DOCUGAMI

The Document Engineering Company

LangChain Webinar: Lessons from Deploying LLMs with LangSmith

Jean Paoli @jeanpaoli

Taqi Jaffri @tjaffri

Mike Palmer

Zubin Wadia @zubinwadia

Topics

1. Intros

2. Real-World Challenges using LLMs with Documents
    i. Real documents are more than flat Text
    ii. Documents are Knowledge Graphs
    iii. Building Complex Chains with the LangChain Expression Language
    iv. Debugging Complex Chain Failures in Production

3. Summary: End to end LLM Ops
    ... deploy / run / trace / correct / finetune / repeat

Who we are...

- Generative AI for Business Documents. Founded 2018.
- Leverage open source and our own LLMs trained on millions of business documents to create a full XML data representation of complex documents in their entirety, delivering immediate value to frontline business users.
- Document as data can be used to generate insights and reports, create new documents, or drive line-of-business applications. Free trial at www.docugami.com/trial
- Customers in many sectors, including insurance, real estate, health, professional services.

Jean Paoli
Co-founder, CEO
Pioneer in document engineering; co-creator of XML, .docx, .xlsx, .pptx; started four $1B+ businesses at Microsoft

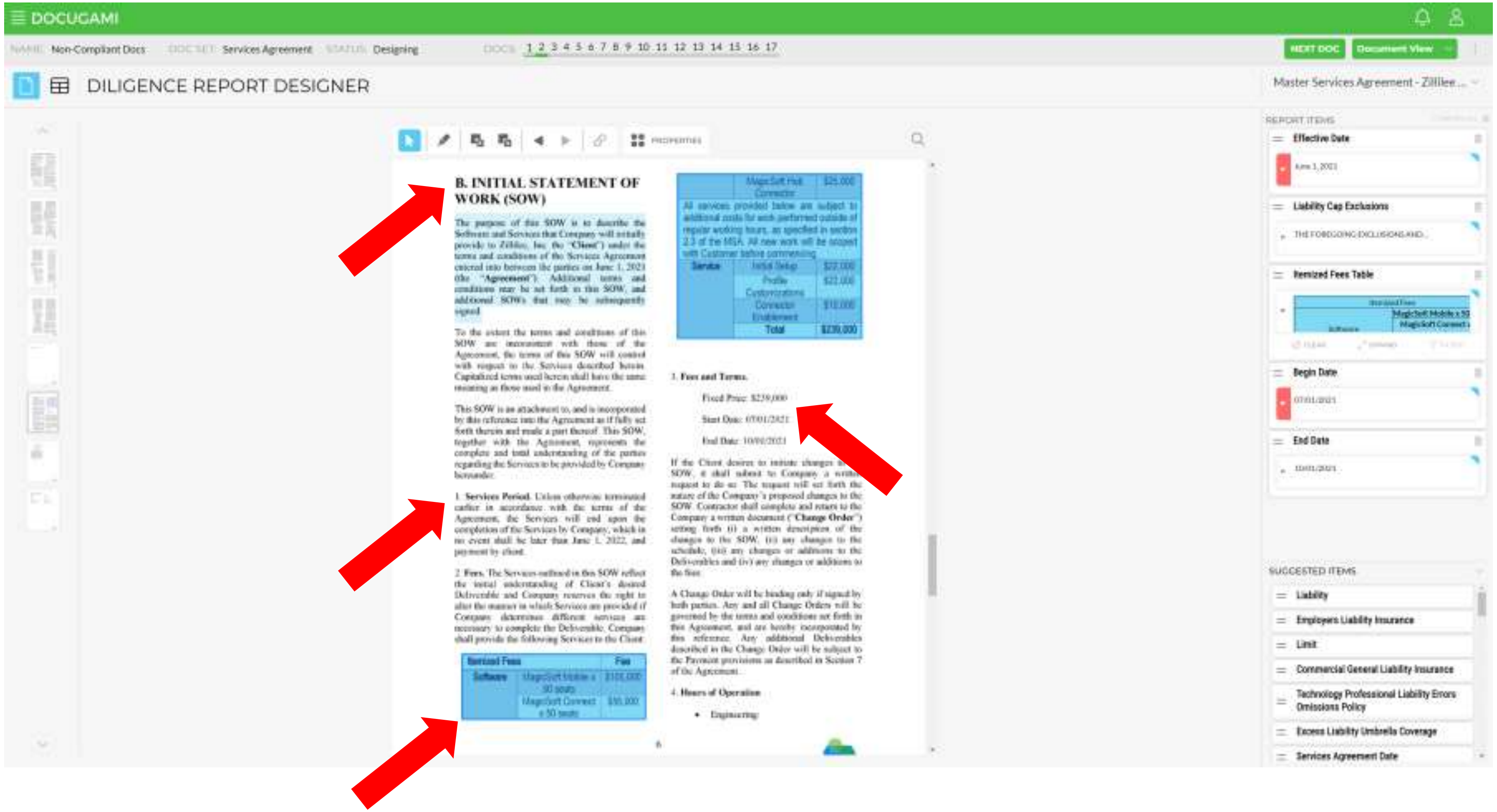
Taqi Jaffri
Co-founder, Head of Product
Principal Product Manager at Microsoft; co-created the AI driving Microsoft presence in physical stores.

Mike Palmer
Co-founder, Head of Technologies
Senior engineering manager at Microsoft; co-created Microsoft InfoPath.

Zubin Wadia
Product Manager
Making documents matter to people and systems – 100+ products over 20 years. MIT, ImageWork, CiviGuard.

Challenge 1: Real documents are more than flat text

Real Documents are Structurally Complex: Headings, Lists, Tables, Headers/Footers, Complex Reading Orders, Figures, etc.

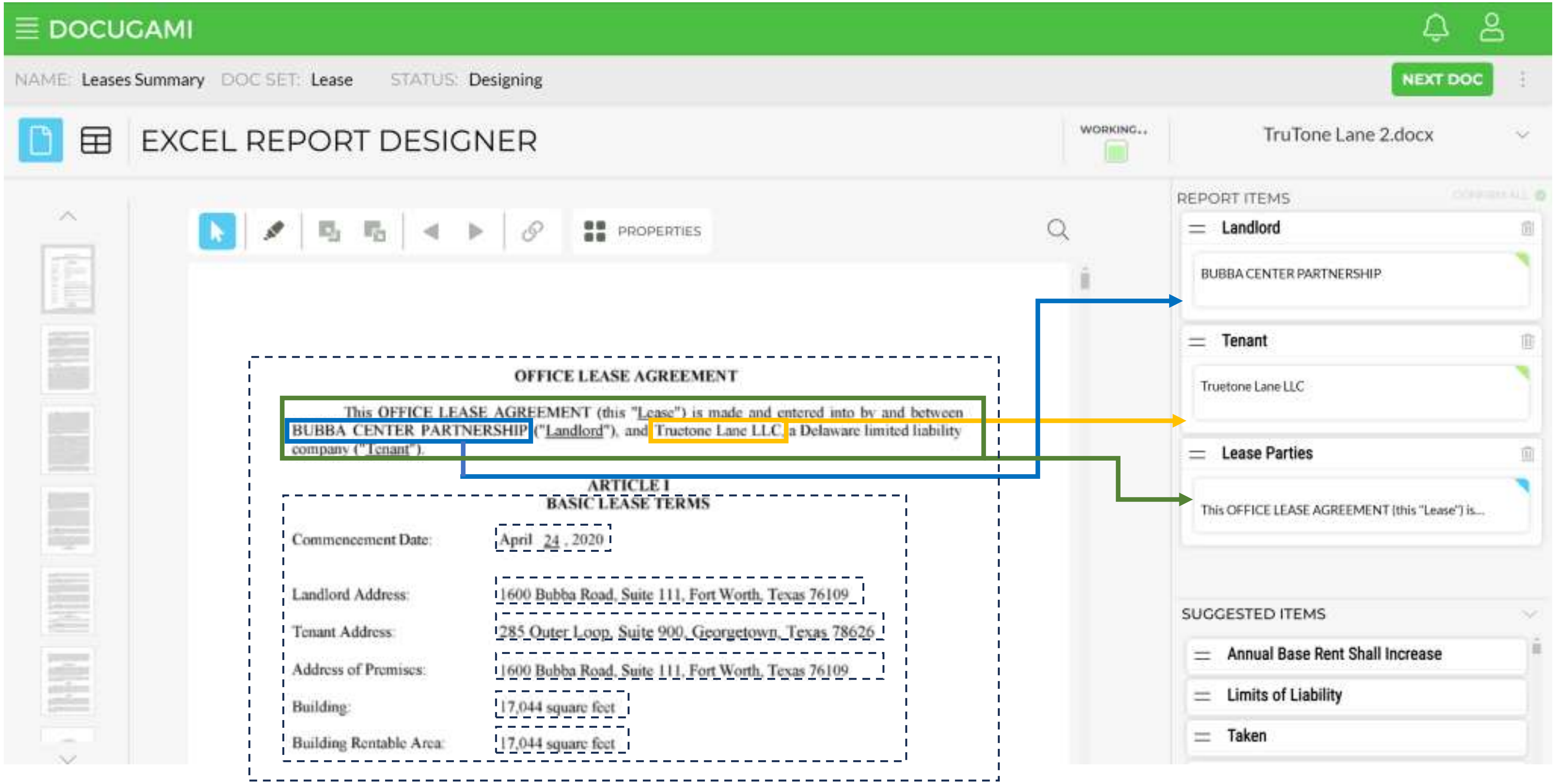


Mitigations:

1. Structurally chunk the document to find structural elements.
2. Stitch together reading order flow with language models.
3. Sub-chunk for RAG using document structural elements rather than simple tiling or text splitting.

Challenge 2: Documents are Knowledge Graphs

Document structures and spatial relationships contain semantics that are often lost in simple text-based retrieval



Mitigations:

1. XML hierarchical Knowledge Graph representation for inherent per-document semantics added to vector store for RAG
2. Schema normalization across sets of similar documents

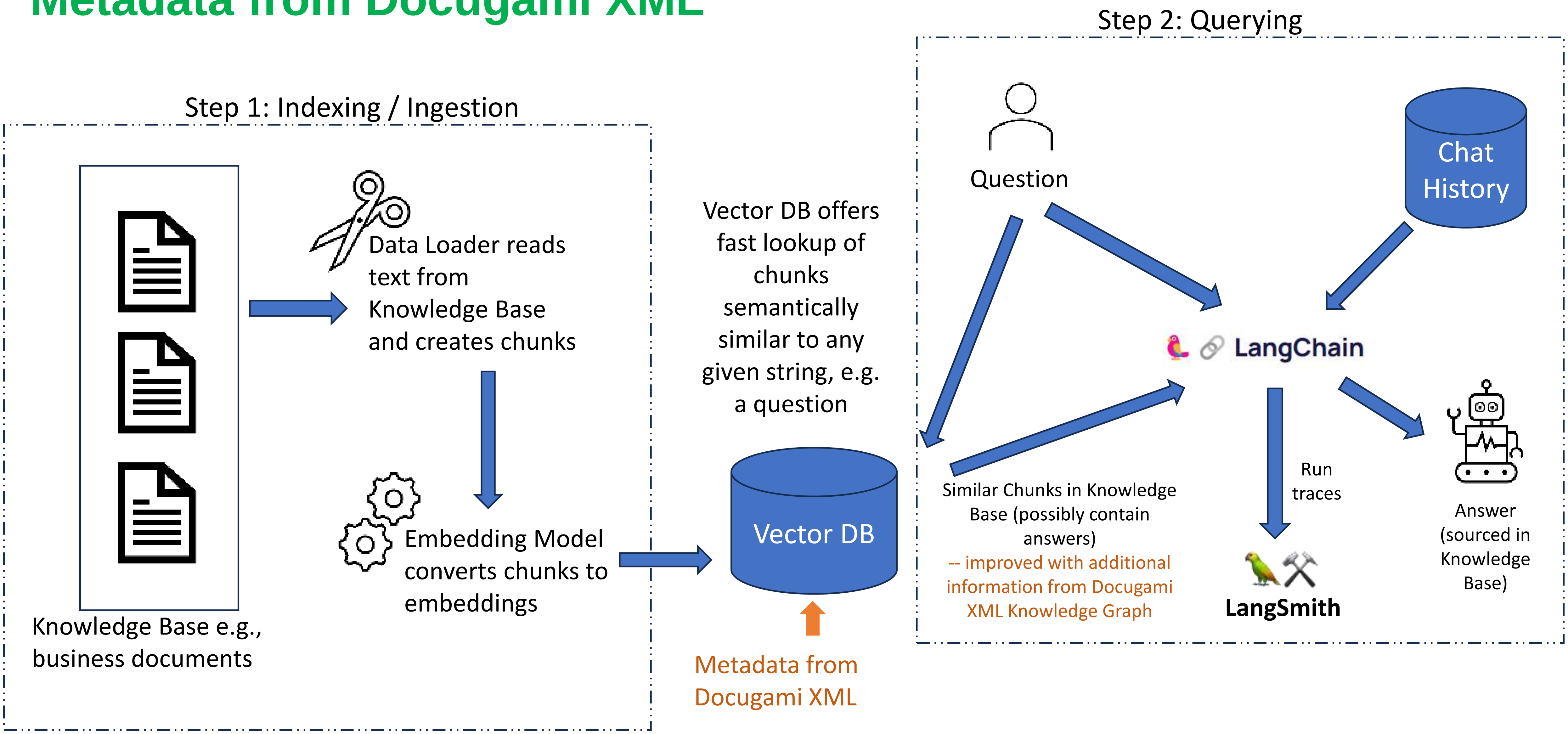
Example XML hierarchical Knowledge Graph

Represents structural relationships as well as per-chunk semantic labels


```xml
<?xml version="1.0" encoding="utf-8"?>
<docset:MUTUALNON-DISCLOSUREAGREEMENT-section xmlns:docset="http://www.docugami.com/2021/dgml/PublishTest/NDA" xmlns:addedChunks="http://www.docugami.com/2021/dgml/PublishTest/NDA/addedChunks" xmlns:dg="http://www.docugami.com/2021/dgml"
    <docset:MutualNon-disclosure structure="h1">
        <dg:chunk>MUTUAL NON-DISCLOSURE AGREEMENT </dg:chunk>
    </docset:MutualNon-disclosure>
    <docset:MUTUALNON-DISCLOSUREAGREEMENT structure="div">
        <docset:ThisMutualNon-disclosureAgreement>
            <docset:Preamble structure="p">
              This
               <dg:chunk>Mutual Non-Disclosure Agreement </dg:chunk>(this "
               <dg:chunk>Agreement</dg:chunk>") is entered into and made effective as of
               <docset:EffectiveDate xsi:type="date">2/4/2018 </docset:EffectiveDate>between
               <dgc:Org>Docugami Inc.</dgc:Org>, a
               <dgc:USState>Delaware </dgc:USState>corporation, whose address is 150 Lake Street South, Suite 221, Kirkland,
               <dgc:USState>Washington </dgc:USState>
               <dg:chunk>Delaware corporation</dg:chunk>, whose address is
               <docset:CompanyAddress>
                   <dgc:Street>
                       <dg:chunk>
                           <dgc:Number>150 </dgc:Number>
                           <dgc:StreetName>Lake Street South</dgc:StreetName>
                       </dg:chunk>,
                       <dgc:Apt>
                           <dg:chunk>Suite </dg:chunk>
                           <docset:Suite>221</docset:Suite>
                       </dgc:Apt>
                   </dgc:Street>,
                   <dgc:City>Kirkland</dgc:City>,
                   <dgc:State>Washington </dgc:State>
                   <dgc:Number>98033</dgc:Number>
               </docset:CompanyAddress>, and
               <docset:Signatory>Leonarda Hosler</docset:Signatory>, an individual, whose address is
               <dgc:Address>
                   <dgc:Street>
                       <dgc:Number>374 </dgc:Number>
                       <dgc:StreetName>William S Canning Blvd</dgc:StreetName>
                   </dgc:Street>,
                   <dg:chunk>
                       <dgc:City>Fall River </dgc:City>
                       <dgc:State>MA </dgc:State>
                   </dg:chunk>
               </dgc:Address>
               <dgc:Number>2721</dgc:Number>
               <docset:SignatoryAddress>374 William S Canning Blvd, Fall River MA 2721</docset:SignatoryAddress>.
            </docset:Preamble>
            <docset:Discussions structure="p">

              The above named parties desire to engage in discussions regarding a potential agreement or other transaction between the parties (the "Purpose"). In connection with such discussions, it may be necessary for the parties to d
            <docset:Consideration>
```

Retrieval Augmented Generation (RAG) w/ Semantic Chunk Metadata from Docugami XML



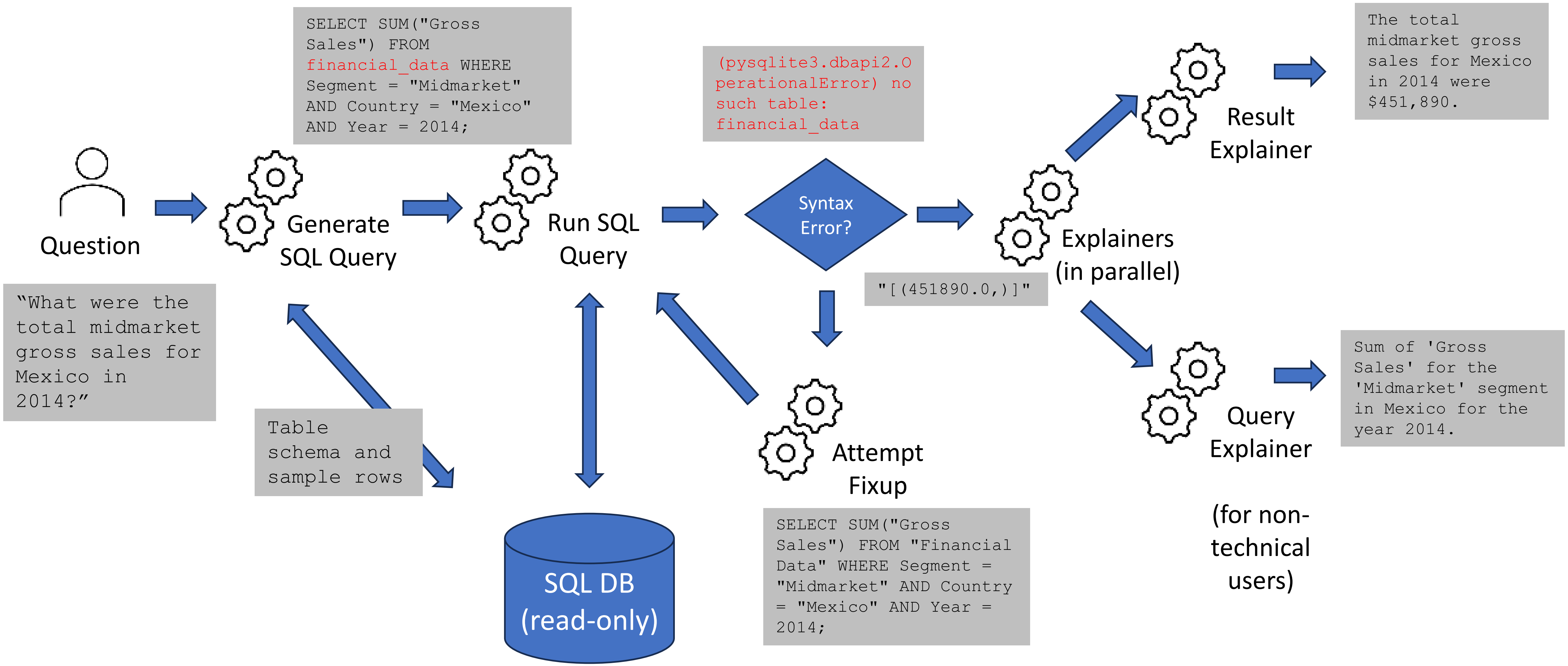
Challenges 1 & 2

CODE WALK-THROUGH

https://rebrand.ly/docugami_semantic_rag

Challenge 3: Building Complex Chains with the LangChain Expression Language

Real-world chains can get complicated with parallel branches, output parsers, few shot examples, conditional sub-chains, etc.



Sample code: https://rebrand.ly/docugami_complex_chain

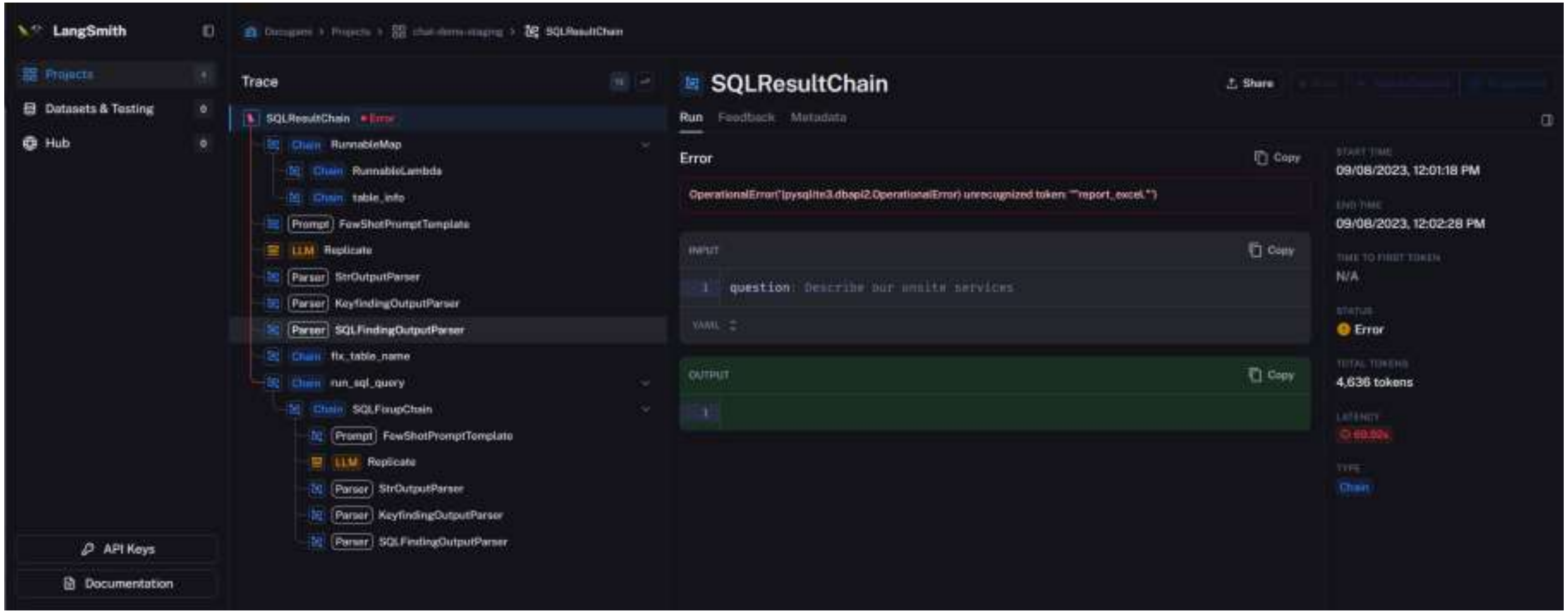
Challenge 3

LANGSMITH WALK-THROUGH

https://smith.langchain.com/public/ab8ef1ec-46d1-4d1f-980a-6dcc8f1943e0/r

Challenge 4: Debugging Complex Chain Failures in Production

Identifying what went wrong, where…



Common Failures:

1. Invalid SQL that could not be automatically fixed, leading to chain failure
2. Token overflow (table schema too large)
3. LLM rate limit reached
4. LLM call failed with GPU OOM (for self-hosted models)
5. Exception in custom python RunnableLambda
6. Exception in custom output parser

TIP: Name all runnable lambdas and pass config to conditionally-invoked runnables to correctly link and name. Reference: cookbook | example

TIP: If you have a link to a detailed call stack for an exception, you can add it to the failed run as metadata for later investigation. You can also add other information useful to investigate the failure later.

Summary: Docugami's End to end LLM Ops with LangChain + LangSmith

1. Deploy Model (we self-host a custom LLM)
2. Regularly look at failed and user-disliked runs
3. Add some problematic runs to dataset
   a) Failed runs e.g. with syntax errors
   b) Runs with negative user feedback
4. Fix runs in dataset
   a) Tip: use larger LLM to propose fixes
   b) Tip: validate that fixes are syntactically correct e.g. by checking SQL syntax
5. Add some examples to few shot set for in-context learning
6. Fine tune model on updated dataset
7. Redeploy (go back to #1)

Q&A

DOCUGAMI

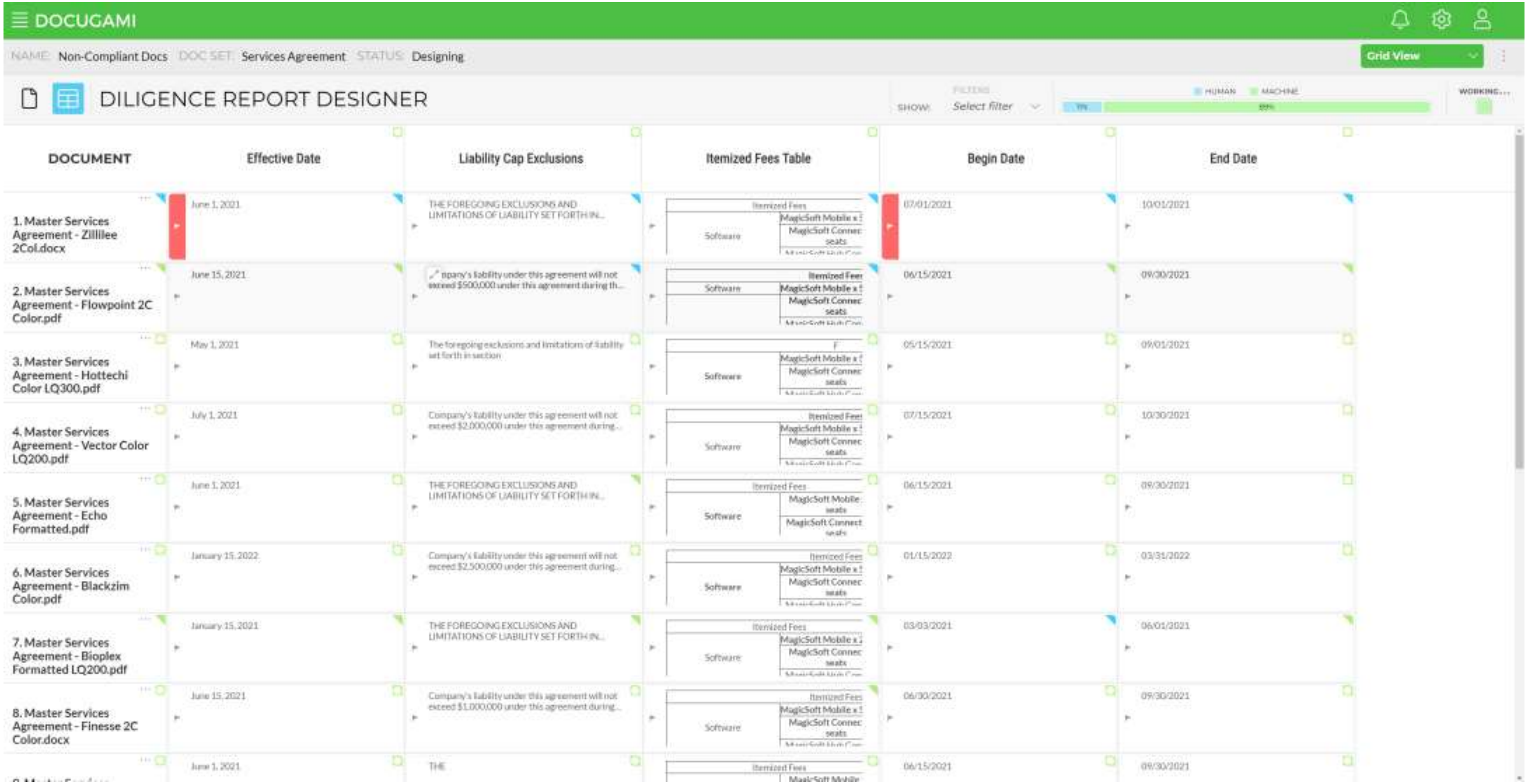
The Document Engineering Company

Docugami API

1. Docs: https://help.docugami.com/home/docugami-api
2. Things to try:
   a) Upload docs (DOCX, DOC, digital or scanned PDF)
   b) Download processed XML (with semantic metadata tags)
   c) Build reports to identify key chunks
   d) Use the LlamaHub Docugami loader to load docs for RAG, with report metadata
3. Free trials available!

Example: Extracting Custom Data from a Document Set

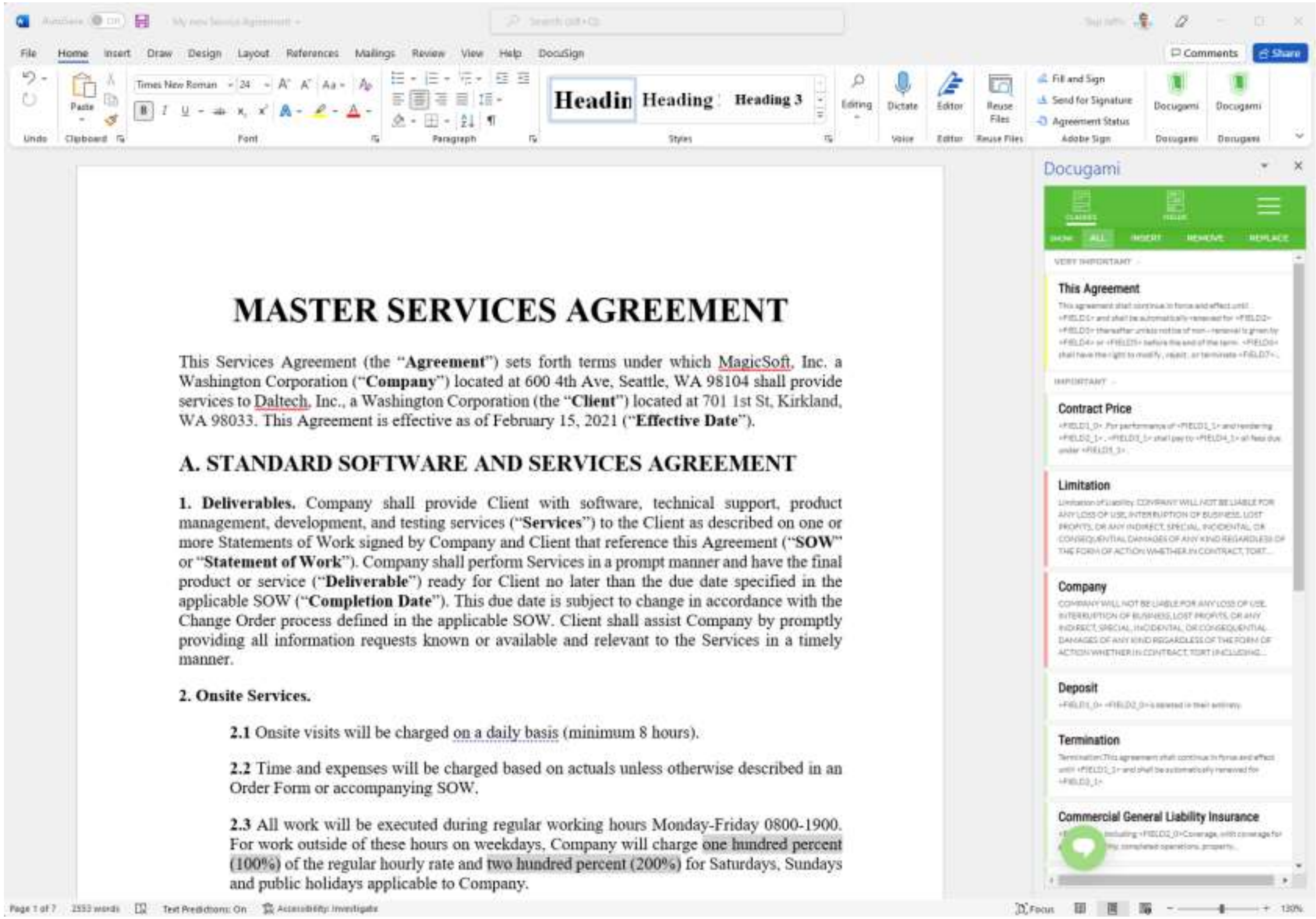
Improved significantly via structural chunking and key metadata associated with all chunks



DOCUMENT	Effective Date	Liability Cap Exclusions	Itemized Fees Table	Begin Date	End Date
1. Master Services Agreement - Zillilee 2Col.docx	June 1, 2021	THE FOREGOING EXCLUSIONS AND LIMITATIONS OF LIABILITY SET FORTH IN...	Itemized Fees; Software; MagicSoft Mobile x; MagicSoft Connect seats	07/01/2021	10/01/2021
2. Master Services Agreement - Flowpoint 2C Color.pdf	June 15, 2021	mpany's liability under this agreement will not exceed $500,000 under this agreement during th...	Itemized Fees; Software; MagicSoft Mobile x; MagicSoft Connect seats	06/15/2021	09/30/2021
3. Master Services Agreement - Hottechi Color LQ300.pdf	May 1, 2021	The foregoing exclusions and limitations of liability set forth in section	Software; MagicSoft Mobile x; MagicSoft Connect seats	05/15/2021	09/01/2021
4. Master Services Agreement - Vector Color LQ200.pdf	July 1, 2021	Company's liability under this agreement will not exceed $2,000,000 under this agreement during...	Itemized Fees; Software; MagicSoft Mobile x; MagicSoft Connect seats	07/15/2021	10/30/2021
5. Master Services Agreement - Echo Formatted.pdf	June 1, 2021	THE FOREGOING EXCLUSIONS AND LIMITATIONS OF LIABILITY SET FORTH IN...	Itemized Fees; Software; MagicSoft Mobile seats; MagicSoft Connect seats	06/15/2021	09/30/2021
6. Master Services Agreement - Blackzim Color.pdf	January 15, 2022	Company's liability under this agreement will not exceed $2,500,000 under this agreement during...	Itemized Fees; Software; MagicSoft Mobile x; MagicSoft Connect seats	01/15/2022	03/31/2022
7. Master Services Agreement - Bioplex Formatted LQ200.pdf	January 15, 2021	THE FOREGOING EXCLUSIONS AND LIMITATIONS OF LIABILITY SET FORTH IN...	Itemized Fees; Software; MagicSoft Mobile x; MagicSoft Connect seats	03/03/2021	06/01/2021
8. Master Services Agreement - Finesse 2C Color.docx	June 15, 2021	Company's liability under this agreement will not exceed $1,000,000 under this agreement during...	Itemized Fees; Software; MagicSoft Mobile x; MagicSoft Connect seats	06/30/2021	09/30/2021
9. Master Services	June 1, 2021	THE	Itemized Fees; MagicSoft Mobile	06/15/2021	09/30/2021

Example: Authoring Assistance Based on Your Documents

Improved significantly via structural chunking and key metadata associated with all chunks



Example: Workflows Triggered by Auto-Extracted Data

Improved significantly via structural chunking and key metadata associated with all chunks