PDF Days Europe 2025

Empowering the Future of PDF with AI

Leveraging AI for Enhanced Accessibility and Content Transformation

Matthew Hardy | Director of Engineering | Adobe

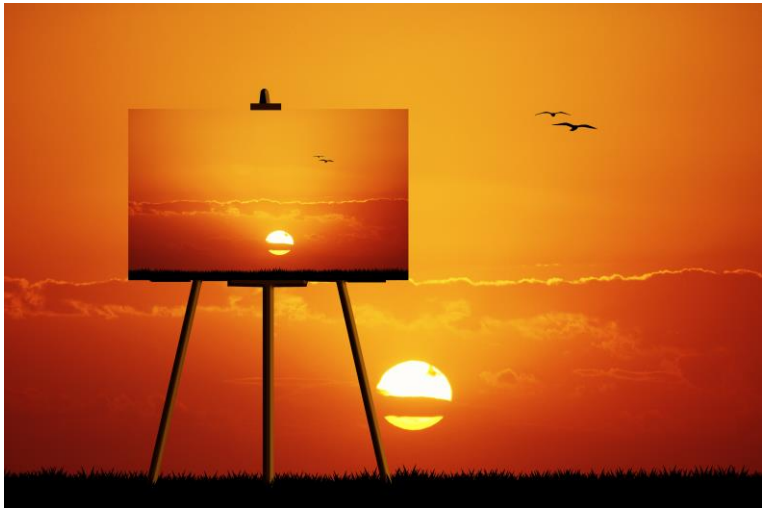
PDF association

Challenges facing PDF

“Print to PDF”

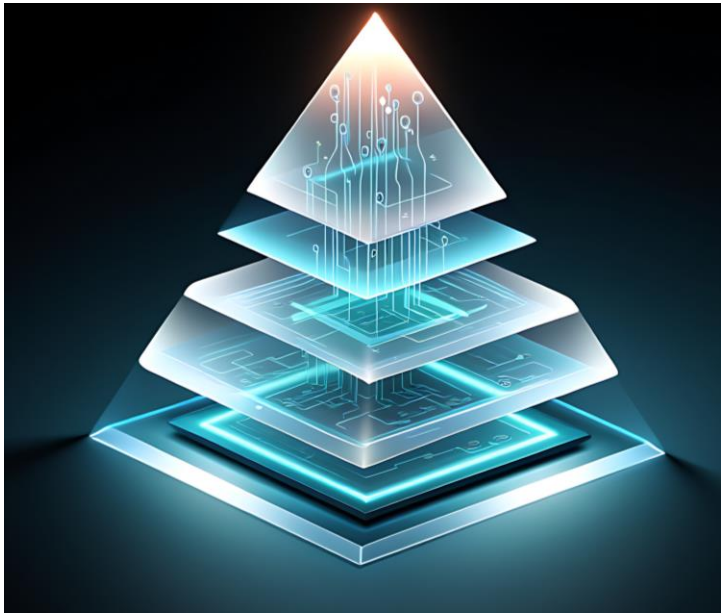
PDF as Page Independent Format with Fixed Layout

- Early versions of PDF
  - Reliable exchange of visual information
  - Screen-oriented at first
  - Rich print capabilities fast followed
- Fixed layout and page independence
  - Optimized for computers of the 90s
  - Pages don’t know about each other
  - Fixed “final form” layout reproducible everywhere



Marked, Structured and Tagged PDF

- Apparent more was required
  - Marked PDF (PDF 1.2)
  - Structured PDF (PDF 1.3)
  - Tagged PDF (PDF 1.4)
- Foundation of semantics and rich data
  - Each built on the last, but not replacements
  - Marking for embedded information
  - Structure for rich semantic exchange
  - Tagged PDF for interchangeable semantics



Print to PDF (Distillation)

- PDF through the Print Driver
  - Ease of Creation
  - Great compatibility (PostScript)
  - Dominant path to creation
  - Zero structure, semantics or data
- Rich Creation paths
  - Much more limited
  - Value not obvious to end users
  - PDF often considered just for display/print

Revolutionary, yet limited

- PDF changed the World
  - Trillions of PDFs today
  - Dominant format
- Unstructured (or poorly) and static
  - Less than 20% of PDF is Tagged
  - Less than 5% is Tagged at a high quality
  - Rich data and content locked inside the visual format
  - Unless…

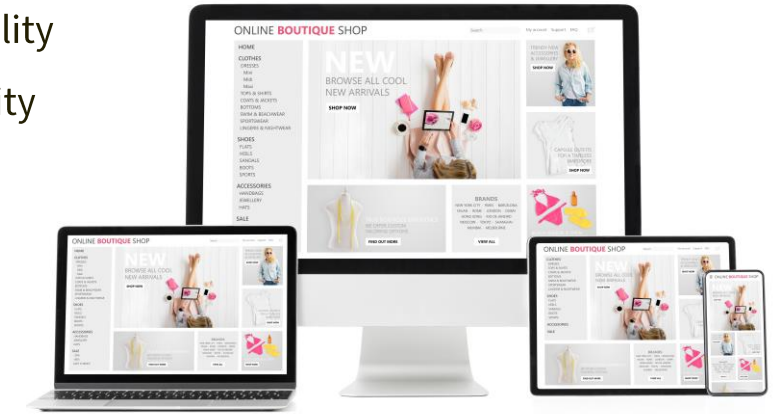
PDF association

AI for Responsive PDF

“The early days”

Early focus on Responsive Layout

- PDF’s fixed layout presents challenges
  - No responsive layout for mobile or tablet (or even large screens)
  - Zoom requires pan and scroll for Accessibility
  - Fonts and color hard to adapt for readability
- Tagged PDF intended as solution
  - Empower adaptability
  - Repurpose and reuse content
  - Road to ubiquitous Tagging never-ending
  - Could it be retrofitted through AI?

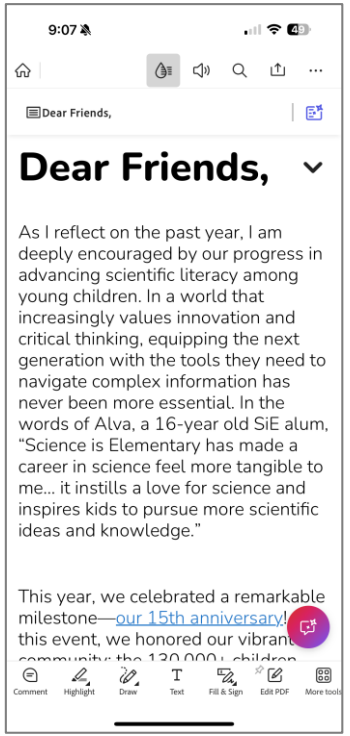


Using classic AI to understand documents

- Object detection model (YOLO Vision Models)
  - Identify the building blocks of a document
  - Paragraphs, lists, tables, figures, headings, etc.
  - Fast processing of pages
- Classic heuristics
  - Assemble the building blocks into an order
  - Break the blocks down into smaller pieces
    - List items, words, table cells, etc.
  - Create a rich tagged PDF

Derivation to HTML (Common Use Case)

- PDF Content Transformed
  - Converted to HTML
  - Original maintained AND connected
- Responsive Web
  - Enable responsiveness to screen size
  - Magnification and reflow
  - CSS can adapt the color and contrast
  - Retain stylistic feel of original
- Workflows
  - Can be maintained
  - Examples include review and commenting



Data Extraction

- Extraction of Content
  - Into structured databases and CMS
  - Reusable for workflows
  - Components of the document
- Structured Data
  - Data locked in tables
  - Extractable and repurposable

Incredible Success, but Data Intensive

- Opportunities are massive
  - Readability improvements for all
  - Accessibility for everyone (with some limitations)
  - Born-accessible PDF is still the gold standard
- Development of models
  - Data intensive and extremely costly
  - Millions of pages of labelled data
  - Brittle to new requirements

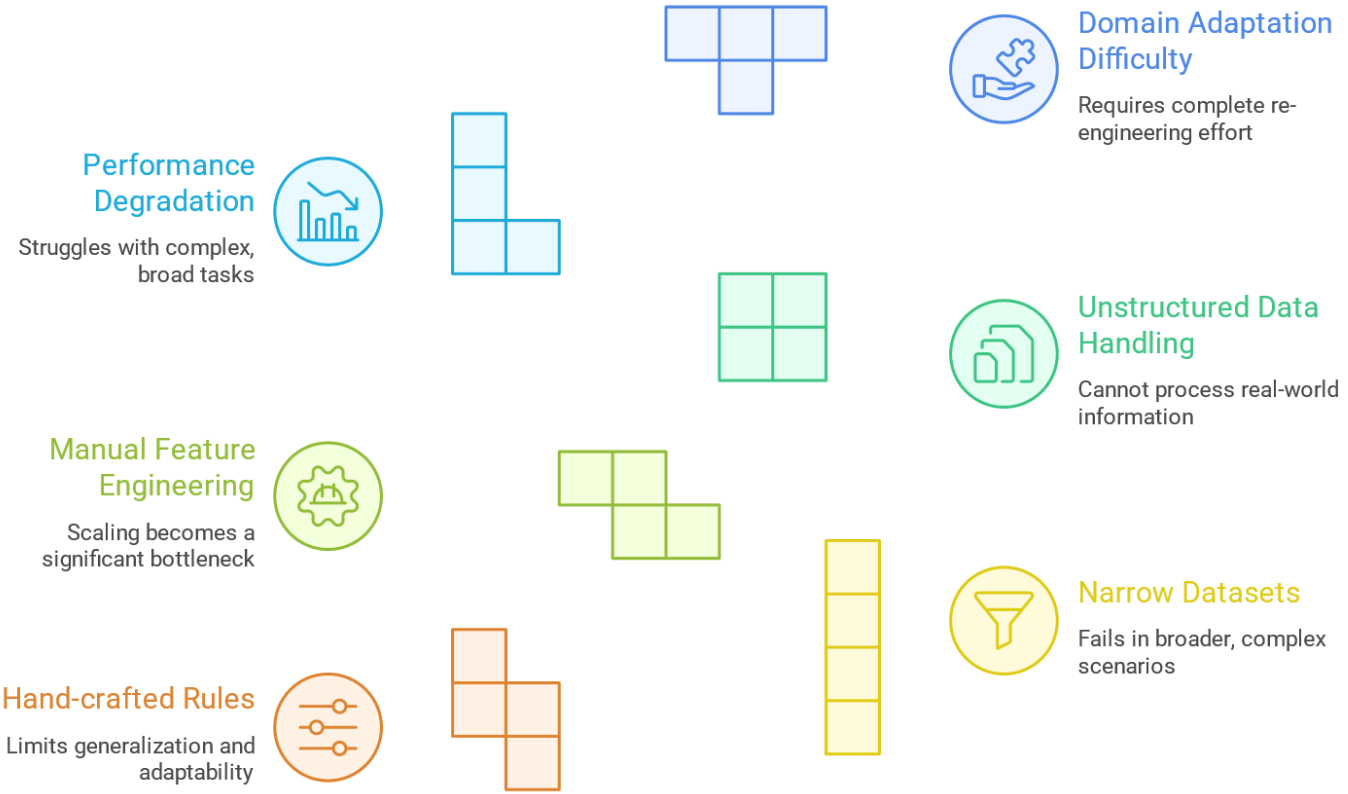
- Everything to this point is just restoring missing information!

PDF association

Beyond Classic ML — LLMs

“Generative AI”

Classic Model Limitations



Performance Degradation

Struggles with complex, broad tasks

Manual Feature Engineering

Scaling becomes a significant bottleneck

Hand-crafted Rules

Limits generalization and adaptability

Domain Adaptation Difficulty

Requires complete re-engineering effort

Unstructured Data Handling

Cannot process real-world information

Narrow Datasets

Fails in broader, complex scenarios

Large Language models

What they are

Trained on text, learn patterns

How they work

Neural networks, predict next word

Why they matter

Summarize, translate, create

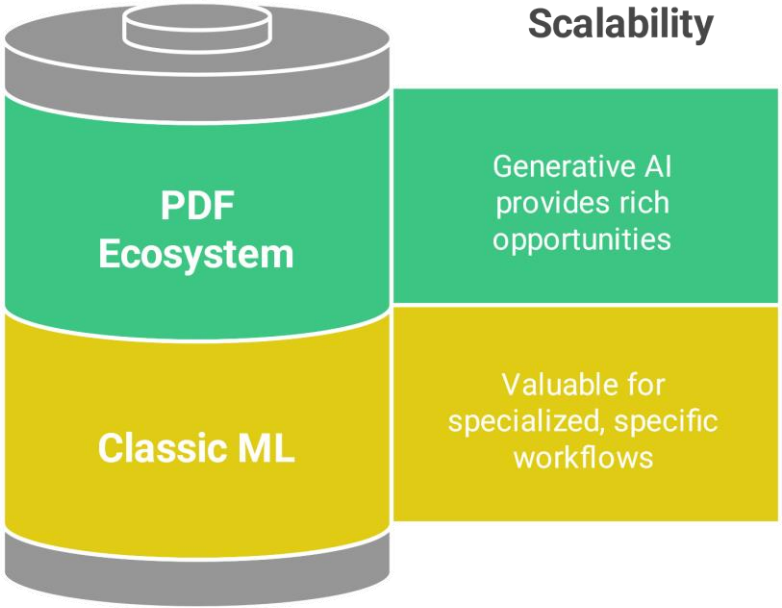
Impact

Assist coding, research, analysis

PDF association

PDF as an Ecosystem vs Specialists

- Classic ML
  - Very valuable for specialized workflows
  - Highly leveraged for specific workflows
  - Still very valuable tool for specific use cases
- PDF as an Ecosystem
  - Users represent vast sets of different use cases
  - Classic ML just doesn't scale to broad use cases
  - Generative AI provides so many rich opportunities



PDF association

Empowering the Future of PDF with AI

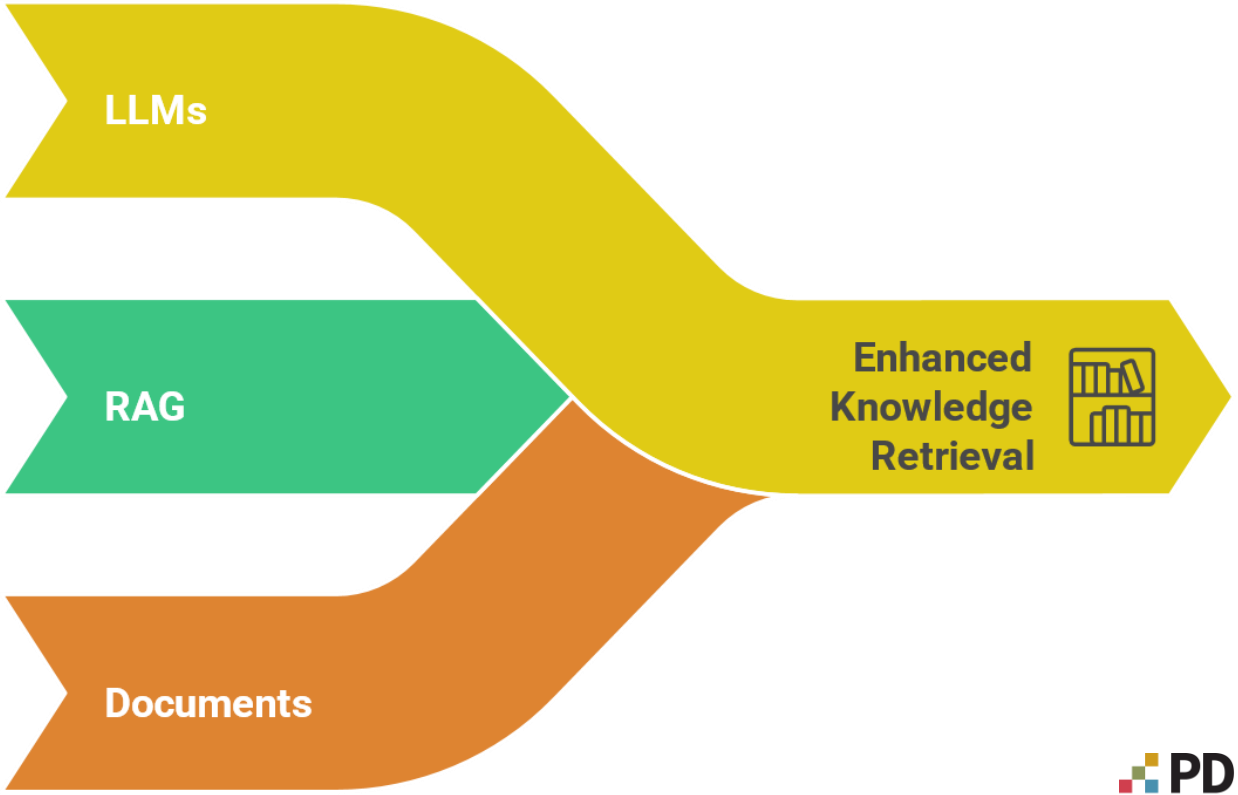
“And maybe documents in general…”

Documents as Knowledge and Context

1 Vast knowledge sources but lack personalization.

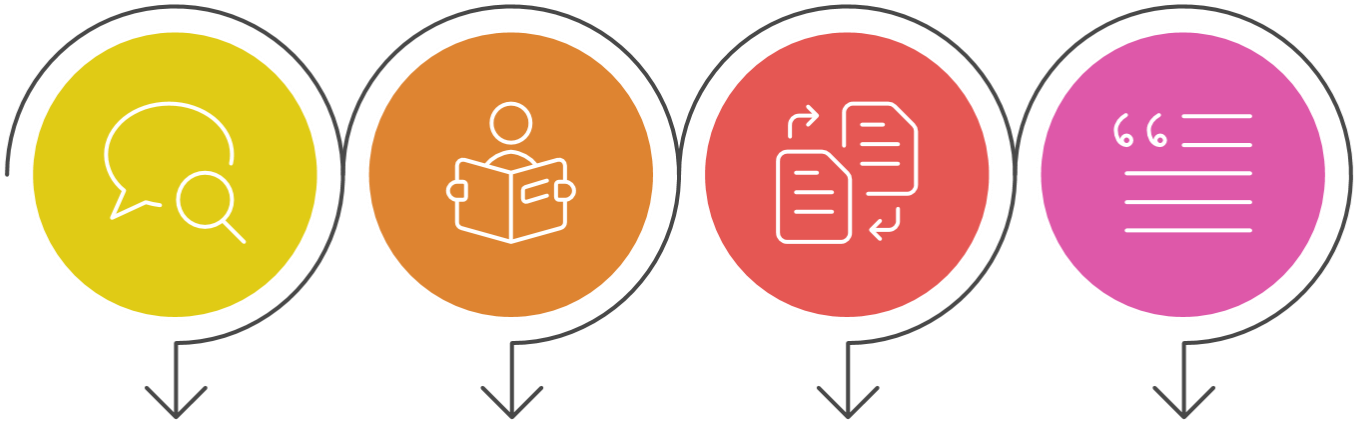
2 Integrates external knowledge for better context.

3 Provides specific context for relevant answers.



PDF association

Conversational PDF: From Static to Interactive



Dynamic Q&A

Enables interactive question answering within documents.

Tailored Summaries

Provides summaries customized for different user roles.

Combined PDFs

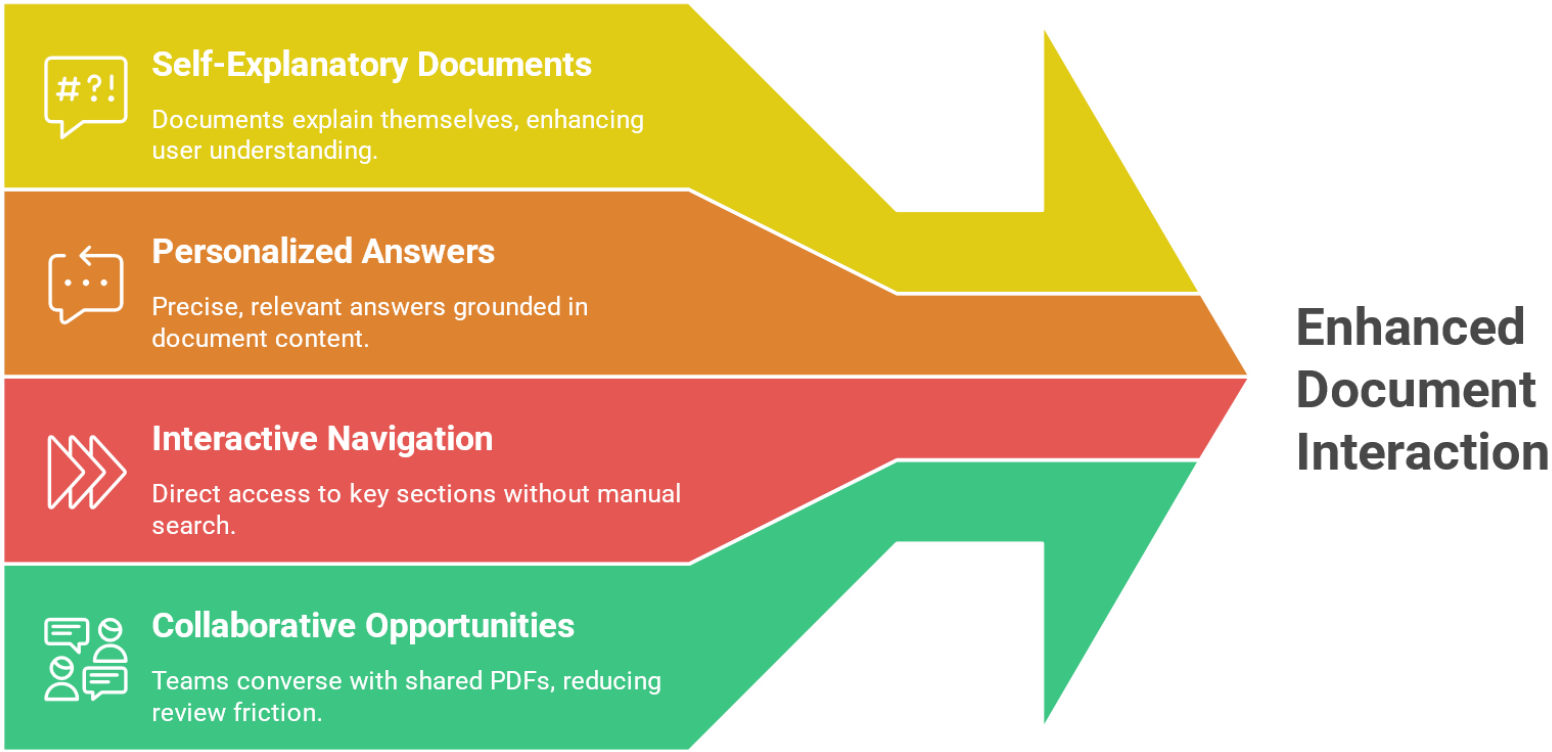
Allows interaction across multiple PDF documents.

Provenance and Citations

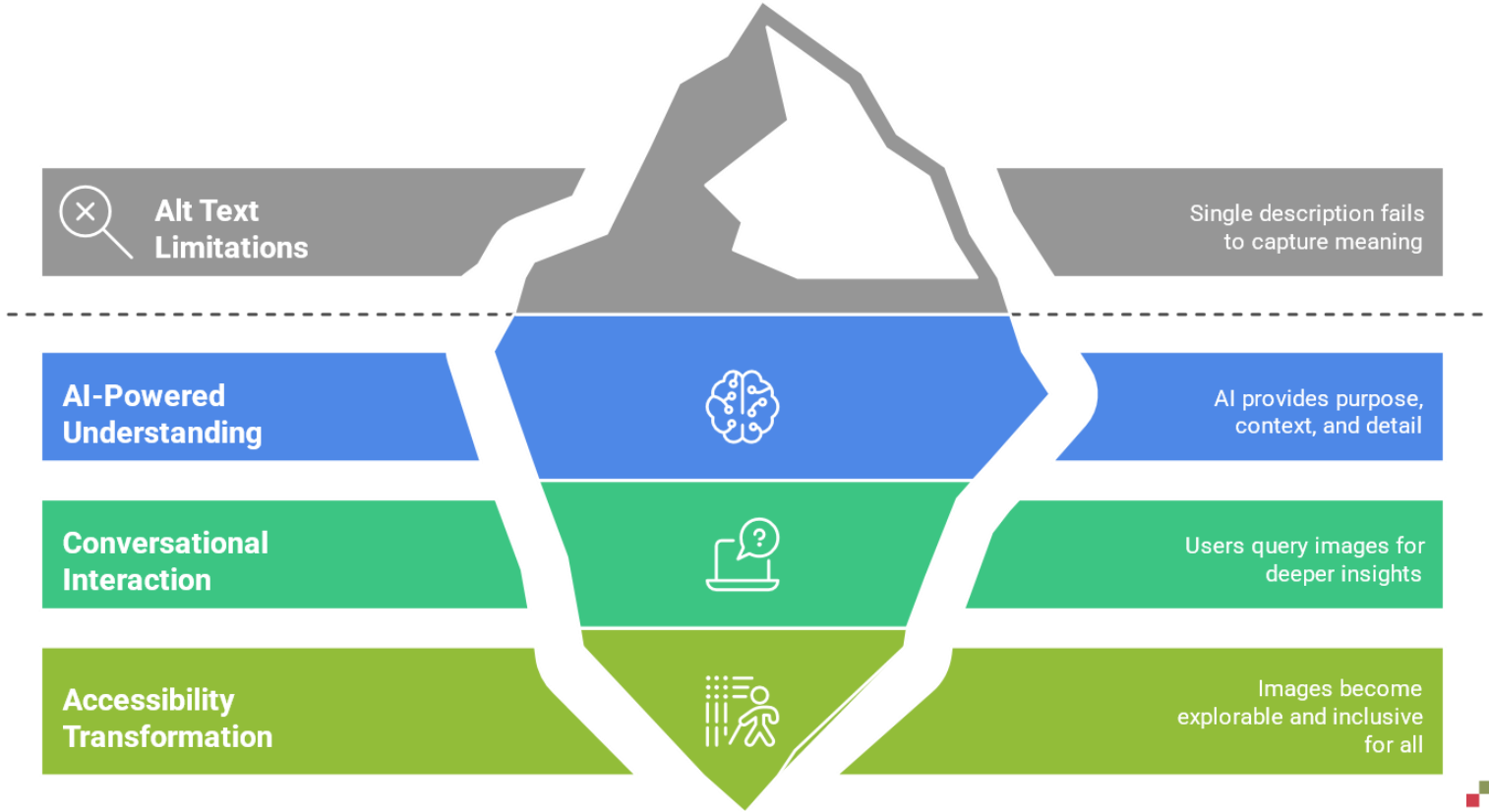
Maintains trust with passage references and citations.

PDF association

Conversational Workflows with PDF



Beyond Alt Text: Conversational Images



PDF association

Cross-Document Synthesis

- Combine multiple PDFs into cohesive reports
  - Merge sources into a unified narrative while identifying overlaps or gaps
- Summarize across domains and sources
  - Extract trends and insights from hundreds of pages in minutes
- Extract and integrate key findings into new narratives
  - Pull structured facts and references with citations for credibility
- Accelerate research, compliance, and due diligence workflows
  - Automates reviews and audits, reducing manual cross-checking effort

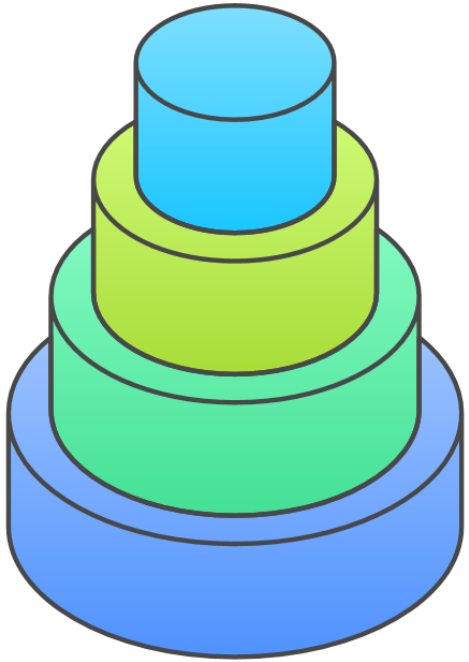
Transforming Data into Insight

AI-Driven Decision Support

Insights integrated into workflows, reducing manual effort

Visual Adaptation

Visuals tailored for different audience levels



Fidelity and Clarity

Source data verifiable, insights easier to grasp

Static Content Enrichment

Documents enriched with summaries and context

PDF association

Wrapping Up

“Bringing it together…”

Empowering the Future of PDF with AI

Knowledge Ecosystems

Connects documents to broader knowledge networks

AI Transformation of PDFs

Converts static PDFs into dynamic knowledge sources

AI-Powered Creativity

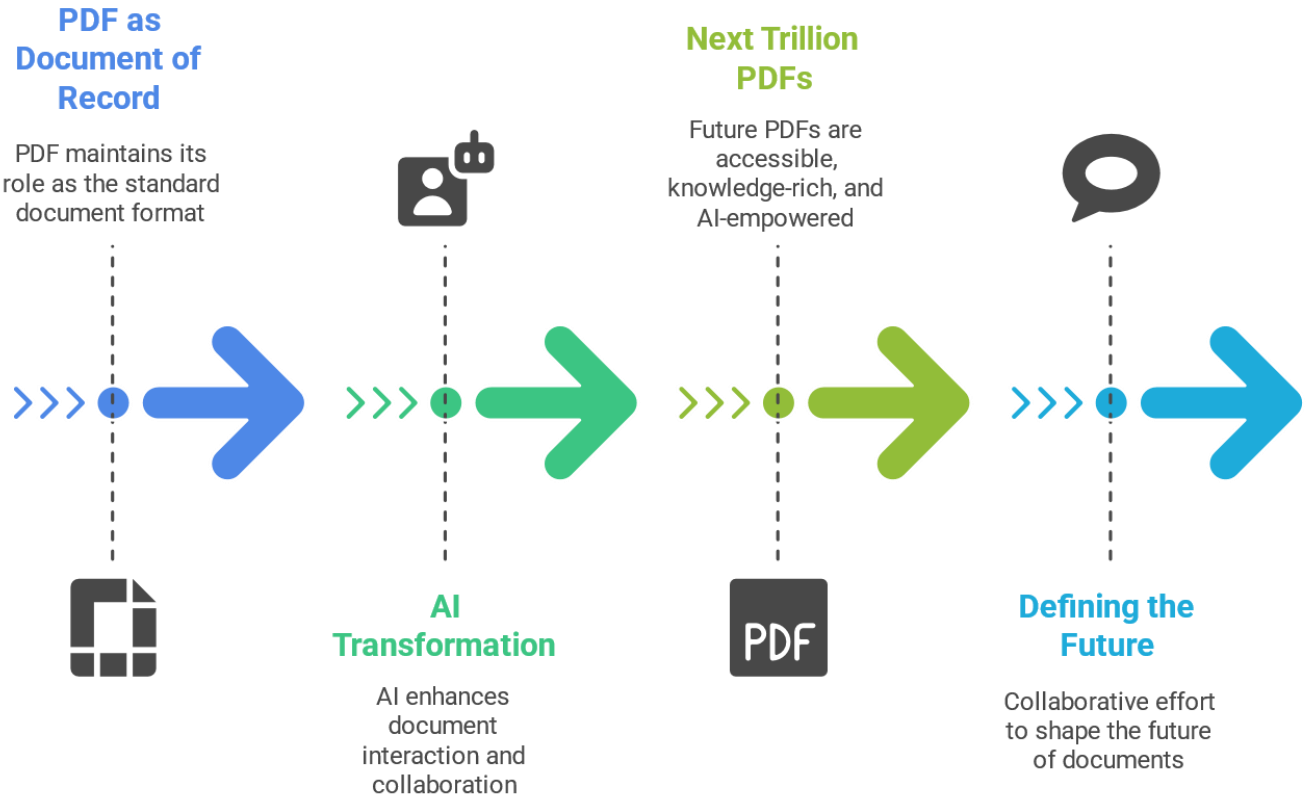
Generates new insights and creative content

Accessibility Workflows

Ensures documents are accessible to all users

PDF association

Creating the Next Chapter



PDF association

PDF association

The next trillion PDFs won’t just preserve knowledge — they’ll empower it.