Aprender con herramientas de inteligencia artificial multimodal

Introducción a los conceptos básicos de la inteligencia artificial

Autora: Kacsur Annamária, ITStudy

Revisora: Hartyányi Mária, ITSTudy

M1-T1-MLC6

Objetivo de aprendizaje

Las soluciones de inteligencia artificial más recientes han ido mucho más allá de la comunicación basada en texto. Los grandes modelos de lenguaje (ChatGPT, Gemini, Claude), desarrollados originalmente para generar texto, no solo son capaces ya de generar imágenes, vídeo y audio de calidad cada vez mayor, sino también de procesar información digital procedente de cualquier fuente externa, como sensores. Los modelos con estas capacidades ampliadas se conocen en conjunto como sistemas de **inteligencia artificial multimodal**.

Este material didáctico analiza las oportunidades que la inteligencia artificial multimodal ofrece a la educación y en qué va más allá de los materiales didácticos multimedia que se han generalizado en las últimas décadas.

¿Cómo aprendemos?

Los científicos llevan mucho tiempo estudiando la naturaleza del aprendizaje, pero todavía solo entendemos en parte cómo aprenden los niños pequeños y cómo las variadas experiencias que adquieren al explorar su entorno se convierten en un conocimiento coherente. Cuando un niño pequeño juega con una pelota, la vista, el tacto, el movimiento y las palabras asociadas se unen en una experiencia común. La comprensión no surge de la suma de estos estímulos, sino de reconocer las conexiones entre ellos.



¿Cómo aprenden los bebés? (Fuente: ChatGPT)

En los niños mayores y en los adultos, el aprendizaje se produce de forma similar. No percibimos el mundo a través de "canales" separados: vemos, oímos, actuamos, hablamos y leemos, e interpretamos estas experiencias relacionándolas entre sí y con nuestros conocimientos previos. Esto concuerda con la teoría constructivista del aprendizaje, según la cual el alumnado no se limita a recibir la información que se le transmite, sino que la interpreta a la luz de sus propias experiencias y conocimientos previos, establece conexiones y le da sentido por sí mismo.

En la década de 1990, los investigadores estudiaron cada vez más cómo los distintos modos de expresión y representación favorecen la comprensión. De ahí surgió la teoría de la **multimodalidad**[1], y fue en este periodo cuando aparecieron por primera vez en la educación los conceptos de **aprendizaje multimodal** y **pedagogía multimodal**, que subrayan que el conocimiento también puede construirse reconociendo y desarrollando conexiones entre diferentes representaciones.

De la multimedia a la multimodalidad

El lenguaje ha sido durante mucho tiempo el principal medio de enseñanza, en particular el texto escrito y el habla: el docente explica, mientras que el alumnado lee, toma apuntes y después demuestra sus conocimientos oralmente o por escrito.

[1] En la palabra multimodalidad, "multi" es un prefijo de origen latino que significa muchos o múltiples; "modalidad" deriva del latín modus (modo) y, en este contexto, se refiere a un modo de representación, comunicación o expresión.



Educación tradicional: el docente explica; los estudiantes leen, toman apuntes y después exponen oralmente o por escrito (Fuente: ChatGPT)

Con el desarrollo de las tecnologías digitales, la **multimedia** (el uso combinado de texto, imágenes, audio, vídeo y animación) se ha vuelto cada vez más accesible. Un "**modo**" o "**modalidad**" se refiere a la forma en que se presenta la información o al tipo de datos que entran en un sistema. Algunos ejemplos son el texto, las imágenes, el audio, el vídeo y los datos procedentes de sensores externos, como la temperatura media diaria registrada por una estación meteorológica.

Los sistemas digitales pueden trabajar con:

- texto;
- imágenes y gráficos;
- habla, música y otros sonidos;
- vídeo;
- datos que describen movimientos y gestos;
- datos de sensores, como mediciones de posición, temperatura o velocidad.

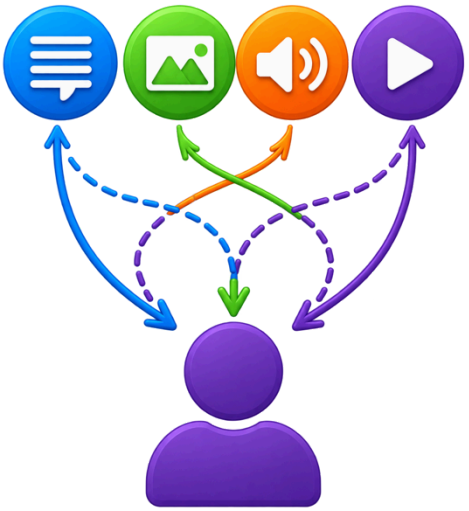
Esto no significa que una modalidad sea lo mismo que un formato de archivo. JPEG y PNG, por ejemplo, son formatos de imagen diferentes, pero ambos pertenecen a la misma **modalidad visual**.



Multimedia en la enseñanza: uso conjunto de varios medios; la información se presenta en paralelo.

La experiencia de las últimas décadas ha demostrado que los CD educativos interactivos y las presentaciones coloridas enriquecidas con imágenes, animaciones y sonidos favorecen el aprendizaje, pero no suponen una mejora cualitativa de la comprensión.

Hoy en día, la mayoría de las clases siguen centrándose principalmente en el texto y el habla (libros de texto, explicaciones a todo el grupo y cuadernos).

	
Presentación a través de varios canales	**Construcción de significado relacionando distintas representaciones**
MULTIMEDIA	MULTIMODALIDAD

Si el alumnado sigue recibiendo principalmente información ya elaborada y relacionar los significados que transmiten los distintos medios en una explicación depende por completo de la iniciativa del docente, los elementos multimedia comunican el mismo mensaje sin contribuir a la comprensión.

> En un enfoque multimodal de la enseñanza, el alumnado no solo recibe información en varias formas, sino que también compara, relaciona, transforma y crea por sí mismo distintas representaciones. Así, el énfasis pasa de la transmisión de información mediante recursos multimedia a la construcción multimodal de significado por parte del alumnado.



Construcción multimodal del conocimiento (Fuente: ChatGPT)

La fórmula de una función, su gráfica y una historia que describe un movimiento real no son simplemente tres "formatos" distintos del mismo contenido de aprendizaje. El alumnado necesita reconocer cómo se reflejan en la gráfica los cambios descritos en la historia y cómo pueden expresarse mediante relaciones matemáticas. La comprensión se desarrolla a medida que se reconocen las conexiones entre las representaciones.

Inteligencia artificial multimodal

La inteligencia artificial multimodal acepta como entrada distintos tipos de datos (texto, imágenes, audio y vídeo) y, a petición, genera distintos tipos de salida.



Inteligencia artificial multimodal (Fuente: ChatGPT)

La **inteligencia artificial multimodal**[2] puede procesar e interpretar varios tipos de datos (texto, imágenes, audio, vídeo y otros) dentro de un único modelo. El término "multimodal" hace referencia a la variedad de datos de entrada y de salida que el modelo puede manejar.

No solo combina distintos formatos multimedia, sino que también "aprende" las relaciones entre ellos.

Ámbitos de aplicación

Sanidad

La inteligencia artificial puede analizar radiografías o resonancias magnéticas junto con las notas clínicas escritas de un paciente, identificar patrones interpretando estas fuentes conjuntamente y utilizarlos para formular un posible diagnóstico.

[2] **Codificación de modalidades**: cada tipo de datos de entrada se procesa primero mediante un codificador especializado. El codificador de imágenes convierte los elementos de la imagen (píxeles) en datos numéricos, el codificador de lenguaje transforma el texto en tokens y el codificador de audio extrae las características importantes del habla o del sonido. Estos codificadores captan las características esenciales de cada modalidad y filtran los detalles irrelevantes.

Integración multimodal (fusión): el modelo combina la información de varias modalidades para formar una representación unificada de los datos de entrada. Los modelos fundacionales multimodales suelen lograr esta integración mediante capas transformer compartidas y mecanismos de atención, lo que permite que la información de una modalidad influya en la interpretación de otra. Por ejemplo, al interpretar la expresión facial de una persona junto con su tono de voz y sus palabras, el modelo puede inferir emociones o intenciones.



IA multimodal en la sanidad (Fuente: ChatGPT)

Atención al cliente y accesibilidad

La IA multimodal permite a los usuarios con discapacidad visual o movilidad reducida hacer preguntas oralmente y solicitar las respuestas en el modo que mejor se adapte a sus necesidades.

Creación de contenidos

Modelos como ChatGPT y Gemini utilizan entradas combinadas de texto e imagen para modificar una imagen o crear nuevos contenidos creativos.

Robótica y sistemas autónomos

Los vehículos autónomos y los sistemas robóticos utilizan la inteligencia artificial multimodal para integrar las imágenes de las cámaras y los datos espaciales, formarse una representación de su entorno y tomar decisiones de navegación en condiciones reales.

Análisis de documentos

Cuando recibe un documento que contiene gráficos, diagramas e imágenes insertadas, la inteligencia artificial multimodal procesa todos los elementos visuales junto con el texto y responde a las preguntas interpretándolos conjuntamente.

Ejemplos de la formación profesional

En un proyecto medioambiental, el alumnado sube a un sistema de IA un vídeo que muestra el estado de un río y un documento con los resultados de unas mediciones. La IA ayuda a establecer la relación entre los fenómenos visibles en el vídeo y las mediciones, así como las conclusiones que pueden extraerse al examinar ambas fuentes conjuntamente.



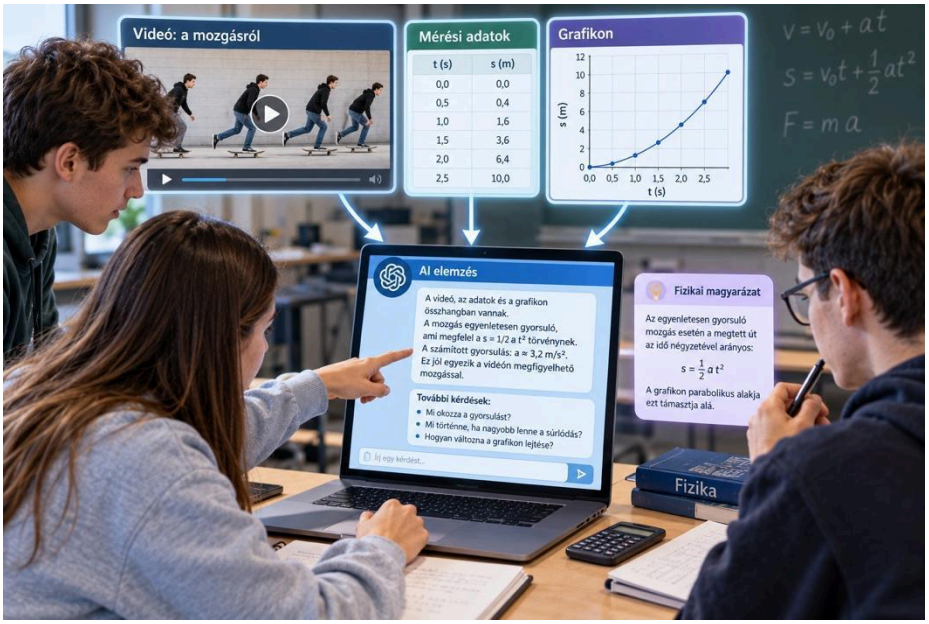
IA multimodal en un proyecto medioambiental (Fuente: ChatGPT)

En una clase de historia, los estudiantes pueden pedir a la IA que responda a sus preguntas analizando conjuntamente una fotografía, un cartel y una fuente histórica escrita de la época.



IA multimodal en una clase de historia (Fuente: ChatGPT)

En física, el alumnado puede examinar el vídeo de un movimiento junto con las mediciones y la gráfica correspondientes: ¿coincide la gráfica con lo que vemos en la grabación?, ¿cómo puede explicarse la relación mediante las leyes de la física?



IA multimodal en una clase de física (Fuente: ChatGPT)

Los estudiantes de formación profesional suben una fotografía o un vídeo de un componente defectuoso junto con su documentación técnica y los datos de medición. Al analizar las fuentes conjuntamente, la IA puede responder a las preguntas de los estudiantes sobre las posibles causas de la avería.



IA multimodal en el aula (Fuente: ChatGPT)

Ideas clave para docentes

Las aplicaciones de IA actuales permiten a los docentes crear de forma rápida y sencilla imágenes, vídeos, presentaciones e infografías llamativos. Elaborar materiales didácticos resulta más fácil, su rápida obsolescencia deja de ser un gran problema y adaptarlos a las necesidades de cada estudiante puede hacerse realidad.

La IA multimodal puede provocar un cambio más profundo si animamos al alumnado a comparar contenidos creados por sí mismo o generados con IA, a pedir que se analicen, a solicitar retroalimentación sobre sus ideas y a pedir que se justifiquen las respuestas recibidas.

La IA multimodal abre nuevas posibilidades para el aprendizaje activo. Con la orientación del docente, puede ayudar al alumnado a convertirse en participante activo en su propio aprendizaje. La función del docente es diseñar y crear un entorno de aprendizaje en el que la IA multimodal impulse al alumnado a pensar, interpretar, comparar y construir su propio significado. Hoy en día, estas tecnologías pueden reducir el predominio del procesamiento de la información basado en texto y acercar la educación a la verdadera naturaleza del aprendizaje.

Referencias

- Simamkele Matuntuta (2026), How AI is Transforming Vocational Education and Training, artículo en cloudassess.com (consulta: 3/7/2026)
- Ayush Chauhan (2025): Multimodal AI: Beyond Text and Images, artículo en https://www.techaheadcorp.com,
- Kamalmeet Singh (2026): How Multimodal AI Works?, artículo de blog en https://medium.com, consulta: 8/7/2026
- Amol Yadhav (2024): Multimodal AI: The Next Frontier in Personalized Education, artículo de blog en https://medium.com/, consulta: 8/7/2026
- Peraschi Selvan Subramanian (2025): Multimodal AI: The future of integrated intelligence, artículo publicado en World

Journal of Advanced Engineering Technology and Sciences, consulta: 9/7/2026

Enlaces

IA multimodal https://www.llamaindex.ai/glossary/multimodal-ai

¿Cómo funcionan los modelos de IA multimodal? Vídeo de YouTube:

https://www.youtube.com/watch?v=WkoytIA3MoQ

Derechos de autor © 2026 Consorcio VITA

Esta obra puede compartirse y adaptarse bajo los términos de la Licencia Internacional Creative Commons Atribución/Reconocimiento-CompartirIgual 4.0 (CC BY-SA 4.0).

Financiado por la Unión Europea. Las opiniones expresadas son únicamente las de los autores y no reflejan necesariamente las de la Unión Europea ni las de la Agencia Nacional.