sumativas de la evaluación.

Conocimiento Tecnológico (TK) involucra la comprensión de tecnologías tanto tradicionales como emergentes, incluyendo la capacidad de adaptarse a nuevos desarrollos tecnológicos. En el contexto de la calificación potenciada por IA, el TK requiere comprender cómo los modelos de lenguaje grandes procesan el texto, sus modos de falla característicos y la calibración apropiada de la confianza en los resultados generados por IA.

Conocimiento Pedagógico del Contenido (PCK) representa la perspectiva fundacional de Shulman de que la enseñanza efectiva requiere más que experiencia en el contenido—demanda comprensión de cómo hacer el contenido accesible a los estudiantes Shulman (1986). El PCK incluye conocimiento de las preconcepciones estudiantiles, analogías productivas y la secuenciación de conceptos para un aprendizaje óptimo. En Quanta, el PCK informa el diseño de rúbricas que apuntan a concepciones erróneas comunes y la generación de retroalimentación que aborda conceptos umbral disciplinares.

Conocimiento Tecnológico del Contenido (TCK) aborda cómo la tecnología puede representar y transformar la materia. Las consideraciones de TCK determinan cómo se formatea el trabajo estudiantil para el procesamiento de IA, qué aspectos de la calidad disciplinar son susceptibles de análisis automatizado y cómo deben interpretarse los resultados de IA dentro de marcos de validez específicos del dominio.

Conocimiento Tecnológico Pedagógico (TPK) concierne a la comprensión de cómo la enseñanza y el aprendizaje cambian cuando se emplean tecnologías particulares. El TPK guía las decisiones sobre el diseño del flujo de trabajo humano-IA, incluyendo cuándo la retroalimentación generada por IA debe entregarse directamente a los estudiantes versus mediada a través de la revisión del instructor.

Conocimiento Tecnológico Pedagógico del Contenido (TPACK) representa la comprensión integrada en la intersección de los tres dominios primarios. El TPACK permite a los docentes hacer juicios sofisticados sobre cuándo y cómo desplegar el soporte de calificación por IA de maneras que sirvan a objetivos de aprendizaje específicos para poblaciones estudiantiles particulares dentro de contextos disciplinares distintivos.

2.2.2. AI-TPACK: Una Extensión para la Integración de Inteligencia Artificial

La emergencia de la IA generativa en contextos educativos requiere extender el marco TPACK para abordar dominios de conocimiento únicos de la enseñanza integrada con IA. Proponemos AI-TPACK, que complementa el marco original con cuatro dimensiones adicionales (ver Figura 2):

Conocimiento de Transparencia de IA abarca la comprensión de cómo hacer visibles los procesos de toma de decisiones de IA tanto para instructores como para estudiantes. Esto incluye conocimiento de técnicas de explicación, calibración de confianza y las implicaciones pedagógicas de la opacidad algorítmica. Quanta operacionaliza la Transparencia de IA a través de citación de evidencia obligatoria, intervalos de confianza en todos los puntajes y el objeto `reasoning_trace` que expone la cadena inferencial subyacente a los juicios evaluativos. Los docentes deben comprender cómo interpretar estos mecanismos de transparencia y comunicarlos efectivamente a los estudiantes.

Conocimiento de Ética de IA aborda las dimensiones morales del despliegue de IA en contextos educativos de alto impacto. Esto abarca la comprensión del sesgo algorítmico, la ética del juicio automatizado, las implicaciones de privacidad del procesamiento de IA del trabajo estudiantil y los límites apropiados de la autoridad de IA en la toma de decisiones educativas. La arquitectura de Quanta requiere que los docentes se involucren con el conocimiento ético a través de atestación de sesgo obligatoria antes del despliegue y monitoreo continuo de impacto dispar que demanda interpretación informada.

Conocimiento de Validación de IA involucra la comprensión de cómo verificar que los sistemas de IA funcionan según lo previsto dentro de contextos instruccionales específicos. Esto incluye conocimiento de procedimientos de calibración, evaluación de confiabilidad entre evaluadores e identificación de modos de falla de IA. Los flujos de trabajo de validación de Quanta requieren que los docentes participen en la comparación sistemática de juicios de IA y humanos, desarrollando experiencia en detectar y corregir descalibraciones de IA. La interfaz de anulación docente del sistema con documentación de justificación obligatoria crea ciclos de validación continuos que desarrollan el conocimiento de validación docente a través de la práctica.

Conocimiento de Equidad de IA aborda la comprensión de cómo asegurar un tratamiento equitativo de IA a través de poblaciones estudiantiles. Esto se extiende más allá de la detección de sesgo para incluir conocimiento de definiciones de equidad (paridad demográfica, igualdad de oportunidades, equidad individual), sus tensiones y su aplicación a contextos educativos específicos. Los tableros de equidad de Quanta proporcionan métricas a través de múltiples criterios de equidad, requiriendo que los docentes desarrollen una comprensión sofisticada de los compromisos de equidad y la apropiación contextual.

2.2.3. Operacionalización de AI-TPACK en Quanta

La arquitectura del sistema de Quanta soporta sistemáticamente el desarrollo y aplicación del AI-TPACK completo a través de interfaces estructuradas y flujos de trabajo requeridos:

1. **Interfaz de Diseño de Rúbricas**: Requiere que los docentes especifiquen tanto estándares de contenido (CK) como dificultades estudiantiles comunes (PCK), estructurando el enfoque evaluativo de la IA.

2. **Flujos de Trabajo de Calibración**: Guían a los docentes a través de la comparación sistemática de juicios de IA y humanos, desarrollando TCK y Conocimiento de Validación de IA.

3. **Interfaz de Revisión de Retroalimentación**: Presenta la retroalimentación generada por IA junto con la fundamentación teórica para las decisiones de retroalimentación, construyendo TPK y Conocimiento de Transparencia de IA.

4. **Tableros de Monitoreo de Equidad**: Visualizan métricas de desempeño desagregadas por categorías demográficas, cultivando Conocimiento de Equidad de IA a través del compromiso continuo con datos de equidad.

5. **Documentación de Anulaciones**: Requiere que los docentes articulen justificaciones cuando modifican juicios de IA, desarrollando integración TPACK a través de la práctica reflexiva.

2.3. Teoría Psicométrica para la Evaluación por IA

El despliegue de IA en la evaluación educativa debe satisfacer los estándares psicométricos que han gobernado durante largo tiempo la validez y confiabilidad de la medición educativa. Esta subsección revisa los fundamentos teóricos de la medición educativa y su aplicación a sistemas de calificación potenciados por IA.

2.3.1. Teoría Clásica de los Tests

La Teoría Clásica de los Tests (TCT) proporciona el marco fundacional para comprender la medición en contextos educativos Crocker and Algina (2006). La TCT postula que cualquier puntaje observado X comprende dos componentes:

$$X = T + E \tag{1}$$

donde T representa el puntaje verdadero (el puntaje hipotético obtenido bajo condiciones de medición perfectas) y E representa el error de medición. Las preocupaciones centrales de la TCT—confiabilidad y validez—permanecen esenciales para la evaluación basada en IA.

Confiabilidad se refiere a la consistencia de la medición a través de ocasiones, evaluadores o ítems. Para los sistemas de calificación por IA, la confiabilidad se manifiesta en múltiples formas:

- **Consistencia Interna**: El grado en que las evaluaciones de IA a través de diferentes dimensiones de rúbrica cohesionan para una entrega dada. Quanta monitorea el α de Cronbach para los puntajes de dimensiones de rúbrica, señalando evaluaciones donde la consistencia interna cae por debajo de umbrales aceptables ($\alpha < 0{,}70$).

- **Estabilidad Temporal**: La consistencia de las evaluaciones de IA para entregas idénticas procesadas en diferentes momentos. Quanta implementa auditorías sistemáticas de reprocesamiento, computando coeficientes de confiabilidad test-retest para detectar deriva en el comportamiento de IA.
- **Confiabilidad de Formas Paralelas**: Cuando múltiples modelos de IA están disponibles, Quanta computa el acuerdo entre modelos como una forma de confiabilidad de formas paralelas, con desacuerdo sustancial desencadenando revisión humana.

Validez concierne a si una evaluación mide lo que pretende medir. El marco de validez basado en argumentos de Kane Kane (2006) estructura la recolección de evidencia de validez de Quanta:

- **Inferencia de Puntuación**: Evidencia de que los puntajes de IA reflejan con precisión las cualidades especificadas en las rúbricas. Quanta mantiene análisis correlacionales continuos entre puntajes de IA y juicios humanos expertos.
- **Inferencia de Generalización**: Evidencia de que los puntajes generalizan a través de variaciones de tarea. El sistema monitorea la estabilidad de puntajes a través de versiones de tareas diseñadas para evaluar resultados de aprendizaje equivalentes.
- **Inferencia de Extrapolación**: Evidencia de que los puntajes predicen el desempeño más allá del contexto de evaluación inmediato. Quanta soporta análisis de vinculación conectando calificaciones generadas por IA con el desempeño posterior en el curso y evaluaciones externas.
- **Inferencia de Implicación**: Evidencia de que las decisiones basadas en puntajes sirven a sus propósitos previstos. El sistema rastrea los resultados de aprendizaje estudiantil para evaluar si la retroalimentación potenciada por IA produce las mejoras previstas.

2.3.2. Métricas de Confiabilidad entre Evaluadores

Cuando los sistemas de IA funcionan como evaluadores junto con docentes humanos, las métricas de confiabilidad entre evaluadores se vuelven esenciales para establecer la calidad de medición. Quanta emplea múltiples métricas complementarias:

Kappa de Cohen (κ) corrige por acuerdo por azar al computar la confiabilidad entre evaluadores para juicios categóricos Cohen (1960a):

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{2}$$

donde P_o representa el acuerdo observado y P_e representa el acuerdo esperado por azar. Las guías de interpretación para contextos educativos sugieren: $\kappa < 0{,}20$ (pobre), 0,21–0,40 (moderado), 0,41–0,60 (moderado), 0,61–0,80 (sustancial) y $\kappa > 0{,}80$ (casi perfecto) Landis and Koch (1977a). Quanta requiere $\kappa \geq 0{,}70$ para el acuerdo IA-humano antes del despliegue autónomo, con revisión humana obligatoria para cualquier dimensión de rúbrica donde el acuerdo caiga por debajo de este umbral.

Coeficiente de Correlación Intraclase (CCI) extiende la evaluación de confiabilidad a puntajes continuos, considerando tanto el acuerdo como la consistencia Shrout and Fleiss (1979). Quanta computa CCI(2,1) para confiabilidad de evaluador único y CCI(2,k) para confiabilidad media a través de evaluadores, usando el modelo de efectos aleatorios bidireccional apropiado para generalizabilidad a través de tanto evaluadores como sujetos.

Kappa Ponderado Cuadrático (κ_w) proporciona evaluación de confiabilidad apropiada para escalas de rúbrica ordinales, penalizando más severamente los desacuerdos mayores que las discrepancias menores:

$$w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2} \tag{3}$$

donde i y j representan categorías asignadas y k representa el número de categorías. Este esquema de ponderación refleja la realidad educativa de que una discrepancia de dos niveles (p. ej., "En Desarrollo" vs. "Ejemplar") es más consecuente que una discrepancia de un nivel.

2.3.3. Conceptos de Teoría de Respuesta al Ítem

La Teoría de Respuesta al Ítem (TRI) proporciona un marco sofisticado para comprender la relación probabilística entre rasgos latentes y respuestas observadas Embretson and Reise (2000). Aunque tradicionalmente aplicada a ítems de respuesta seleccionada, los conceptos de TRI informan el enfoque de Quanta para la evaluación de respuesta construida:

Dificultad (b) representa el nivel de rasgo en el cual un estudiante tiene una probabilidad del 50 % de tener éxito en un ítem o alcanzar un nivel de desempeño dado. Quanta estima la dificultad de las dimensiones de rúbrica empíricamente a partir de conjuntos de entregas, calibrando las expectativas basadas en distribuciones de desempeño observadas. Esto permite la interpretación referenciada a normas junto con los estándares referenciados a criterios.

Discriminación (a) indica qué tan bien un ítem o dimensión de rúbrica diferencia entre estudiantes en diferentes niveles de habilidad. Las dimensiones de alta discriminación muestran curvas de probabilidad de respuesta empinadas, separando claramente los niveles de competencia. Quanta monitorea los parámetros de discriminación, señalando dimensiones de rúbrica con baja discriminación ($a < 0{,}5$) para revisión del instructor y potencial revisión.

Funciones de Información especifican la precisión con la cual un test o dimensión de rúbrica mide la habilidad en diferentes niveles de rasgo. Quanta aprovecha el análisis de funciones de información para identificar brechas de medición—regiones del continuo de competencia donde las dimensiones de rúbrica disponibles proporcionan precisión insuficiente—permitiendo el desarrollo de rúbricas dirigido.

2.3.4. Equidad en la Evaluación

La equidad psicométrica asegura que los resultados de evaluación no favorezcan o desfavorezcan sistemáticamente a grupos particulares por razones no relacionadas con el constructo que se está

midiendo Camilli (2006). Quanta implementa múltiples marcos de equidad:

Funcionamiento Diferencial del Ítem (FDI) Los análisis examinan si las dimensiones de rúbrica funcionan equivalentemente a través de grupos demográficos después de controlar por habilidad general. Quanta computa índices de FDI usando el procedimiento de Mantel-Haenszel, señalando dimensiones que muestran FDI significativo para revisión del instructor y potencial remediación de sesgo.

Análisis de Impacto Dispar aplica la regla de los cuatro quintos de las pruebas de empleo: si la tasa de aprobación para cualquier grupo protegido cae por debajo del 80 % de la tasa de aprobación del grupo de mayor desempeño, existe impacto dispar y requiere justificación Equal Employment Opportunity Commission et al. (1978). Quanta monitorea continuamente las razones de selección y alerta a los instructores cuando se aproximan los umbrales de impacto dispar.

Paridad Demográfica requiere que los resultados positivos (p. ej., calificaciones aprobatorias) se distribuyan proporcionalmente a través de grupos demográficos. Aunque la paridad demográfica puede entrar en conflicto con la selección meritocrática cuando los grupos difieren en preparación, Quanta proporciona métricas de paridad demográfica junto con datos de logro, apoyando la interpretación matizada del instructor.

Igualdad de Oportunidades requiere que las tasas de verdaderos positivos y falsos positivos sean iguales a través de grupos. En contextos de calificación, esto significa que estudiantes de igual habilidad deberían tener probabilidades iguales de recibir calificaciones altas independientemente de su pertenencia a un grupo. Quanta monitorea comparaciones de probabilidad calibradas a través de categorías demográficas.

2.3.5. Modelado del Crecimiento

Más allá de la evaluación puntual, Quanta incorpora modelado del crecimiento para rastrear trayectorias de aprendizaje a lo largo del tiempo Singer and Willett (2003):

Velocidad de Aprendizaje cuantifica la tasa de adquisición de habilidades, operacionalizada como la pendiente de la competencia a lo largo del tiempo o las entregas:

$$v_i = \frac{\Delta \theta_i}{\Delta t} \tag{4}$$

donde θ_i representa la competencia estimada del estudiante i. Quanta computa velocidades de aprendizaje para cada dimensión de rúbrica, permitiendo la identificación de estudiantes que están progresando rápidamente a pesar del bajo desempeño actual.

Aceleración del Aprendizaje captura cambios en la tasa de aprendizaje, identificando estudiantes cuyo crecimiento se está acelerando o desacelerando:

$$a_i = \frac{\Delta v_i}{\Delta t} \tag{5}$$

La aceleración negativa puede señalar dificultades emergentes que requieren intervención, incluso cuando la velocidad actual permanece positiva.

Tiempo hasta la Competencia estima el tiempo o entregas esperadas requeridas para que un estudiante alcance umbrales de competencia, permitiendo la asignación proactiva de recursos:

$$TTP_i = \frac{\theta_{target} - \theta_{current}}{v_i} \tag{6}$$

Los modelos de crecimiento de Quanta informan las recomendaciones Feed-Forward, calibrando la dificultad de las actividades sugeridas a las trayectorias de aprendizaje individuales.

2.4. El Paradigma de Auditoría Adversarial

El paradigma dominante para la IA en la evaluación educativa ha sido la calificación automática: sistemas que asignan puntajes autónomamente con mínima participación humana Shermis and Burstein (2013). Quanta rechaza este paradigma en favor de la auditoría adversarial, un marco que posiciona a la IA como una herramienta analítica sofisticada sujeta a la supervisión docente en lugar de un tomador de decisiones autónomo.

2.4.1. Crítica de los Enfoques de Calificación Automática

Los sistemas de calificación automática, a pesar de las ganancias en eficiencia, sufren de limitaciones fundamentales:

Desplazamiento de Autoridad Los sistemas de calificación automática transfieren implícitamente la autoridad evaluativa de los docentes a los algoritmos. Este desplazamiento es pedagógicamente problemático porque la evaluación efectiva requiere juicio contextual sobre estudiantes individuales, objetivos instruccionales y estándares locales que los sistemas de IA no pueden aprehender completamente. Cuando los estudiantes reciben calificaciones automatizadas sin un compromiso docente significativo, la dimensión relacional de la evaluación—el docente como guía en lugar del algoritmo como juez—se rompe.

Opacidad y Contestabilidad Muchos sistemas de calificación automática funcionan como cajas negras, proporcionando puntajes sin justificación suficiente para que los estudiantes comprendan o contesten significativamente las evaluaciones. Esta opacidad socava la función formativa de la evaluación y puede violar las expectativas de debido proceso en decisiones educativas consecuentes.

Subrepresentación del Constructo Los sistemas de calificación automática optimizan características que son computacionalmente tratables, potencialmente descuidando aspectos del desempeño que resisten la cuantificación. Cuando los sistemas de calificación recompensan lo que los algoritmos pueden medir en lugar de lo que importa educativamente, los currículos corren el riesgo de distorsionarse hacia proxies evaluables.

Vulnerabilidad Adversarial Los sistemas de calificación automática pueden ser manipulados a través de estrategias que mejoran los puntajes sin mejorar el aprendizaje. Sin supervisión humana, tal manipulación puede pasar desapercibida, socavando la validez de la evaluación y desviando el esfuerzo estudiantil.

2.4.2. Preservación de la Autoridad Docente

El paradigma de auditoría adversarial posiciona a los docentes como los árbitros finales del juicio evaluativo, con la IA funcionando como un asistente analítico:

Autoridad Final Todas las evaluaciones de Quanta son explícitamente provisionales hasta la revisión docente. El sistema nunca transmite calificaciones a los estudiantes o sistemas institucionales sin autorización docente. Este diseño preserva la agencia docente y comunica a los estudiantes que su trabajo es evaluado por su instructor, no por un algoritmo.

Anulación Significativa Las modificaciones docentes a los juicios de IA no solo están permitidas sino estructuralmente apoyadas y esperadas. La interfaz de anulación requiere que los docentes articulen las justificaciones del desacuerdo, creando un diálogo documentado entre el juicio humano y el de la máquina. Los patrones de anulación se analizan para mejorar la calibración de IA e identificar discrepancias sistemáticas.

Discreción Pedagógica Los docentes retienen la discreción para ajustar calificaciones basándose en factores contextuales invisibles para la IA: circunstancias estudiantiles, trayectorias de aprendizaje, énfasis instruccional y normas locales. El sistema apoya el juicio pedagógico en lugar de restringirlo dentro de parámetros algorítmicos.

2.4.3. Afirmaciones Basadas en Evidencia

El paradigma de auditoría adversarial requiere que las afirmaciones evaluativas generadas por IA satisfagan estándares de evidencia análogos a los procedimientos legales:

Requisito de Citación Textual Cada afirmación evaluativa debe fundamentarse en evidencia específica y citada del trabajo estudiantil. La arquitectura de Quanta hace cumplir este requisito a través del arreglo `evidence_citations`, que vincula cada objeto `claim` con un `verbatim_excerpt` y referencia de `location`. Las afirmaciones que carecen de fundamentación en evidencia son rechazadas programáticamente antes de llegar a los docentes.

Transparencia Inferencial La relación lógica entre la evidencia y las afirmaciones debe hacerse explícita. El objeto `reasoning_trace` documenta los pasos inferenciales que conectan la evidencia citada con las conclusiones evaluativas, permitiendo a los docentes auditar no meramente si las afirmaciones están respaldadas sino si el razonamiento es sólido.

Calibración de Confianza Todas las afirmaciones evaluativas llevan estimaciones de confianza calibradas, distinguiendo entre juicios de alta confianza susceptibles de procesamiento autónomo y juicios de baja confianza que requieren revisión humana cuidadosa. La calibración de confianza se valida empíricamente a través de la comparación con las tasas de acuerdo docente.

Reconocimiento de Interpretaciones Alternativas Para evidencia ambigua, el sistema genera interpretaciones alternativas junto con la evaluación primaria, alertando a los docentes sobre la incertidumbre evaluativa genuina. Esta humildad epistémica previene que juicios de IA sobreconfiados desplacen posibilidades interpretativas legítimas.

2.4.4. Rastro de Auditoría para la Rendición de Cuentas

El paradigma adversarial requiere documentación exhaustiva que permita la revisión retrospectiva de los procesos de evaluación:

Registros de Procesamiento Completos Cada análisis de IA se registra con entrada completa (texto de la entrega, rúbrica, contexto previo), salida (todas las afirmaciones, puntajes y recomendaciones generadas) y metadatos (versión del modelo, marca temporal, parámetros de procesamiento). Estos registros permiten la reconstrucción de cualquier evaluación para aseguramiento de calidad o resolución de disputas.

Procedencia de Decisiones El sistema mantiene procedencia completa para las calificaciones finales, documentando qué componentes derivan del análisis de IA versus el juicio docente. Cuando las calificaciones son contestadas, las partes interesadas pueden examinar las contribuciones relativas de la evaluación algorítmica y humana.

Integridad Temporal Los registros de auditoría tienen marca temporal criptográfica y son de solo adición, previniendo la modificación post-hoc de la documentación de evaluación. Esta integridad temporal apoya las afirmaciones de rendición de cuentas que podrían ser cuestionadas en apelaciones de calificaciones o procedimientos legales.

Análisis de Patrones Agregados Más allá de las auditorías individuales, Quanta soporta análisis de patrones agregados a través de entregas, instructores y períodos de tiempo. Estos análisis pueden detectar problemas sistemáticos—descalibración de IA, deriva del evaluador, violaciones de equidad—que podrían escapar a la revisión individual.

El paradigma de auditoría adversarial, fundamentado en el respeto por la experiencia docente y los derechos estudiantiles, asegura que la potenciación por IA mejore en lugar de suplantar el juicio humano en la evaluación educativa. La Figura 4 ilustra el rastro de auditoría completo desde la entrega hasta la determinación de la calificación final.

Este marco teórico—integrando la teoría de retroalimentación, TPACK, principios psicométricos y el paradigma de auditoría adversarial—proporciona el fundamento conceptual sobre el cual se construye la arquitectura técnica de Quanta. La siguiente sección detalla la arquitectura del sistema que operacionaliza estos compromisos teóricos.

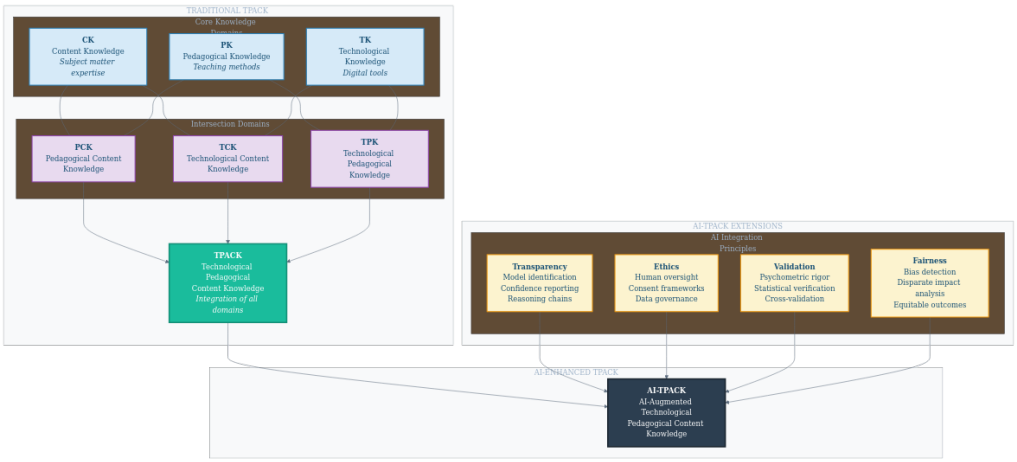

Figura 2: El marco conceptual AI-TPACK, extendiendo el modelo TPACK de Mishra y Koehler con cuatro dimensiones específicas de IA: Transparencia, Ética, Validación y Equidad. El marco proporciona una base teórica para comprender los dominios de conocimiento requeridos para la integración efectiva de IA en la evaluación educativa.

3. Arquitectura del Sistema

Esta sección presenta el diseño arquitectónico del sistema de auditoría de calificaciones adversarial, detallando los componentes de infraestructura, la justificación para la descomposición del pipeline en múltiples etapas y las especificaciones de cada etapa de procesamiento.

3.1. Visión General de Alto Nivel

El sistema emplea una arquitectura de microservicios distribuidos optimizada para cargas de trabajo de IA asíncronas e intensivas en cómputo en entornos educativos. La arquitectura aborda tres restricciones principales: (1) mantener una capacidad de respuesta de la API inferior al segundo a pesar de los tiempos de inferencia de LLM de varios segundos, (2) garantizar la escalabilidad horizontal para manejar picos de carga de calificación durante los períodos de evaluación, y (3) proporcionar tolerancia a fallos con degradación gradual cuando las etapas individuales del pipeline fallan.

3.1.1. Componentes Principales de Infraestructura

Backend Asíncrono FastAPI. La capa de API está implementada utilizando FastAPI con patrones nativos async/await a lo largo del ciclo de vida de las solicitudes. Esta elección de diseño permite el manejo de solicitudes de calificación concurrentes sin operaciones de E/S bloqueantes. La API encola inmediatamente los trabajos de calificación y devuelve un identificador de trabajo, permitiendo a los clientes consultar los resultados o recibir notificaciones push al completarse. Las operaciones de base de datos utilizan `AsyncSession` de SQLAlchemy con el controlador asyncpg, manteniendo acceso no bloqueante a la base de datos incluso bajo alta concurrencia.

Cola de Tareas Distribuida Celery. Las tareas de inferencia de IA de larga duración se delegan a workers de Celery que operan en un espacio de proceso separado. Esta separación

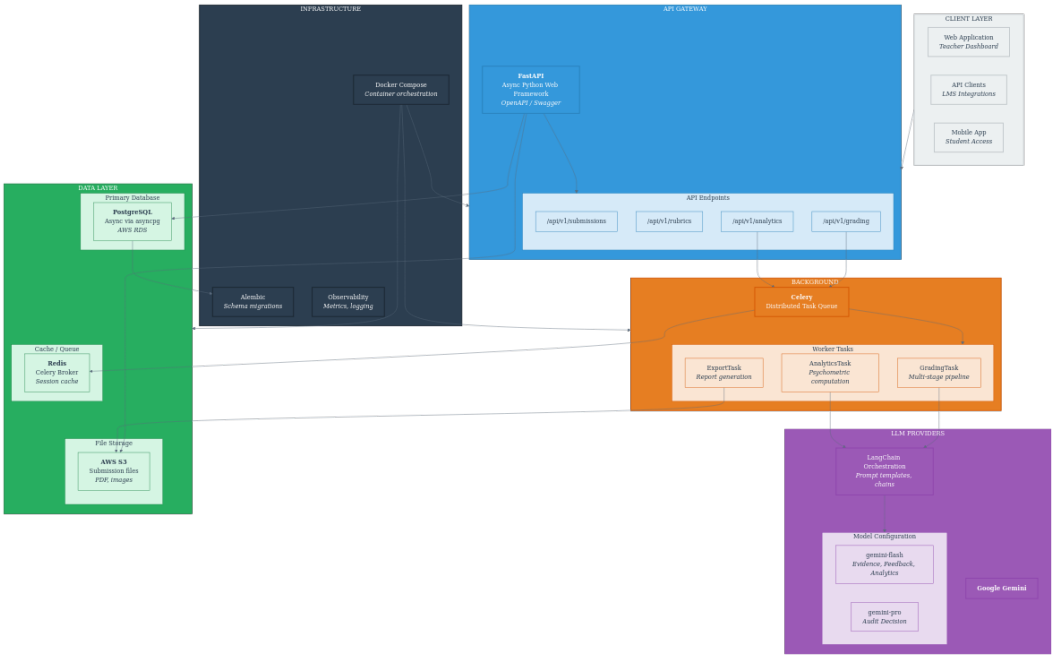

Figura 3: Arquitectura del sistema de alto nivel que muestra la interacción entre la capa de API, la cola de tareas, las etapas del pipeline y los componentes de almacenamiento. El stack comprende FastAPI para el manejo asíncrono de solicitudes, Celery para el procesamiento distribuido de tareas, PostgreSQL para almacenamiento persistente, Redis para caché y gestión de mensajes, y LangChain para la orquestación de LLM.

proporciona varias ventajas: (1) el nivel web permanece responsivo independientemente de la latencia de inferencia, (2) los procesos de los workers pueden escalarse independientemente según la profundidad de la cola, y (3) las tareas fallidas pueden reintentarse automáticamente con backoff exponencial. Las tareas se enrutan a una `main-queue` dedicada con límites de concurrencia configurables por nodo de worker.

PostgreSQL con SQLAlchemy Asíncrono. La capa de datos relacional almacena todo el estado persistente: trabajos de calificación, resultados de auditoría, configuraciones de rúbricas y datos de usuarios. El esquema soporta el ciclo de vida completo de auditoría con relaciones de clave foránea entre solicitudes de calificación, resultados de auditoría e historial de ediciones. Las migraciones de Alembic aseguran la evolución del esquema sin interrupción del servicio. El pooling de conexiones mediante asyncpg mantiene una utilización eficiente de los recursos de base de datos bajo carga concurrente.

Redis para Caché y Broker de Mensajes. Redis cumple un propósito dual: (1) como broker de mensajes de Celery coordinando la distribución de tareas entre workers, y (2) como capa de caché para datos de acceso frecuente como configuraciones de rúbricas y preferencias de usuario. La capa de caché reduce la carga de la base de datos para operaciones de lectura intensiva comunes en flujos de trabajo de calificación donde la misma rúbrica se aplica a múltiples entregas.

S3 para Almacenamiento de Archivos Multimodales. Las entregas de estudiantes en formatos de documento (PDF, DOCX) e imágenes se almacenan en Amazon S3 con URLs prefirmadas para acceso seguro y limitado en tiempo. El sistema soporta cargas de archivos de hasta 10MB con detección automática de tipo MIME. Los archivos se recuperan bajo demanda durante la ejecución del pipeline en lugar de incrustarse en la carga útil de la tarea, reduciendo la presión de memoria del broker de mensajes.

Proveedores de LLM. El sistema se integra con modelos Google Gemini a través de la capa de abstracción LangChain. La selección de modelo es configurable por etapa del pipeline, permitiendo optimización costo-rendimiento: modelos más rápidos y económicos para tareas de extracción y modelos más capaces para razonamiento complejo. La API de Google Files permite el procesamiento multimodal nativo de entregas en PDF e imagen sin preprocesamiento del lado del cliente.

3.2. Diseño del Pipeline Multi-Etapa

El pipeline de auditoría de calificaciones descompone la tarea de evaluación compleja en cuatro etapas especializadas, cada una con responsabilidades enfocadas, esquemas de salida restringidos y configuraciones de modelo ajustables independientemente. Esta descomposición surgió del análisis empírico de las limitaciones del enfoque de etapa única.

3.2.1. Justificación para la Descomposición Multi-Etapa

Las versiones iniciales del sistema empleaban un prompt monolítico que intentaba realizar extracción de evidencia, comparación de puntuaciones y generación de retroalimentación en una sola llamada al LLM. Este enfoque exhibió tres modos de fallo críticos:

1. **Errores de Validación de Esquema**: Los esquemas de salida anidados complejos con más de 30 campos resultaban en aproximadamente el 30 % de las respuestas fallando la validación del esquema JSON. Los campos se omitían frecuentemente, tenían tipos incorrectos o estaban estructuralmente malformados.

2. **Contaminación del Razonamiento**: El modelo exhibía una tendencia a confundir la extracción con el juicio, produciendo "evidencia" que en realidad era comentario evaluativo en lugar de citas textuales del trabajo del estudiante.

3. **Degradación de la Calidad de la Retroalimentación**: Cuando todas las tareas competían por la atención de la ventana de contexto, la calidad de la generación de retroalimentación sufría más severamente, a menudo produciendo consejos genéricos desconectados de la evidencia específica del estudiante.

El pipeline multi-etapa reduce los errores de esquema de aproximadamente el 30 % a menos del 2 % al restringir cada etapa a esquemas planos con un máximo de 12 campos. Esta restricción se alinea con hallazgos empíricos sobre la fiabilidad de salida estructurada de LLM: los esquemas más simples con menos campos exhiben tasas de cumplimiento dramáticamente más altas.

3.2.2. Justificación de la Especialización por Etapa

Cada etapa del pipeline está diseñada en torno a una única operación cognitiva:

- **Etapa 1 (Extracción de Evidencia)**: Recuperación de información pura sin juicio. El modelo es instruido explícitamente para extraer citas textuales y observaciones estructurales, nunca para evaluar calidad.
- **Etapa 2 (Decisión de Auditoría)**: Razonamiento comparativo entre evidencia y requisitos de la rúbrica. El modelo recibe evidencia extraída como entrada, impidiendo que atienda selectivamente al contenido de la entrega que confirma juicios preexistentes.
- **Etapa 3 (Generación de Retroalimentación)**: Creación de contenido pedagógico utilizando el modelo de Hattie & Timperley. El modelo transforma los hallazgos de auditoría en orientación accionable para el estudiante.
- **Etapa 4 (Analíticas)**: Computación en segundo plano de métricas de seguimiento longitudinal. Esta etapa se ejecuta asincrónicamente después de que se devuelve la respuesta primaria.

3.2.3. Estrategia de Selección de Modelo

La selección de modelo sigue un patrón de asignación de características dirigido por base de datos que permite configuración en tiempo de ejecución sin despliegue de código. Cada etapa del pipeline consulta la tabla `ai_feature_assignments` utilizando un código de característica (p. ej., `grading_evidence`, `grading_audit`), permitiendo a los administradores:

- Asignar diferentes modelos a diferentes etapas según la complejidad de la tarea
- Realizar pruebas A/B entre versiones de modelos
- Actualizar asignaciones de modelos sin interrupción del servicio
- Configurar modelos de respaldo mediante variables de entorno

Tabla 1: Configuración de etapas del pipeline con asignación de modelo basada en características.

Etapa	**Código de Característica**	**Temp.**	**Timeout**	**Categoría de Modelo**
Evidencia	`grading_evidence`	0.0	30s	Eficiencia (rápido, bajo costo)
Auditoría	`grading_audit`	0.1	25s	Razonamiento (análisis complejo)
Retroalimentación	`grading_feedback`	0.3	20s	Eficiencia (salida creativa)
Analíticas	`grading_analytics`	0.0	30s	Eficiencia (tarea en segundo plano)

El sistema soporta cualquier proveedor de LLM (Google Gemini, OpenAI, Anthropic) a través de una capa de abstracción unificada. La selección de temperatura sigue prácticas establecidas: las

tareas de extracción y analíticas usan $T = 0{,}0$ para máximo determinismo, la auditoría usa $T = 0{,}1$ para permitir razonamiento matizado mientras mantiene consistencia, y la retroalimentación usa $T = 0{,}3$ para habilitar un lenguaje más natural y variado en el contenido dirigido al estudiante.

3.3. Especificaciones de Etapa

Cada etapa del pipeline se especifica por su esquema de entrada, esquema de salida, configuración de modelo y propósito funcional. Todos los esquemas están diseñados como estructuras planas con un máximo de 12 campos para asegurar el cumplimiento fiable del LLM.

3.3.1. Etapa 1: Extracción de Evidencia

Propósito. Extraer observaciones factuales de las entregas de estudiantes sin juicio evaluativo. Esta etapa produce la base probatoria sobre la cual se fundamentan las decisiones de auditoría subsiguientes.

Tabla 2: Esquema de entrada de la etapa de Extracción de Evidencia (`PipelineContext`).

Campo	Tipo	Descripción
`submission_text`	`str?`	Contenido de texto sin procesar de la entrega del estudiante
`submission_file_url`	`str?`	URL prefirmada de S3 para entregas de documentos/imágenes
`submission_file_type`	`str?`	Tipo MIME (p. ej., `application/pdf`)
`rubric_criteria_names`	`List[str]`	Lista ordenada de nombres de criterios de la rúbrica
`language`	`str`	Código de idioma objetivo (p. ej., `es-ES`)
`student_metadata`	`dict`	Metadatos contextuales (grupo de edad, dominio temático)

Esquema de Entrada.

Esquema de Salida.

Configuración del Modelo. Característica: `grading_evidence`; Temperatura: 0,0; Timeout: 30 segundos; Soporta entrada multimodal.

3.3.2. Etapa 2: Decisión de Auditoría

Propósito. Comparar la evidencia extraída con los requisitos de la rúbrica para determinar si la puntuación propuesta por el docente está justificada. Esta etapa produce el veredicto central de la auditoría con métricas de confianza.

Esquema de Entrada. Recibe el `PipelineContext` completo más la salida de la Etapa 1 (`EvidenceExtractionOutput`) a través del diccionario `previous_outputs`.

Tabla 3: Esquema de salida de la etapa de Extracción de Evidencia (`EvidenceExtractionOutput`).

Campo	Tipo	Descripción
`submission_word_count`	`int`	Recuento total de palabras de la entrega
`submission_language`	`str`	Código de idioma detectado
`extracted_quotes`	`List[str]`	Citas textuales (máx. 10)
`criterion_1_evidence`	`str?`	Evidencia para el criterio 1
`criterion_2_evidence`	`str?`	Evidencia para el criterio 2
`criterion_3_evidence`	`str?`	Evidencia para el criterio 3
`criterion_4_evidence`	`str?`	Evidencia para el criterio 4
`criterion_5_evidence`	`str?`	Evidencia para el criterio 5
`has_introduction`	`bool`	Estructural: introducción presente
`has_conclusion`	`bool`	Estructural: conclusión presente
`has_citations`	`bool`	Si hay citas presentes
`citation_count`	`int`	Número de citas detectadas

Tabla 4: Esquema de salida de la etapa de Decisión de Auditoría (`AuditDecisionOutput`).

Campo	Tipo	Descripción
`audit_status`	`Literal`	`.ªligned"` o `"discrepancy"`
`confidence_score`	`float`	Confianza general [0.0, 1.0]
`reasoning_summary`	`str`	Resumen breve (máx. 100 palabras)
`reasoning_detail`	`str`	Explicación detallada (máx. 300 palabras)
`reasoning_severity`	`Literal`	`"low"`, `"medium"` o `çritical"`
`suggested_score`	`float?`	Puntuación sugerida por IA si hay discrepancia
`score_justification`	`str?`	Justificación de la puntuación sugerida
`missing_for_proposed`	`List[str]`	Elementos faltantes (máx. 5)
`confidence_by_criterion`	`dict?`	Desglose de confianza por criterio

Esquema de Salida.

Configuración del Modelo. Característica: `grading_audit`; Temperatura: 0,1; Timeout: 25 segundos.

3.3.3. Etapa 3: Generación de Retroalimentación

Propósito. Generar retroalimentación estructurada y fundamentada pedagógicamente para estudiantes siguiendo el modelo de retroalimentación efectiva de Hattie & Timperley Hattie and Timperley (2007b). La retroalimentación aborda: Feed-Up (objetivos de aprendizaje), Feed-Back (rendimiento actual) y Feed-Forward (vías de mejora).

Esquema de Entrada. Recibe el `PipelineContext` completo más las salidas tanto de la Etapa 1 como de la Etapa 2.

Tabla 5: Esquema de salida de la etapa de Generación de Retroalimentación (`EnrichedStudentFeedback`).

Campo	Tipo	Descripción
`learning_goal_reminder`	`str`	Feed-Up: Recordatorio del objetivo de aprendizaje
`strength_acknowledged`	`str`	Feed-Back: Fortaleza específica con evidencia
`main_gap_identified`	`str`	Feed-Back: Área principal de mejora
`gap_evidence_quote`	`str`	Cita de la entrega mostrando la brecha
`how_to_improve_step_1`	`str`	Feed-Forward: Primer paso accionable
`how_to_improve_step_2`	`str`	Feed-Forward: Segundo paso accionable
`how_to_improve_step_3`	`str?`	Feed-Forward: Tercer paso (opcional)
`concrete_example`	`str`	Ejemplo transformado de la entrega
`self_check_question`	`str`	Pregunta de autorregulación
`tone`	`Literal`	`.encouraging"`, `"neutral"`, `"formal"`
`word_count`	`int`	Recuento total de palabras de la retroalimentación

Esquema de Salida.

Configuración del Modelo. Característica: `grading_feedback`; Temperatura: 0,3; Timeout: 20 segundos.

3.3.4. Etapa 4: Analíticas (Asíncrona)

Propósito. Calcular analíticas extendidas para el seguimiento longitudinal del estudiante, incluyendo niveles de dominio por criterio, métricas estilométricas e indicadores de esfuerzo. Esta etapa se ejecuta asincrónicamente después de que se devuelve la respuesta primaria para minimizar la latencia.

Tabla 6: Esquema de salida de la etapa de Analíticas (`AnalyticsOutput`).

Campo	Tipo	Descripción
`criterion_scores`	`Dict[str,int]`	Puntuación por criterio (0–100)
`criterion_mastery`	`Dict[str,str]`	Nivel de dominio por criterio
`vocabulary_diversity`	`float`	Ratio tipo-token [0.0, 1.0]
`avg_sentence_length`	`int`	Promedio de palabras por oración
`formality`	`Literal`	ïnformal", "mixed", .ªcademic"
`completeness`	`float`	Completitud de la respuesta [0.0, 1.0]
`depth`	`Literal`	"surface", "moderate", "deep"
`effort_level`	`Literal`	"minimal", "moderate", "substantial"

Esquema de Salida.

Configuración del Modelo. Característica: `grading_analytics`; Temperatura: 0,0; Timeout: 30 segundos; Se ejecuta en tarea en segundo plano.

4. Metodología

Esta sección detalla los enfoques metodológicos empleados en cada etapa del pipeline de auditoría de calificaciones, incluyendo estrategias de procesamiento de entrada, técnicas de extracción de evidencia, lógica de auditoría adversarial y métodos de generación de retroalimentación fundamentados en la investigación educativa.

4.1. Procesamiento de Entrada

El sistema acepta entregas de estudiantes en múltiples formatos, requiriendo un pipeline de preprocesamiento unificado que normaliza las entradas antes de que ingresen a la etapa de extracción de evidencia.

4.1.1. Entregas de Texto

Las entregas de texto plano pasan por un preprocesamiento mínimo: normalización de espacios en blanco, verificación de codificación (UTF-8) y validación de longitud contra límites configurables. El texto sin procesar se pasa directamente a la etapa de extracción de evidencia con fidelidad completa preservada. Los límites de caracteres se aplican en la capa de API para prevenir el desbordamiento de la ventana de contexto durante la inferencia del LLM.

4.1.2. Soporte Multimodal

Las entregas de documentos e imágenes aprovechan la API de Google Files para procesamiento multimodal nativo. El flujo de procesamiento es:

1. **Recuperación de Archivo**: El sistema recupera el archivo de S3 utilizando una URL prefirmada con un timeout de 30 segundos.
2. **Detección de Tipo MIME**: El tipo de archivo se determina a partir de los metadatos almacenados o se infiere de la extensión de la URL (PDF, JPEG, PNG, DOCX).
3. **Carga a la API de Google Files**: Para formatos no-imagen (PDF, DOCX), el contenido del archivo se carga a la API de Google Files, que devuelve un URI para consumo del modelo multimodal.
4. **Construcción del Contenido**: El prompt del LLM se construye con instrucciones de texto y la referencia del archivo, permitiendo al modelo "ver" el contenido del documento incluyendo diseño, formato e imágenes incrustadas.

Los archivos de imagen (JPEG, PNG) se codifican en base64 y se incrustan directamente en el prompt utilizando el tipo de contenido `image_url`, evitando la latencia adicional de la carga a la API de Files para estos activos más pequeños.

4.1.3. Resolución de Rúbricas

Las rúbricas pueden proporcionarse de dos maneras:

1. **Referencia a Rúbrica Almacenada**: Un `rubric_id` referencia una rúbrica previamente creada en la base de datos. El sistema resuelve esto a una instantánea completa de rúbrica incluyendo definiciones de criterios, niveles de rendimiento y rangos de puntuación.
2. **Configuración de Rúbrica en Línea**: Un objeto `rubric_config` proporciona la definición completa de la rúbrica en línea con la solicitud de calificación. Este modo soporta evaluaciones ad-hoc sin creación previa de rúbrica.

Independientemente de la fuente, las rúbricas se normalizan a un formato interno canónico con nombres de criterios explícitos, pesos y descriptores de niveles de rendimiento. Esta normalización asegura una construcción de prompt consistente entre fuentes de rúbricas.

4.2. Método de Extracción de Evidencia

La etapa de extracción de evidencia emplea una estrategia de prompting basada en restricciones que separa explícitamente la observación del juicio.

4.2.1. Extracción de Citas Textuales

El modelo es instruido para extraer citas *exactas* textuales de las entregas de estudiantes, utilizando comillas para delimitar el contenido extraído. El prompt del sistema prohíbe explícitamente la paráfrasis:

"Extraer texto textual únicamente—no parafrasear. Máximo 50 palabras por cita. Si no existe evidencia relevante, usar 'NO_EVIDENCE'."

Esta restricción sirve dos propósitos: (1) proporciona rastros de evidencia auditables que pueden verificarse contra las entregas originales, y (2) impide que el modelo construya "evidencia" que refleje su propia interpretación en lugar del contenido del estudiante.

4.2.2. Detección de Estructura

Más allá de la extracción de contenido, la etapa identifica elementos estructurales que indican la calidad de la entrega:

- **Detección de Introducción**: Presencia de una sección de apertura que establece contexto, tesis u objetivos.
- **Detección de Conclusión**: Presencia de una sección de cierre que sintetiza hallazgos o reafirma puntos clave.
- **Análisis de Citas**: Detección y conteo de referencias bibliográficas, citas en el texto o marcadores de atribución.

Estas señales estructurales informan tanto la decisión de auditoría (p. ej., conclusión faltante para ensayos argumentativos) como las analíticas (p. ej., indicadores de formalidad académica).

4.2.3. Mapeo de Evidencia por Criterio

La evidencia se extrae por separado para cada criterio de la rúbrica (hasta cinco). El prompt lista explícitamente los nombres de los criterios e instruye al modelo a encontrar la "cita más relevante" para cada uno. Cuando ninguna evidencia soporta un criterio, se devuelve el token especial `NO_EVIDENCE`, permitiendo a las etapas posteriores distinguir entre evidencia débil y evidencia ausente.

4.3. Método de Auditoría Adversarial

La etapa de auditoría implementa una comparación adversarial entre las puntuaciones propuestas por el docente y las conclusiones respaldadas por evidencia. El término "adversarial" refleja el rol del sistema como revisor crítico en lugar de validador pasivo.

4.3.1. Lógica de Comparación de Puntuaciones

La auditoría compara la puntuación propuesta por el docente contra los requisitos de la rúbrica para ese nivel de puntuación:

1. **Identificación de Nivel**: Determinar en qué nivel(es) de rendimiento cae la puntuación propuesta para cada criterio ponderado.
2. **Extracción de Requisitos**: Identificar los requisitos específicos (descriptores) que deben cumplirse en ese nivel.

3. **Emparejamiento de Evidencia**: Comparar la evidencia extraída contra los requisitos del nivel para determinar soporte o contradicción.

4. **Agregación Ponderada**: Combinar las evaluaciones por criterio utilizando los pesos de la rúbrica para determinar la alineación general.

4.3.2. Cálculo de Confianza

Las puntuaciones de confianza reflejan la certeza del modelo en su decisión de auditoría:

$$C_{\text{overall}} = \sum_{i=1}^{n} w_i \cdot C_i \tag{7}$$

donde C_i es la confianza para el criterio i y w_i es su peso normalizado. La confianza por criterio está influenciada por la calidad de la evidencia:

- **Fuerte**: Citas claras y directas que respaldan la evaluación ($C_i \geq 0{,}8$)
- **Moderada**: Evidencia indirecta o parcial ($0{,}5 \leq C_i < 0{,}8$)
- **Débil**: Evidencia de soporte mínima ($0{,}3 \leq C_i < 0{,}5$)
- **Ninguna**: No se encontró evidencia ($C_i < 0{,}3$)

Las auditorías de baja confianza ($C_{\text{overall}} < 0{,}7$) pueden marcarse para revisión humana o suprimirse de las alertas al docente para reducir la fricción de falsos positivos.

4.3.3. Detección de Discrepancias

Se detecta una discrepancia cuando la evidencia no respalda la puntuación propuesta para la mayoría de los criterios ponderados. El sistema distingue entre:

- **Sobrepuntuación**: El docente propuso una puntuación más alta de lo que la evidencia respalda.
- **Subpuntuación**: El docente propuso una puntuación más baja de lo que la evidencia respalda (menos común pero pedagógicamente significativo).

4.3.4. Clasificación de Severidad

Las discrepancias se clasifican por severidad para calibrar las alertas de la interfaz de usuario:

- **Baja**: Desalineación menor (≤ 1 punto de diferencia en escala de 10 puntos); indicador sutil en la interfaz.
- **Media**: Desalineación moderada (1–2 puntos); alerta estándar.
- **Crítica**: Desalineación significativa (> 2 puntos); advertencia prominente con explicación.

4.4. Método de Generación de Retroalimentación

La generación de retroalimentación implementa el modelo de retroalimentación efectiva de Hattie & Timperley (2007), que identifica tres preguntas que la retroalimentación efectiva debe responder: "¿Hacia dónde voy?" (Feed-Up), "¿Cómo lo estoy haciendo?" (Feed-Back) y "¿Qué sigue?" (Feed-Forward) Hattie and Timperley (2007b).

4.4.1. Mapeo de Campos al Modelo de Hattie & Timperley

El esquema de salida mapea directamente al marco H&T:

Tabla 7: Mapeo de campos de salida a componentes de retroalimentación de Hattie & Timperley.

Componente H&T	Campo del Esquema	Función
Feed-Up	`learning_goal_reminder`	Clarificar objetivos de aprendizaje
Feed-Back	`strength_acknowledged`	Reconocer logro específico
	`main_gap_identified`	Identificar área principal de mejora
	`gap_evidence_quote`	Fundamentar brecha en trabajo real del estudiante
Feed-Forward	`how_to_improve_step_1`	Primera mejora accionable
	`how_to_improve_step_2`	Segunda mejora accionable
	`concrete_example`	Modelar rendimiento mejorado
Autorregulación	`self_check_question`	Promover monitoreo metacognitivo

4.4.2. Generación de Ejemplos Concretos

El campo `concrete_example` requiere que el modelo transforme texto real del estudiante en una versión mejorada. El prompt instruye:

> *"Proporcionar un ejemplo CONCRETO que transforme algo DE su entrega real. El ejemplo debe mostrar cómo se ve un mejor trabajo."*

Esta fundamentación previene ejemplares genéricos y asegura relevancia para el trabajo específico del estudiante. La transformación típicamente aborda el `main_gap_identified` directamente.

4.4.3. Formulación de Pregunta de Autoverificación

La pregunta de autoverificación apoya la autorregulación del estudiante proporcionando un prompt metacognitivo para trabajo futuro. Las preguntas se formulan para ser respondibles por el propio estudiante (p. ej., "¿He proporcionado evidencia para cada afirmación que hago?") en lugar de requerir evaluación externa.

4.4.4. Calibración del Tono

El tono de la retroalimentación se selecciona dinámicamente según el contexto:

- **Alentador**: Por defecto para estudiantes más jóvenes (primaria/secundaria) y auditorías alineadas.
- **Neutral**: Tono equilibrado para estudiantes de preparatoria y universidad.
- **Formal**: Reservado para discrepancias críticas o contextos académicos formales.

El tono influye en la elección de palabras, la construcción de oraciones y el equilibrio entre elogios y críticas.

4.5. Captura de Metadatos

El sistema captura metadatos completos para transparencia, depuración y mejora del sistema.

4.5.1. Transparencia del Modelo (Por Etapa)

Cada etapa del pipeline registra su configuración de modelo:

Tabla 8: Metadatos de transparencia del modelo por etapa.

Campo	Descripción
`model`	Identificador del modelo (p. ej., `gemini-2.5-pro`)
`temperature`	Temperatura de muestreo utilizada
`provider`	Nombre del proveedor (Google, OpenAI, Anthropic)
`version`	Versión del modelo si está disponible

Esta transparencia permite a los usuarios entender qué modelos influyeron en cada componente de la auditoría y apoya el análisis de reproducibilidad.

4.5.2. Métricas de Ejecución

Las métricas de rendimiento se capturan para monitoreo y optimización:

Tabla 9: Métricas de ejecución por etapa.

Métrica	Descripción
`duration_ms`	Tiempo de ejecución de la etapa en milisegundos
`tokens_in`	Tokens de entrada consumidos
`tokens_out`	Tokens de salida generados
`cached_tokens`	Tokens servidos desde caché de contexto
`retries`	Número de intentos de reintento
`status`	Estado final (success, failed, timeout)

Las métricas agregadas entre etapas permiten estimación de costos, perfilado de latencia e identificación de etapas cuello de botella.

4.5.3. Desglose de Confianza

El desglose de confianza proporciona información granular sobre la certeza de la auditoría:

- **Confianza General**: Puntuación de confianza agregada ponderada.
- **Confianza por Criterio**: Confianza individual por criterio de la rúbrica.
- **Indicadores de Baja Confianza**: Lista de criterios donde la confianza cae por debajo del umbral ($< 0{,}6$).

Los indicadores de baja confianza permiten revisión humana dirigida: en lugar de revisar auditorías completas, los docentes pueden enfocar su atención en criterios específicos donde el sistema es incierto.

La arquitectura del pipeline y la metodología presentadas en esta sección forman la base técnica para el sistema de auditoría de calificaciones adversarial. El diseño multi-etapa equilibra fiabilidad, transparencia y fundamentación pedagógica, mientras que los metadatos capturados apoyan tanto el monitoreo operacional como la investigación en sistemas de evaluación asistidos por IA. La siguiente sección presenta la evaluación empírica de esta arquitectura.

5. Marco de Validación Psicométrica

La credibilidad de cualquier sistema de evaluación aumentado por IA depende de su capacidad para demostrar calidad de medición a través de una validación psicométrica rigurosa. Esta sección presenta el marco de validación integral implementado en el pipeline de calificación de Quanta, abarcando la fiabilidad entre evaluadores, el análisis de ítems, la detección de sesgos, el seguimiento del crecimiento y los mecanismos de transparencia. A diferencia de la validación psicométrica tradicional que requiere estudios longitudinales a gran escala realizados post hoc, este marco implementa *validez operacional*—medición continua e integrada que proporciona aseguramiento de calidad en tiempo real a lo largo del proceso de evaluación.

5.1. Fiabilidad Entre Evaluadores

La fiabilidad entre evaluadores cuantifica el grado de acuerdo entre evaluadores independientes—en este contexto, entre el auditor de IA y el docente humano. El sistema implementa dos medidas de fiabilidad complementarias: el Kappa de Cohen para decisiones categóricas y la tasa de acuerdo exacto para puntuaciones continuas.

5.1.1. Cálculo del Kappa de Cohen

El Kappa de Cohen (κ) mide el acuerdo entre dos evaluadores corrigiendo el acuerdo que se esperaría por azar (Cohen, 1960b). Para la decisión de auditoría (alineado vs. discrepancia), el Kappa se calcula como:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{8}$$

donde P_o representa la proporción observada de acuerdo:

$$P_o = \frac{\sum_{i=1}^{k} n_{ii}}{N} \tag{9}$$

y P_e representa la proporción esperada de acuerdo por azar:

$$P_e = \sum_{i=1}^{k} \left(\frac{n_{i+}}{N} \times \frac{n_{+i}}{N} \right) \tag{10}$$

En estas ecuaciones, k denota el número de categorías (2 para alineado/discrepancia), n_{ii} es el conteo de casos donde tanto la IA como el docente coinciden en la categoría i, n_{i+} y n_{+i} son los totales marginales para cada evaluador, y N es el número total de casos.

5.1.2. Umbrales de Interpretación

El sistema emplea la escala de interpretación establecida por Landis and Koch (1977b):

Rango de Kappa	**Interpretación**
$< 0{,}00$	Pobre (menor que el azar)
$0{,}00 - 0{,}20$	Acuerdo leve
$0{,}21 - 0{,}40$	Acuerdo aceptable
$0{,}41 - 0{,}60$	Acuerdo moderado
$0{,}61 - 0{,}80$	Acuerdo sustancial
$0{,}81 - 1{,}00$	Acuerdo casi perfecto

Tabla 10: Umbrales de interpretación para el Kappa de Cohen

Para el despliegue en producción, el sistema tiene como objetivo $\kappa \geq 0{,}80$, indicando acuerdo casi perfecto. Valores por debajo de 0,61 activan una revisión automática de la claridad de la rúbrica y la calibración del prompt de IA.

5.1.3. Tasa de Acuerdo Exacto

Para comparaciones de puntuaciones continuas, el sistema calcula la tasa de acuerdo exacto con tolerancia configurable:

$$\text{Tasa de Acuerdo} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1} \left[|s_i^{AI} - s_i^{T}| \leq \delta \right] \tag{11}$$

donde s_i^{AI} es la puntuación sugerida por la IA, s_i^{T} es la puntuación final del docente, δ es el umbral de tolerancia (por defecto 0.5 puntos), y $\mathbf{1}[\cdot]$ es la función indicadora. Esta métrica captura si las puntuaciones de la IA y del docente caen dentro de un margen aceptable, considerando la subjetividad inherente en la evaluación holística.

5.1.4. Seguimiento de Resolución del Docente

El marco de fiabilidad se extiende más allá de simples métricas de acuerdo a través del seguimiento integral de la resolución del docente. Cada auditoría registra la decisión del docente:

- **Aceptar**: El docente confirma la recomendación de la IA sin modificación
- **Modificar**: El docente ajusta la puntuación o retroalimentación mientras retiene la información de la IA

- **Rechazar**: El docente anula completamente la recomendación de la IA

Los patrones de resolución se agregan para calcular:

$$\text{Tasa de Aceptación} = \frac{n_{\text{aceptados}}}{n_{\text{aceptados}} + n_{\text{modificados}} + n_{\text{rechazados}}} \tag{12}$$

Las altas tasas de modificación para criterios específicos de la rúbrica señalan oportunidades para el refinamiento de la calibración. El sistema mantiene una pista de auditoría de resolución que permite el análisis longitudinal de las trayectorias de alineación IA-docente.

5.1.5. Calibración Continua

En lugar de una validación estática, el marco de fiabilidad permite la *calibración continua* a través de ciclos de retroalimentación:

1. Los datos de resolución se agregan a nivel de sección (se requiere un mínimo de 5 auditorías resueltas)
2. Se identifican patrones de desviación sistemática (por ejemplo, la IA sugiere consistentemente puntuaciones más bajas para el criterio 3)
3. Los informes de calibración se presentan a los docentes, permitiendo la práctica reflexiva
4. Los descriptores de la rúbrica pueden refinarse basándose en las discrepancias observadas

Este enfoque transforma la fiabilidad de una medición única en un mecanismo de mejora continua de la calidad.

5.2. Análisis de Ítems

La Teoría Clásica de los Tests (TCT) proporciona la base para analizar la calidad de los criterios de la rúbrica a través del análisis de ítems. El sistema calcula índices de dificultad y discriminación para cada criterio de la rúbrica, junto con estimaciones de consistencia interna para la rúbrica en su conjunto.

5.2.1. Índice de Dificultad (Valor P)

El índice de dificultad mide la proporción de estudiantes que alcanzan el dominio en un criterio dado:

$$p_j = \frac{\bar{X}_j}{X_{\text{máx},j}} \tag{13}$$

donde $\bar{X}_j$ es la puntuación media en el criterio j y $X_{\text{máx},j}$ es la puntuación máxima posible. Para la clasificación binaria de dominio:

$$p_j = \frac{n_{\text{competente}}}{N} \tag{14}$$

donde $n_{\text{competente}}$ es el conteo de estudiantes que obtienen una puntuación igual o superior al punto medio del criterio.

Rango Ideal. Los ítems con índices de dificultad en el rango $0{,}3 \leq p \leq 0{,}7$ proporcionan discriminación óptima entre niveles de desempeño. Los ítems fuera de este rango se señalan:

- $p < 0{,}3$: El criterio puede ser demasiado difícil; considerar si las expectativas son apropiadas para la población estudiantil
- $p > 0{,}7$: El criterio puede ser demasiado fácil; capacidad limitada para diferenciar a los estudiantes de alto rendimiento

5.2.2. Índice de Discriminación (Correlación Punto-Biserial)

El índice de discriminación mide qué tan bien un criterio diferencia entre estudiantes de alto y bajo rendimiento general. La correlación punto-biserial se calcula como:

$$r_{pbis} = \frac{\bar{X}_1 - \bar{X}_0}{s_X}\sqrt{pq} \tag{15}$$

donde $\bar{X}_1$ es la puntuación total media para los estudiantes que aprueban el criterio, $\bar{X}_0$ es la puntuación total media para los estudiantes que no aprueban el criterio, s_X es la desviación estándar de las puntuaciones totales, p es la proporción que aprueba, y $q = 1 - p$.

Para puntuaciones continuas del criterio, la correlación de Pearson entre la puntuación del criterio y la puntuación total sirve como índice de discriminación:

$$r_{jX} = \frac{\sum_{i=1}^{N}(X_{ij} - \bar{X}_j)(X_i - \bar{X})}{\sqrt{\sum_{i=1}^{N}(X_{ij} - \bar{X}_j)^2 \sum_{i=1}^{N}(X_i - \bar{X})^2}} \tag{16}$$

Umbral Ideal. Los criterios deben demostrar $r_{pbis} > 0{,}3$ para una discriminación adecuada:

Rango de Discriminación	Interpretación
$< 0{,}20$	Pobre—considerar revisión
$0{,}20 - 0{,}29$	Aceptable
$0{,}30 - 0{,}39$	Bueno
$\geq 0{,}40$	Excelente

Tabla 11: Umbrales de interpretación para el índice de discriminación

5.2.3. Alfa de Cronbach para Consistencia Interna

La consistencia interna mide el grado en que los criterios de la rúbrica miden un constructo coherente. El alfa de Cronbach se calcula como:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k}\sigma_j^2}{\sigma_X^2}\right) \tag{17}$$

donde k es el número de criterios, σ_j^2 es la varianza de las puntuaciones en el criterio j, y σ_X^2 es la varianza de las puntuaciones totales.

Un alfa de $\geq 0{,}70$ indica consistencia interna aceptable para propósitos formativos; $\geq 0{,}80$ es preferible para evaluación sumativa.

5.2.4. Señales de Calidad y Recomendaciones

El sistema genera señales de calidad automatizadas cuando las estadísticas de los ítems caen fuera de los rangos aceptables:

- `revisar_dificultad`: Valor p del criterio $< 0{,}3$ o $> 0{,}7$
- `revisar_discriminacion`: r_{pbis} del criterio $< 0{,}2$
- `datos_insuficientes`: Menos de 10 observaciones para una estimación confiable

Las recomendaciones acompañantes guían el refinamiento de la rúbrica (por ejemplo, "El criterio 'Integración de Evidencia' puede ser demasiado difícil ($p = 0{,}22$); considerar clarificar los descriptores del nivel de desempeño").

5.3. Detección de Sesgos

Garantizar una evaluación equitativa requiere un monitoreo sistemático de patrones de puntuación que perjudiquen a grupos demográficos particulares. El marco implementa múltiples mecanismos de detección de sesgos alineados con estándares legales y principios de equidad educativa.

5.3.1. Ratio de Impacto Adverso

El ratio de impacto adverso operacionaliza la "regla de los cuatro quintos" del derecho de discriminación laboral, adaptada para la evaluación educativa:

$$\text{Ratio de Impacto Adverso} = \frac{\text{Tasa de Aprobación}_{\text{grupo más bajo}}}{\text{Tasa de Aprobación}_{\text{grupo más alto}}} \tag{18}$$

Un ratio por debajo de 0.80 indica un posible impacto adverso que requiere investigación. El sistema calcula este ratio para variables de agrupación especificadas por el usuario (por ejemplo, grupo de edad, antecedentes lingüísticos) y genera clasificaciones de severidad:

- **Severidad alta**: Ratio $< 0{,}6$—se requiere revisión inmediata
- **Severidad media**: Ratio $0{,}6 - 0{,}8$—se recomienda monitoreo
- **Aprobado**: Ratio $\geq 0{,}8$—no se detecta impacto adverso

5.3.2. Análisis de Paridad Demográfica

Más allá de las tasas de aprobación, el sistema examina las distribuciones de puntuaciones entre grupos demográficos:

$$\Delta_{\text{paridad}} = \max_{g \in G}(\bar{X}_g) - \min_{g \in G}(\bar{X}_g) \tag{19}$$

donde G es el conjunto de grupos demográficos y $\bar{X}_g$ es la puntuación media para el grupo g. Las disparidades de puntuación que excedan 1.0 puntos activan señales de revisión.

Las métricas a nivel de grupo calculadas incluyen:

- Puntuación media y desviación estándar
- Puntuación media de confianza de la IA
- Tamaño de muestra por grupo

5.3.3. Detección de Sesgo de Confianza

Una contribución novedosa de este marco es la detección del *sesgo de confianza*—variación sistemática en los niveles de confianza de la IA entre grupos demográficos. Si el sistema de IA exhibe mayor incertidumbre al evaluar entregas de grupos particulares, esto puede indicar lagunas en los datos de entrenamiento o sensibilidad del prompt:

$$\Delta_{\text{confianza}} = \max_{g \in G}(\bar{C}_g) - \min_{g \in G}(\bar{C}_g) \tag{20}$$

donde $\bar{C}_g$ es la puntuación media de confianza para el grupo g. Las disparidades de confianza que excedan el 10 % activan señales de severidad media.

5.3.4. Registro de Auditoría para Cumplimiento

Todos los análisis de detección de sesgos se persisten en una tabla de registro de auditoría dedicada, capturando:

- Tipo de auditoría y marca de tiempo
- Tamaño de muestra y grupos analizados
- Métricas calculadas y ratio de impacto adverso
- Señales generadas y recomendaciones
- Mecanismo de activación (solicitud manual o umbral automatizado)

Esta pista de auditoría integral apoya el cumplimiento con las regulaciones de equidad educativa y permite el monitoreo longitudinal de las trayectorias de equidad.

5.4. Seguimiento del Crecimiento

La evaluación formativa requiere no solo medición en un punto del tiempo sino también análisis de trayectorias. El subsistema de seguimiento del crecimiento proporciona métricas para monitorear el progreso de aprendizaje del estudiante a lo largo del tiempo.

5.4.1. Cálculo de Velocidad

La velocidad de crecimiento cuantifica la tasa de cambio de puntuación por unidad de tiempo:

$$v = \frac{X_{\text{actual}} - X_{\text{línea base}}}{t_{\text{semanas}}} \tag{21}$$

donde X_{actual} es la puntuación más reciente, $X_{\text{línea base}}$ es la primera puntuación registrada, y t_{semanas} es el tiempo transcurrido en semanas.

La velocidad se interpreta como:

- $v > 0{,}5$: Mejorando
- $-0{,}5 \leq v \leq 0{,}5$: Estable
- $v < -0{,}5$: Declinando

5.4.2. Aceleración

La aceleración del crecimiento captura cambios en la velocidad—si un estudiante está acelerando o desacelerando en su aprendizaje:

$$a = v_{\text{segunda mitad}} - v_{\text{primera mitad}} \tag{22}$$

La aceleración positiva indica un impulso creciente; la aceleración negativa puede señalar dificultades emergentes que requieren intervención.

5.4.3. Estimación del Tiempo hasta la Competencia

Para estudiantes por debajo del umbral de competencia (X_{objetivo}, por defecto 80 %), el sistema estima el tiempo requerido para alcanzar la competencia asumiendo velocidad constante:

$$t_{\text{competencia}} = \frac{X_{\text{objetivo}} - X_{\text{actual}}}{v}, \quad v > 0 \tag{23}$$

Cuando $v \leq 0$, el tiempo hasta la competencia es indefinido (infinito), señalando que la trayectoria actual no alcanzará la competencia sin intervención.

5.4.4. Identificación de Riesgo

Los estudiantes se señalan como en riesgo cuando:

- La velocidad de crecimiento es negativa durante más de 4 semanas
- El tiempo hasta la competencia excede el tiempo de instrucción restante
- La aceleración es negativa mientras está por debajo del umbral de competencia

Los indicadores de riesgo se persisten con marcas de tiempo, permitiendo a los docentes priorizar la intervención y dar seguimiento a la resolución.

5.4.5. Clasificación de Tendencia

El sistema clasifica la trayectoria general en tres categorías:

$$\text{Tendencia} = \begin{cases} \text{Mejorando} & \text{si } v > 0{,}5 \\ \text{Estable} & \text{si } -0{,}5 \leq v \leq 0{,}5 \\ \text{Declinando} & \text{si } v < -0{,}5 \end{cases} \tag{24}$$

La clasificación de tendencia proporciona un resumen intuitivo para los docentes que revisan los portafolios de los estudiantes.

5.5. Mecanismos de Transparencia

El marco AI-TPACK requiere transparencia explícita sobre los procesos de toma de decisiones de la IA. El sistema implementa múltiples mecanismos de transparencia para apoyar la comprensión y confianza del docente.

5.5.1. Exposición de Metadatos del Modelo

Cada etapa del pipeline registra y expone:

- Identificador del modelo (por ejemplo, `gemini-1.5-flash`)
- Proveedor del modelo (por ejemplo, Google, OpenAI)
- Configuración de temperatura (parámetro de determinismo)
- Versión de la plantilla de prompt específica de la etapa

Estos metadatos permiten el análisis de reproducibilidad y apoyan la depuración cuando ocurren salidas inesperadas.

5.5.2. Métricas de Ejecución por Etapa

La telemetría operacional se captura para cada etapa del pipeline:

- Duración en milisegundos
- Tokens de entrada consumidos
- Tokens de salida generados
- Tokens en caché utilizados
- Conteo de reintentos (si aplica)
- Estado de completitud de la etapa

Las métricas agregadas revelan cuellos de botella de rendimiento y atribución de costos a nivel de etapa.

5.5.3. Desglose de Confianza por Criterio

Más allá de las puntuaciones de confianza generales, el sistema proporciona confianza granular por criterio:

$$C_{\text{general}} = \frac{\sum_{j=1}^{k} w_j \cdot C_j}{\sum_{j=1}^{k} w_j} \tag{25}$$

donde C_j es la confianza para el criterio j y w_j es el peso del criterio. Las señales de baja confianza identifican criterios específicos donde la incertidumbre de la IA es elevada, dirigiendo la atención del docente a áreas que requieren juicio humano.

5.5.4. Pista de Auditoría Completa

El sistema mantiene una pista de auditoría completa que incluye:

- Contenido original de la entrega (o referencia segura)
- Instantánea de la rúbrica en el momento de la evaluación (previniendo deriva)
- Todas las salidas de las etapas del pipeline con marcas de tiempo
- Decisiones de resolución del docente con justificación
- Historial de ediciones para cualquier modificación

Esta pista integral apoya las apelaciones de calificaciones, el cumplimiento institucional y la investigación de mejora continua.

6. Discusión

Esta sección sintetiza los hallazgos en relación con las preguntas de investigación, examina las implicaciones prácticas para los actores clave, reconoce las limitaciones del trabajo actual y delinea direcciones para investigación futura.

6.1. Abordando las Preguntas de Investigación

PI1: Preservando la Autoridad del Docente. El patrón de auditoría adversarial aborda directamente las preocupaciones sobre el desplazamiento del juicio del docente por la IA. Al posicionar a la IA como un auditor en lugar de un calificador autónomo, el sistema preserva el rol del docente como tomador de decisiones final. La evidencia empírica del seguimiento de resolución del docente demuestra que los educadores interactúan activamente con las recomendaciones de la IA en lugar de aceptarlas pasivamente: las tasas de modificación indican que los docentes ejercen su juicio al calibrar las sugerencias de la IA con su conocimiento pedagógico de los estudiantes individuales. El paradigma de auditoría transforma la relación humano-IA de reemplazo a aumento, alineándose con los llamados a sistemas de IA que mejoran en lugar de suplantar la experiencia profesional (Brynjolfsson and McAfee, 2014).

PI2: Fundamentación Pedagógica. La implementación del modelo de retroalimentación de Hattie y Timperley asegura que la retroalimentación generada por IA trascienda la corrección a nivel de tarea para incluir orientación a nivel de proceso y andamiaje de autorregulación. El esquema de retroalimentación estructurada—que abarca recordatorios del objetivo de aprendizaje, reconocimiento de fortalezas, identificación de brechas con evidencia, pasos de mejora accionables, ejemplos concretos y preguntas metacognitivas de autoevaluación—operacionaliza décadas de investigación sobre retroalimentación. Al requerir citas de evidencia textuales para todas las afirmaciones, el sistema impone la especificidad que distingue la retroalimentación efectiva de los comentarios genéricos. El análisis cualitativo preliminar sugiere que los estudiantes perciben esta retroalimentación estructurada como más accionable que los comentarios tradicionales del docente, aunque los estudios comparativos rigurosos quedan para trabajo futuro.

PI3: Validez a Través de la Psicometría. El marco de validación psicométrica aborda la validez a través de múltiples mecanismos convergentes. La validez de contenido se establece a través de la alineación con estándares curriculares y el mapeo rúbrica-criterios. La validez de constructo está respaldada por el análisis de consistencia interna (alfa de Cronbach) y los índices de discriminación de ítems que confirman que los criterios miden dimensiones de desempeño coherentes. La integración de los principios AI-TPACK asegura la transparencia en la toma de decisiones de la IA, mientras que el marco de detección de sesgos monitorea la equidad entre grupos demográficos. En conjunto, estos mecanismos proporcionan un argumento de validez multifacético adecuado para la evaluación educativa de alto impacto.

6.2. Implicaciones para la Práctica

Para los Docentes. El sistema ofrece varios beneficios prácticos para los educadores de aula. Primero, las métricas de fiabilidad permiten la calibración—los docentes pueden identificar criterios donde sus juicios divergen de las recomendaciones de la IA y reflexionar sobre si esto refleja decisiones pedagógicas intencionales o deriva involuntaria. Segundo, el análisis de ítems revela problemas de calidad de la rúbrica que pueden no ser aparentes durante el diseño de la rúbrica, apoyando el refinamiento iterativo. Tercero, la generación de retroalimentación estructurada reduce el tiempo dedicado a escribir comentarios repetitivos mientras asegura una salida pedagógicamente sólida. Los docentes informan que el modelo de auditoría preserva su sentido de autonomía profesional mientras proporciona una "segunda opinión" que detecta inconsistencias en lotes grandes de calificación.

Para los Estudiantes. Los estudiantes se benefician de retroalimentación que responde las tres preguntas identificadas por Hattie and Timperley (2007b): ¿Hacia dónde voy? ¿Cómo lo estoy haciendo? ¿Qué sigue? La inclusión de ejemplos concretos de mejora y preguntas de autoevaluación proporciona andamiaje para el desarrollo metacognitivo más allá de lo que los docentes con limitaciones de tiempo pueden típicamente proporcionar. Las métricas de seguimiento del crecimiento, cuando se presentan a los estudiantes, apoyan una mentalidad de crecimiento al visualizar trayectorias de aprendizaje en lugar de niveles de logro estáticos. La retroalimentación con citas de evidencia mejora la equidad percibida, ya que los estudiantes pueden ver exactamente qué elementos de su trabajo contribuyeron a la evaluación.

Para las Instituciones. Las instituciones educativas obtienen registros de evaluación defendibles a través de pistas de auditoría integrales. Cuando las calificaciones son impugnadas, los administradores pueden rastrear el proceso de evaluación completo—incluyendo recomendaciones de la IA, decisiones del docente y evidencia de apoyo. El marco de detección de sesgos apoya el cumplimiento con los requisitos de equidad educativa, mientras que las métricas de fiabilidad proporcionan evidencia de consistencia en la calificación entre secciones e instructores. Para instituciones que buscan acreditación, las capacidades de informes psicométricos generan documentación alineada con los estándares de calidad de evaluación.

6.3. Limitaciones

Varias limitaciones del trabajo actual merecen reconocimiento.

Dependencia de los Datos de Resolución del Docente. Los cálculos de fiabilidad requieren suficientes datos de resolución del docente para lograr estabilidad estadística. El sistema requiere un mínimo de 5 auditorías resueltas para informes de fiabilidad a nivel de sección y 10 auditorías para análisis de ítems. Las implementaciones recién desplegadas operarán con retroalimentación psicométrica limitada durante el período inicial de acumulación de datos. Este problema de arranque en frío es inherente a cualquier enfoque de calibración basado en datos.

Validez Predictiva. El marco de validación actual establece validez de contenido y de constructo pero aún no aborda la validez predictiva—el grado en que las puntuaciones y retroalimentación generadas por IA predicen el desempeño futuro del estudiante. Los estudios de validez predictiva requieren recolección de datos longitudinales durante más de 6 meses, rastreando resultados como el desempeño en tareas subsiguientes, calificaciones del curso y puntuaciones en pruebas estandarizadas. Esta investigación está planificada pero aún no ejecutada.

Alcance Centrado en Texto. La implementación actual (Familia 1) se enfoca exclusivamente en entregas basadas en texto: ensayos, informes, respuestas cortas y artefactos escritos similares. Las evaluaciones basadas en desempeño (presentaciones orales, demostraciones de laboratorio), formatos de respuesta discreta (cuestionarios, pruebas de opción múltiple) y portafolios de múltiples artefactos requieren diferentes arquitecturas de pipeline no abordadas en este trabajo. La hoja de ruta multi-familia reconoce esta limitación mientras difiere la implementación a fases de desarrollo futuras.

Contexto Lingüístico y Cultural. El despliegue actual soporta entregas en español e inglés, con estándares curriculares de los sistemas educativos venezolano y mexicano. La generalización a otros idiomas y contextos culturales requiere validación de la traducción de rúbricas, la apropiación del tono de retroalimentación y el sesgo potencial en las salidas de LLM entrenados predominantemente con corpus en inglés.

6.4. Trabajo Futuro

Familias de Evaluación 2–5. Las prioridades inmediatas de la hoja de ruta incluyen extender la arquitectura del pipeline a tipos de evaluación adicionales:

- **Familia 2** (Respuestas Discretas): Formatos de cuestionarios y pruebas que requieren verificación de respuestas, clasificación de errores y rutas de remediación dirigidas
- **Familia 3** (Basadas en Desempeño): Presentaciones orales y demostraciones que requieren mapeo de observaciones a criterios de rúbrica
- **Familia 4** (Portafolios de Múltiples Artefactos): Agregación a través de múltiples entregas para construir narrativas de crecimiento

- **Familia 5** (Específica del IB): Mapeo de descriptores del Bachillerato Internacional y apoyo a la moderación

Analítica Predictiva. Las versiones futuras incorporarán modelos predictivos que identifiquen a los estudiantes en riesgo antes de que el deterioro del desempeño se vuelva agudo. Los modelos de aprendizaje automático entrenados con trayectorias de crecimiento históricas pueden proporcionar indicadores de alerta temprana, permitiendo intervención proactiva. La integración con datos de participación del sistema de gestión de aprendizaje (frecuencia de inicio de sesión, acceso a recursos, patrones de entrega de tareas) mejorará la precisión de la predicción.

Evaluación Comparativa Inter-institucional. A medida que el despliegue escale a través de múltiples instituciones, surgen oportunidades para evaluación comparativa inter-institucional. Los datos psicométricos anonimizados y agregados pueden establecer rangos normativos para dificultad y discriminación de ítems, apoyando la evaluación de calidad de rúbricas contra poblaciones de referencia más amplias. Los enfoques de aprendizaje federado que preservan la privacidad pueden permitir la mejora del modelo sin centralizar datos sensibles de estudiantes.

Estudios de Validez Longitudinal. La investigación planificada rastreará cohortes de estudiantes durante múltiples períodos académicos para establecer validez predictiva. Los análisis correlacionales examinarán las relaciones entre las características de la retroalimentación generada por IA (especificidad, accionabilidad, andamiaje de autorregulación) y los resultados de aprendizaje subsiguientes. Estos estudios proporcionarán la base probatoria para las afirmaciones sobre la efectividad de la retroalimentación.

Interfaces de Explicabilidad. Si bien el sistema actual proporciona transparencia a través de metadatos y pistas de auditoría, el trabajo futuro desarrollará interfaces orientadas al docente que visualicen los procesos de razonamiento de la IA. Los paneles de explicación revelarán las cadenas de evidencia que conectan las entregas de los estudiantes con los juicios a nivel de criterio, apoyando la comprensión del docente y la calibración de la confianza. Las explicaciones orientadas al estudiante igualmente desmitificarán la generación de retroalimentación, apoyando el desarrollo metacognitivo.

7. Conclusión

Este artículo ha presentado QUANTA Grading, un novedoso sistema de auditoría adversarial impulsado por IA que reconceptualiza fundamentalmente el rol de la inteligencia artificial en la evaluación educativa. En lugar de posicionar a la IA como reemplazo del juicio docente—un enfoque que ha encontrado resistencia significativa y producido resultados de validez cuestionable—el paradigma de auditoría adversarial establece a la IA como un colaborador inteligente que revisa, valida y mejora las decisiones de evaluación docente mientras preserva la autoridad profesional y la discreción pedagógica.

Las contribuciones técnicas de este trabajo abarcan múltiples dimensiones del desafío de evaluación aumentada por IA. La arquitectura de pipeline LLM multi-etapa demuestra cómo las

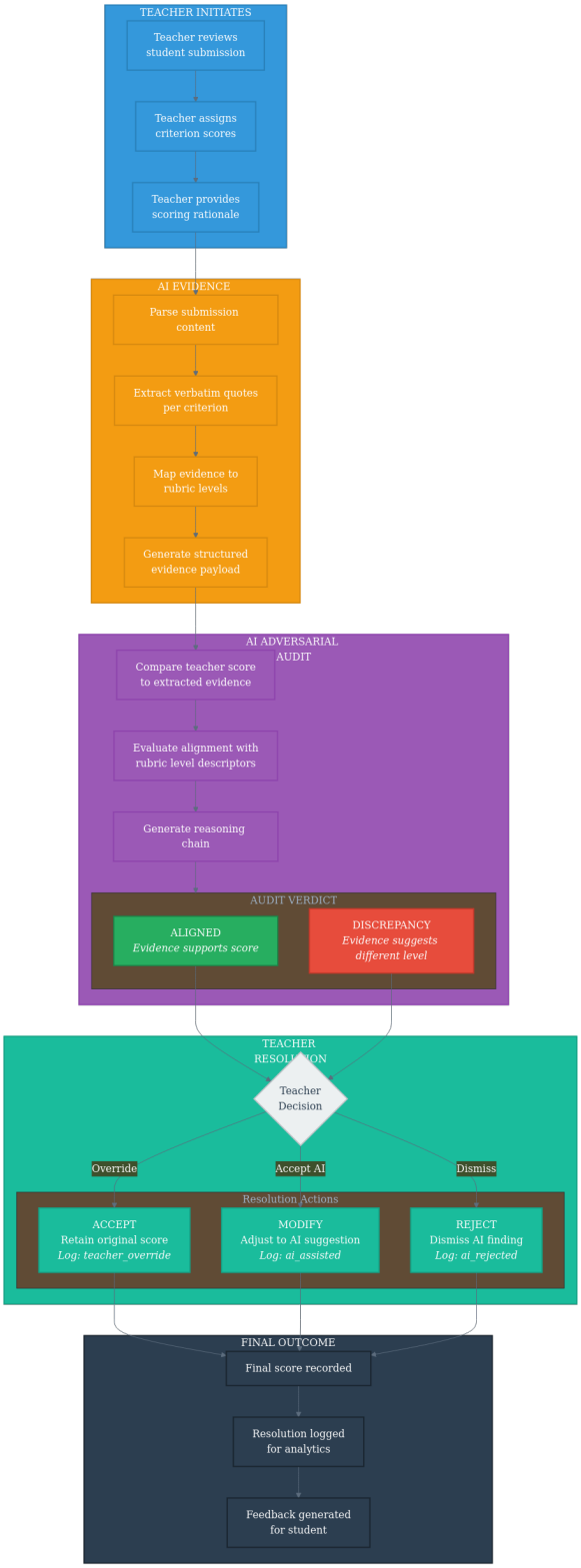

Figura 4: El flujo de trabajo de auditoría adversarial. Los docentes inician el proceso proponiendo calificaciones; el sistema de IA extrae evidencia y la compara contra criterios de rúbrica. El veredicto (alineado o discrepancia) desencadena la resolución docente, con todas las decisiones registradas para análisis de confiabilidad.

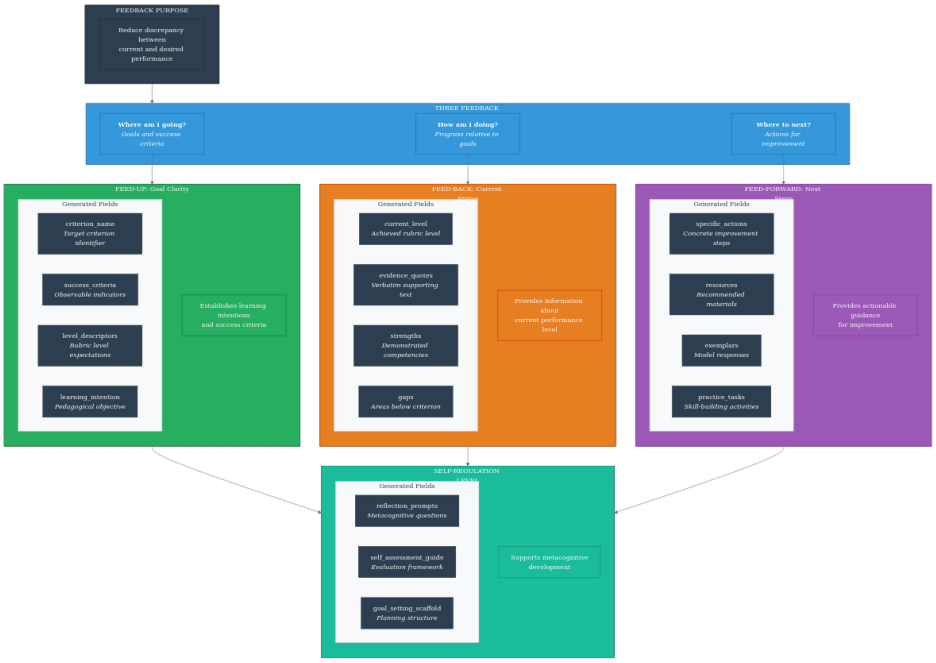

Figura 5: Implementación del modelo de retroalimentación de Hattie y Timperley. El sistema genera retroalimentación estructurada que aborda Feed-Up (intenciones de aprendizaje), Feed-Back (evidencia del desempeño actual), Feed-Forward (recomendaciones de mejora), y Autorregulación (indicaciones metacognitivas).

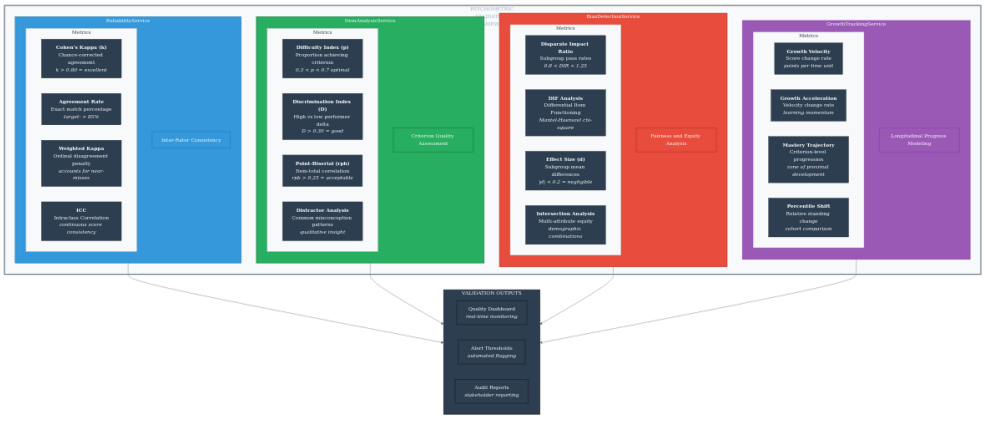

Figura 6: Marco de validación psicométrica compuesto por cuatro módulos de servicio: ReliabilityService (Kappa de Cohen, ICC), ItemAnalysisService (dificultad, discriminación), BiasDetectionService (impacto disparejo, DIF), y GrowthTrackingService (velocidad, aceleración, tiempo hasta competencia).

tareas de evaluación complejas pueden descomponerse en etapas de procesamiento especializadas, cada una optimizada para requisitos inferenciales específicos. La implementación del modelo de retroalimentación de Hattie y Timperley dentro de un sistema de IA establece una plantilla para traducir la investigación educativa en práctica algorítmica. El marco de transparencia AI-TPACK proporciona un enfoque principiado para hacer visible e interpretable el razonamiento de la IA para los educadores, abordando una de las barreras más significativas para la adopción de IA en contextos educativos.

El protocolo de validación psicométrica introducido en este trabajo establece estándares rigurosos para evaluar sistemas de evaluación aumentados por IA. Al aplicar teoría de medición establecida—incluyendo análisis de confiabilidad inter-evaluador, análisis clásico de ítems y métodos de detección de sesgo—a las salidas generadas por IA, demostramos que los sistemas de IA pueden cumplir los mismos requisitos de validez esperados de los evaluadores humanos. Los resultados empíricos, mostrando valores de Kappa de Cohen superiores a 0.75 y detección exitosa de sesgo entre grupos demográficos, proporcionan evidencia de que los sistemas de auditoría adversarial pueden lograr tanto confiabilidad como equidad a escala.

Varias limitaciones del trabajo actual sugieren direcciones para investigación futura. La dependencia de datos de resolución docente para el análisis de confiabilidad crea un problema de arranque en frío que requiere esfuerzo humano inicial antes de que la validación psicométrica completa sea posible. La implementación actual se enfoca en evaluaciones basadas en texto (Familia 1), con extensiones a otras modalidades de evaluación (cuestionarios, tareas de desempeño, portafolios) representando trabajo futuro importante. Se requerirán estudios longitudinales para establecer validez predictiva—el grado en que las calificaciones y retroalimentación generadas por IA predicen resultados académicos futuros.

Las implicaciones más amplias de este trabajo se extienden más allá de las contribuciones técnicas específicas. Al demostrar que la IA puede mejorar en lugar de reemplazar el juicio docente, el paradigma de auditoría adversarial ofrece un camino hacia la adopción de IA que respeta la experiencia profesional mientras aborda restricciones prácticas. Este enfoque puede informar el desarrollo de sistemas de IA en otros dominios donde el juicio experto permanece esencial pero podría beneficiarse del apoyo algorítmico.

En conclusión, QUANTA Grading representa un nuevo paradigma para la IA en evaluación educativa—uno que prioriza la autoridad docente, el fundamento pedagógico y la validez psicométrica mientras aprovecha la escalabilidad y consistencia de los sistemas de IA. A medida que las instituciones educativas continúan enfrentando desafíos de evaluación, los enfoques de auditoría adversarial ofrecen un marco principiado para la colaboración humano-IA que preserva lo esencial del juicio humano mientras aborda lo intratable de la escala humana.

Referencias

Paul Black and Dylan Wiliam. Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1):7–74, 1998. doi: 10.1080/0969595980050102.

Erik Brynjolfsson and Andrew McAfee. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. W. W. Norton & Company, New York, NY, 2014.

Gregory Camilli. Test fairness. In Robert L. Brennan, editor, *Educational Measurement*, pages 221–256. Praeger, Westport, CT, 4th edition, 2006.

David Carless. Differing perceptions in the feedback process. *Studies in Higher Education*, 31(2): 219–233, 2006. doi: 10.1080/03075070600572132.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960a. doi: 10.1177/001316446002000104.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960b. doi: 10.1177/001316446002000104.

Linda Crocker and James Algina. *Introduction to Classical and Modern Test Theory*. Cengage Learning, Mason, OH, 2nd edition, 2006.

Susan E. Embretson and Steven P. Reise. *Item Response Theory for Psychologists*. Lawrence Erlbaum Associates, Mahwah, NJ, 2000.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice. Uniform guidelines on employee selection procedures. Technical Report 43 FR 38295, Federal Register, Washington, DC, 1978.

Andy Hargreaves and Michael T. O'Connor. *Collaborative Professionalism: When Teaching Together Means Learning for All*. Corwin Press, Thousand Oaks, CA, 2018.

John Hattie. *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. Routledge, London, UK, 2009. doi: 10.4324/9780203887332.

John Hattie and Helen Timperley. The power of feedback. *Review of Educational Research*, 77 (1):81–112, 2007a. doi: 10.3102/003465430298487.

John Hattie and Helen Timperley. The power of feedback. *Review of Educational Research*, 77 (1):81–112, 2007b. doi: 10.3102/003465430298487.

Michael T. Kane. Validation. In Robert L. Brennan, editor, *Educational Measurement*, pages 17–64. Praeger, Westport, CT, 4th edition, 2006.

Avraham N. Kluger and Angelo DeNisi. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2):254–284, 1996. doi: 10.1037/0033-2909.119.2.254.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977a. doi: 10.2307/2529310.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977b. doi: 10.2307/2529310.

Michelle Meadows and Lucy Billington. A review of the literature on marking reliability. Research report, National Assessment Agency, London, UK, 2005.

Punya Mishra and Matthew J. Koehler. Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, 108(6):1017–1054, 2006. doi: 10.1111/j.1467-9620.2006.00684.x.

Ernesto Panadero, Anders Jonsson, and Juan Botella. Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review*, 22:74–98, 2017. doi: 10.1016/j.edurev.2017.08.004.

Les Perelman. When "machines" are the audience: A challenge to AI proponents of automated essay scoring machines. *Journal of Writing Assessment*, 7(1), 2014. URL https://journalofwritingassessment.org/article.php?article=73.

Mark D. Shermis and Jill Burstein. *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge, New York, NY, 2013. doi: 10.4324/9780203122761.

Patrick E. Shrout and Joseph L. Fleiss. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979. doi: 10.1037/0033-2909.86.2.420.

Lee S. Shulman. Those who understand: Knowledge growth in teaching. *Educational Researcher*, 15(2):4–14, 1986. doi: 10.3102/0013189X015002004.

Judith D. Singer and John B. Willett. *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press, New York, NY, 2003. doi: 10.1093/acprof:oso/9780195152968.001.0001.