

Our choice was based on **ROI (Return on Investment) and Accuracy**, not the lowest per-page price.

Factor	Traditional OCR	Gemini 2.5 Flash (VLM)	ROI Justification
Accuracy Target	Low (30-60%)	High (85%+)	VLM is required to solve the 40-80% inaccuracy problem in provider directories.
Output	Raw text dump	Structured JSON	VLM directly outputs the standardized data format , eliminating hours of manual coding and validation.
Administrative Cost	High (Requires human review and significant programming)	Low (Automated processing)	The high accuracy of VLM reduces the \$2.4M annual administrative losses tied to manual review and rework.
Complexity	Fails on multi-column forms	Handles visual layout and context (multimodal)	VLM can handle the complex, unstructured Molina Healthcare form used in the demo.

Why Gemini 2.5 Flash?

Feature	Gemini 2.5 Flash (Our Choice)	Gemini 2.5 Pro	Gemini 2.5 Flash-Lite
Primary Use Case	High-Volume VLM & Extraction	Complex Reasoning & Coding	High-Throughput Text Tasks
Input Price (per 1M)	\$0.30	\$1.25	\$0.10
Output Price (per 1M)	\$2.50	\$10.00	\$0.40
Visual Reasoning	Excellent (79.7% MMMU)	Superior (82.9% MMMU)	Good (72.9% MMMU)
Context Window	1 Million Tokens	1 Million Tokens	1 Million Tokens
Thinking Budget	Controllable	Adaptive (Fixed High)	Minimal (Off by default)
Latency	Very Low	Medium	Blazing Fast