Multimodal AI explained: text, image and audio in one system

26 October 2026 · 4 min read · blogs.iictp.org/multimodal-ai-explained

A multimodal model takes more than one kind of input, such as text with images or audio, and can sometimes produce more than one kind of output. It lets a single system read a chart, describe a photo or answer a spoken question.

TL;DR The short version

- A multimodal model takes more than one kind of input, such as text with images or audio.
- Typical uses are reading documents with tables and diagrams and describing images for accessibility.
- Input quality drives accuracy, so verify extracted numbers and test extraction and description separately on real inputs.

Read online, listen to it, or change the text size: https://blogs.iictp.org/multimodal-ai-explained

Typical uses

- Reading documents with tables, forms and diagrams.
- Describing images for accessibility.
- Transcribing and summarising audio.
- Answering questions about a screenshot.

Design points

- **Check the input quality.** Blurry or cropped images and noisy audio lower accuracy.
- **Verify the reading.** Extracted numbers from a chart or form should be checked.
- **Mind the privacy.** Images and audio can contain people and personal data.
- **Cost.** Images and audio often cost more to process than text.

Testing

Build an evaluation set from your real inputs, including bad ones, and score extraction and description separately. See how to evaluate an LLM application.

Related credential: AI Practitioner. Explore the exam facts and syllabus.

Sources and further reading

1. IICTP. AI Practitioner credential page. iictp.org/certifications/aip.

Published by IICTP, an independent technical credential body, on 26 October 2026. The current version of this post is at https://blogs.iictp.org/multimodal-ai-explained. More guides: blogs.iictp.org.