SKILLNAVORA · AI PRODUCT MANAGEMENT & GENAI LAUNCHPAD

Multimodal AI & AI Media

Use AI that sees, hears and creates: vision understanding, image generation, video and voice, multimodal product use cases, and rights, safety, deepfakes and content labelling.

Edition 2026 · skillnavora.in

How to use this guide

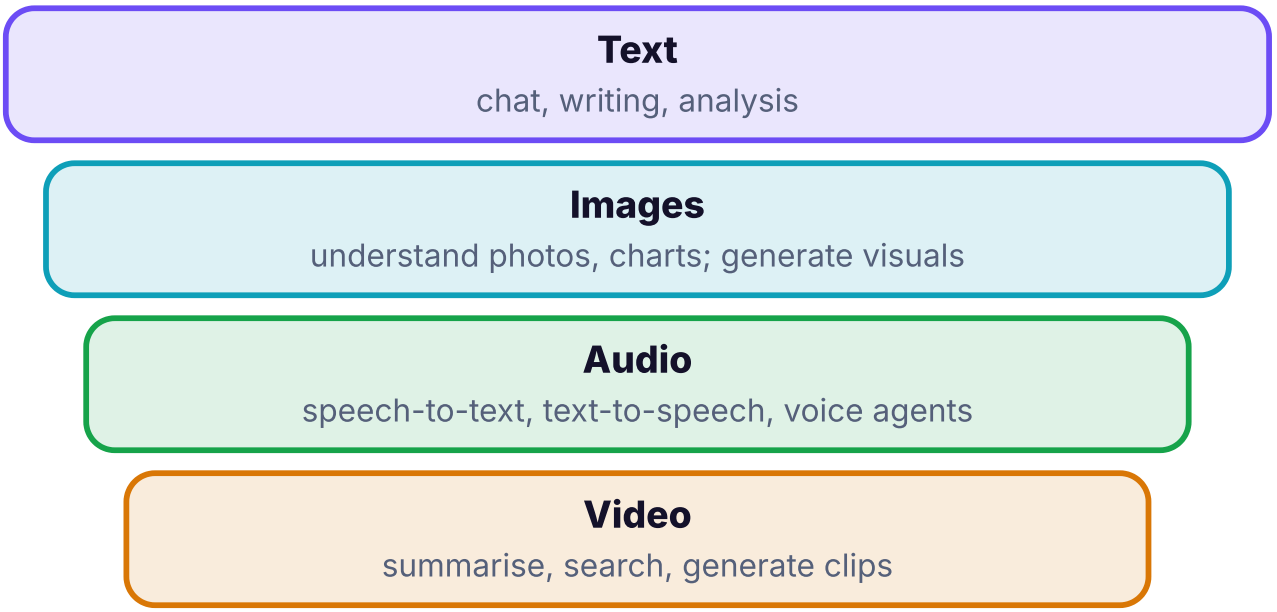
Each class has goals, concepts with Indian examples, key takeaways and a practice exercise, from basic to advanced. Finish with the final project to build something you can show in interviews.

	Class	Level
Class 1	**Multimodal basics**	Basic
Class 2	**Vision understanding use cases**	Basic
Class 3	**Image generation in products**	Intermediate
Class 4	**Voice and video**	Intermediate
Class 5	**Rights, safety, deepfakes and labelling**	Advanced
Class 6	**Final project**	Advanced
	Toolkit: templates, cheat sheet, glossary, 4 interview questions, further reading	

BASIC · CLASS 1

Multimodal basics

Know what modern models can see, hear and create.



Modalities

Key takeaways

- Models now take and produce many modalities.
- Choose the modality users already use.
- Voice matters in India (literacy, languages).

Practice exercise: List 5 features where a photo or voice input beats typing.

BASIC · CLASS 2

Vision understanding use cases

Use AI to read and interpret images.

- OCR of invoices, IDs (with consent), forms; handwritten notes.
- Chart and screenshot understanding; UI review.
- Quality inspection (manufacturing), crop disease, damage assessment (insurance).
- Accessibility: describing images for visually impaired users.

Key takeaways

- Vision unlocks offline-to-digital workflows.
- Mask personal data.
- Measure accuracy on your own images.

Practice exercise: Design an AI feature that reads shop receipts for expense tracking.

INTERMEDIATE · CLASS 3

Image generation in products

Use generative images responsibly.

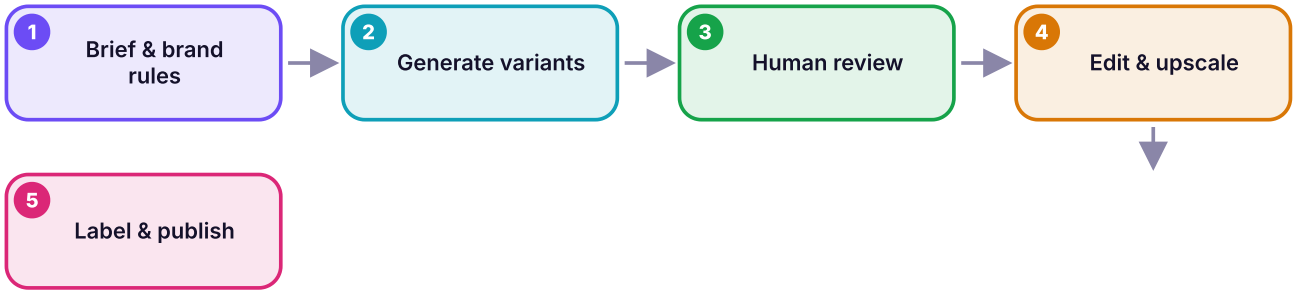


Image generation workflow

Tools: Midjourney, Adobe Firefly, Ideogram, Leonardo, DALL·E, Imagen, Canva AI. Use cases: ads, product mock-ups, thumbnails, illustrations.

Key takeaways

- Keep brand rules in the brief.
- Human review before publishing.
- Prefer tools with clear commercial-use terms.

Practice exercise: Create a brand-safe image brief template.

INTERMEDIATE · CLASS 4

Voice and video

Build with speech and video AI.

- Speech-to-text and text-to-speech in many Indian languages; voice agents (see Voice Agents lesson).
- Video: AI summaries and search of recordings; avatar and generative video tools (Runway, Veo, Sora, HeyGen, Synthesia) for explainers.
- Latency, accents, background noise and cost are key constraints.

Key takeaways

- Test with real accents and noise.
- Video AI is great for explainers and summaries.
- Watch latency and cost.

Practice exercise: Plan a Hindi voice FAQ assistant for a clinic.

ADVANCED · CLASS 5

Rights, safety, deepfakes and labelling

Manage the risks of AI media.

- Copyright and licences: check tool terms for commercial use; avoid copying artists or brands.
- Consent for faces and voices; never clone voices without permission.
- Label AI-generated media; content credentials (C2PA) and watermarking.
- Moderation of user-generated AI media; follow Indian advisories on deepfakes.

Key takeaways

- Consent for likeness and voice.
- Label AI media.
- Check commercial-use rights.

Practice exercise: Write an AI media policy for a marketing team.

ADVANCED · CLASS 6

Final project

Ship a multimodal feature.

Design a multimodal feature (photo receipts, voice ordering, or video course summaries): user flow, model/tool choice, accuracy eval on 30 samples, privacy and consent, labelling, cost estimate.

Key takeaways

- Evaluate on real samples.
- Design for consent and privacy.
- Estimate cost per use.

Practice exercise: Complete the multimodal project.

TOOLKIT

Templates

AI media brief

Field	Fill in
Purpose & audience	
Brand rules	
Tool & licence	
Consent needed?	
Review steps	
Labelling	
Accessibility (alt text, captions)	

TOOLKIT

Cheat sheet

Term / tool	Meaning / use
OCR	Reading text from images
TTS / STT	Text-to-speech / speech-to-text
C2PA	Content credentials standard for media provenance
Deepfake	Realistic fake media of real people

Glossary

Watermark	Hidden or visible mark identifying AI content

INTERVIEW PREP

4 interview questions with answers

Q1. What is multimodal AI?

Models that understand and/or generate several types of data: text, images, audio and video.

Q2. A good vision use case?

Reading receipts or forms from photos to remove manual entry, with accuracy checks and privacy.

Q3. Risks of AI media?

Copyright, consent for likeness and voice, deepfakes, misinformation; mitigate with consent, labelling, C2PA and moderation.

Q4. Why voice in India?

Many users prefer speaking, many languages and literacy levels; voice widens access.

LEARN MORE

Further reading and tools

- C2PA content credentials: https://c2pa.org/
- Google AI Studio (Gemini multimodal): https://aistudio.google.com/
- Adobe Firefly: https://www.adobe.com/products/firefly.html
- ElevenLabs docs: https://elevenlabs.io/docs
- Runway Academy: https://academy.runwayml.com/

© SkillNavora · skillnavora.in · Part of the AI Product Management and GenAI Launchpad programs. Practise with the free AI Tutor, AI PM Toolkit and Practice Lab on skillnavora.in.