KnowledgeLake

WHITE PAPER

Intuitive AI – Leading the Journey to 100% Accuracy in Intelligent Document Processing

By Bart Peluso III

Microsoft Partner
Gold Application Development
Gold Cloud Platform
Gold ISV
Microsoft

DIRECT 314.898.0500
TOLL-FREE 888.898.0555
WEBSITE knowledgelake.com
EMAIL info@knowledgelake.com

INTRODUCTION

Intelligent Document Processing (IDP) has revolutionized the way organizations manage and process vast amounts of unstructured data. From college transcripts and land records to invoices and driver's licenses, IDP systems can automate document handling with remarkable efficiency. However, the accuracy of these systems is paramount. Inaccurate data extraction can lead to compliance risks, financial losses, and operational inefficiencies. This whitepaper explores how the accuracy levels of IDP systems can be significantly enhanced through four key strategies: Day Zero Accuracy, Disambiguation, Honing AI Hints, and Retrieval-Augmented Generation (RAG). All four are elements of KnowledgeLake's unique Intuitive AI based approach to processing complex documents — be they handwritten forms, ancient land records or graphical government issued identification documents.

Microsoft Partner
Microsoft
Gold Application Development
Gold Cloud Platform
Gold ISV

KnowledgeLake

DIRECT 314.898.0500
TOLL-FREE 888.898.0555
WEBSITE knowledgelake.com
EMAIL info@knowledgelake.com

1. Day Zero Accuracy

DEFINITION & IMPORTANCE:

Day Zero Accuracy refers to the system's ability to achieve high levels of accuracy from the moment it is deployed, minimizing the need for initial manual corrections. In the context of IDP, this means that the system should accurately extract and process data from documents like college transcripts or invoices without extensive pre-training or configuration.

STRATEGIES FOR IMPROVEMENT:

High-Quality Pre-Training Data:
Utilizing a rich and diverse dataset during the pre-training phase can significantly improve the system's initial accuracy. For example, when processing college transcripts, the model should be exposed to a variety of transcript formats from different institutions to ensure it recognizes and accurately extracts data across all variations.

Advanced OCR Integration:
Integrating advanced Optical Character Recognition (OCR) technologies that are optimized for different document types can improve the system's ability to correctly interpret and extract data, particularly from complex documents like land records or handwritten notes on driver's licenses.

Example Application:

For example, a college admissions office utilizing IDP to process transcripts can achieve high accuracy from day one if the system is trained on a comprehensive dataset of transcripts from various schools, including those with non-standard formats. This reduces the need for manual corrections and accelerates the admission process.

2. Disambiguation

DEFINITION & IMPORTANCE:

Disambiguation is the process of clarifying and resolving ambiguities in data extraction. This is critical in IDP, as many documents contain terms, abbreviations, or formats that can have multiple meanings or interpretations. Proper disambiguation ensures that the extracted data is contextually accurate.

STRATEGIES FOR IMPROVEMENT:

Contextual Analysis:
Implementing algorithms that consider the broader context of the document can help the system distinguish between similar terms or data points. For instance, when processing land records, the system should be able to differentiate between "lot" as a piece of land and "lot" as in a batch of items.

User Feedback Mechanism:
Allowing users to provide feedback on ambiguous extractions can help the system learn and improve over time. For example, if an IDP system extracts "GPA" from a transcript, it should prompt the user to confirm whether it refers to the Grade Point Average or another metric, and then use that feedback to refine future extractions.

Example Application:

In the case of processing driver's licenses, where abbreviations like "DOB" (Date of Birth) and "DOE" (Date of Expiry) might be confused, a disambiguation system that understands the context of these fields can prevent errors in data extraction.

3. Honing AI Hints

DEFINITION & IMPORTANCE:

Honing AI Hints involves guiding the IDP system with subtle cues or instructions to improve its decision-making process. This is particularly useful when processing documents with complex structures or variable formats, such as invoices or legal contracts.

STRATEGIES FOR IMPROVEMENT:

Guided Templates: Using templates that provide the system with hints about where specific data is likely to be found in a document can significantly improve accuracy. For example, an invoice template might indicate that the total amount is usually located near the bottom right of the page, guiding the AI to focus its extraction efforts there.

Dynamic Learning: The system should adapt to the hints provided by users over time, refining its understanding of document layouts and improving its extraction accuracy with each iteration.

Example Application:

For instance, in processing invoices, if the system initially misidentifies the “Total Amount” field, a user can provide a hint or correction. The system then learns from this hint and improves its accuracy in processing subsequent invoices.

4. Retrieval-Augmented Generation (RAG)

DEFINITION & IMPORTANCE:

RAG combines the generative capabilities of AI with the retrieval of specific, relevant information from external databases or document repositories. This approach ensures that the IDP system's outputs are not only accurate but also enriched with up-to-date and contextually relevant data.

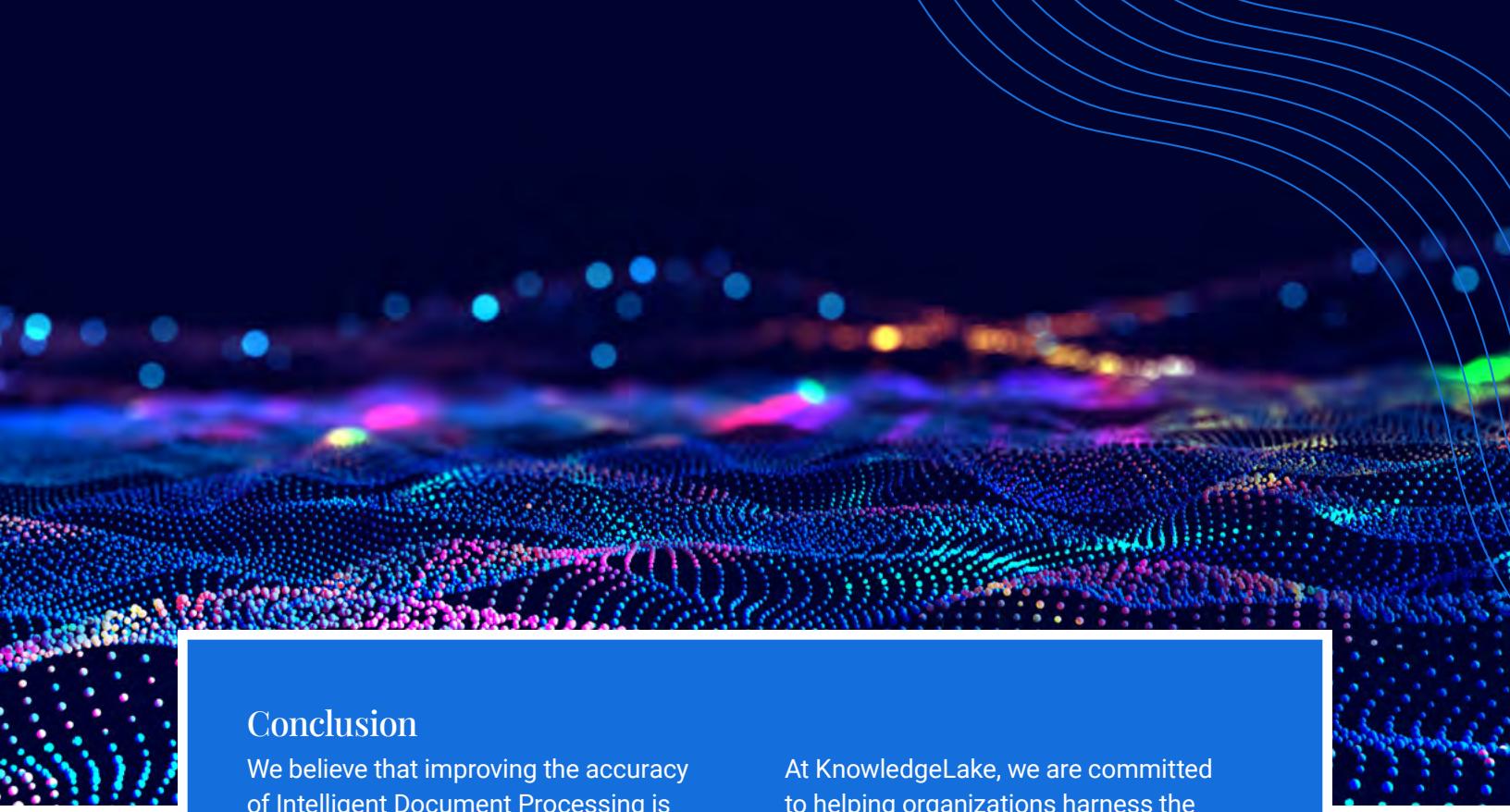
STRATEGIES FOR IMPROVEMENT:

Integration with External Data Sources:
Connecting the IDP system to external databases allows it to retrieve additional information that can confirm or complement the extracted data. For instance, when processing land records, the system can pull up related property databases to verify details like ownership history or zoning information.

Real-Time Data Retrieval:
Ensuring that the system can access and incorporate real-time data during the document processing phase enhances the accuracy and relevance of the extracted information.

Example Application:

A real estate firm processing land records can leverage RAG to not only extract details from the documents but also cross-reference them with a property database to ensure accuracy and completeness, reducing the risk of errors in property transactions.



Conclusion

We believe that improving the accuracy of Intelligent Document Processing is critical for organizations that handle a high volume of unstructured data. While we're not at 100% accuracy yet, but Intuitive is getting us closer every day. By focusing on Day Zero Accuracy, Disambiguation, Honing AI Hints, and Retrieval-Augmented Generation (RAG), organizations can significantly enhance the reliability and efficiency of their IDP systems. Whether dealing with college transcripts, land records, invoices, or driver's licenses, these strategies provide a robust framework for achieving superior document processing outcomes.

At KnowledgeLake, we are committed to helping organizations harness the full potential of Intelligent Document Processing. Our solutions are designed to deliver unparalleled accuracy and efficiency, ensuring that your document processing workflows are both seamless and effective.

For more information on how KnowledgeLake can help you optimize your document processing systems, contact us today at for a free PoC by an expert AI & Automation Solutions Architect.

Please contact:
leonard.spicer@knowledgelake.com