Snorkel

Snorkel

Data development for GenAI: A systems-level view

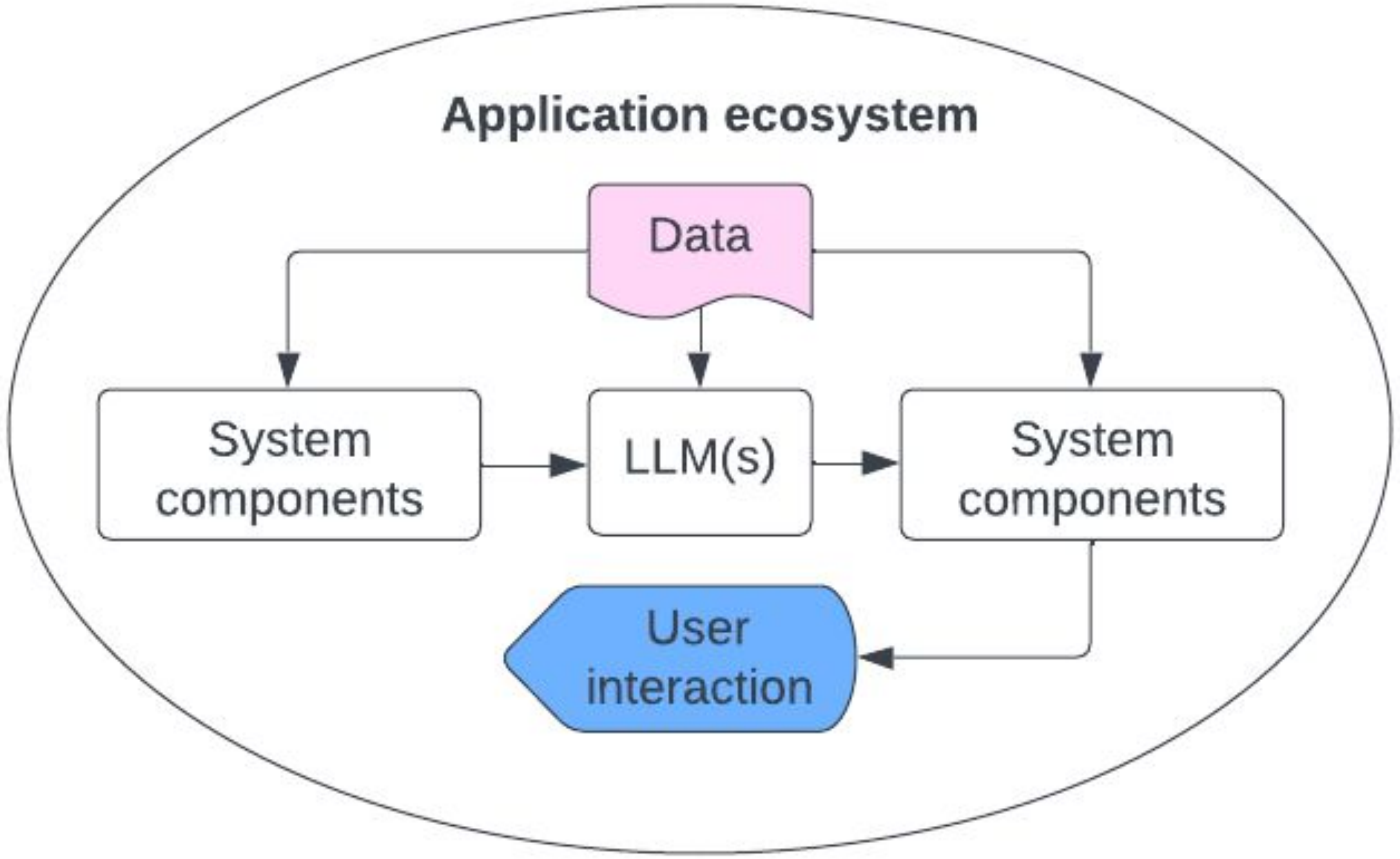
Chris Glaze
Staff Research Scientist

Summary

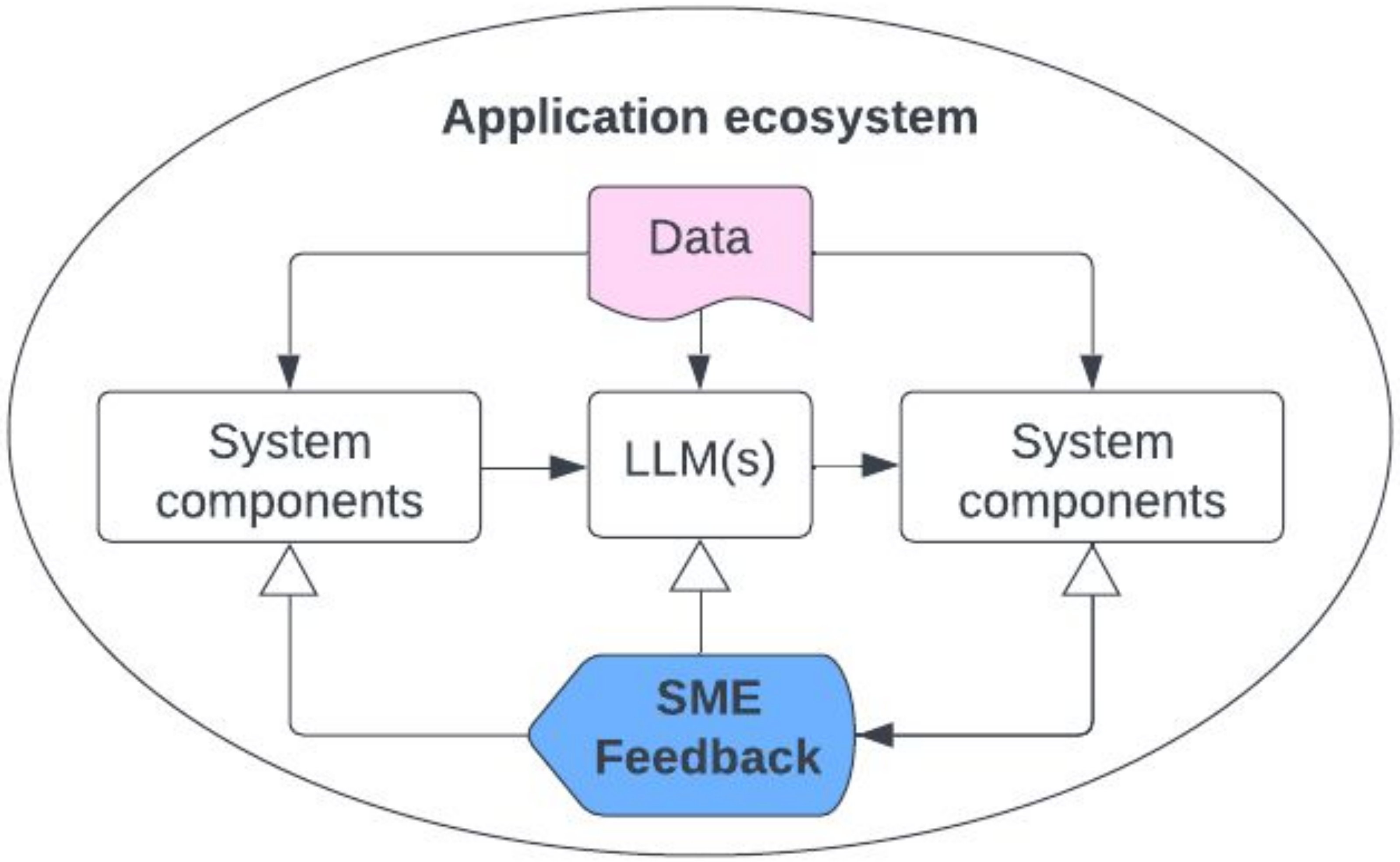
- For large language models (LLMs) in targeted business use cases:
  - Accuracy depends on the larger ecosystem in which they live
  - All system components benefit from fine-tuning with subject-matter expert (SME) feedback

- Snorkel is developing methods that efficiently incorporate SMEs in these development loops
  - Case Study: retrieval-augmented generation (RAG) for a top global bank, 54 point increase in question-answering accuracy in 3 weeks

Large language models do not exist in vacuums



Many components require fine-tuning



But subject-matter expert feedback has a scalability problem.

Keeping subject-matter experts in the loop

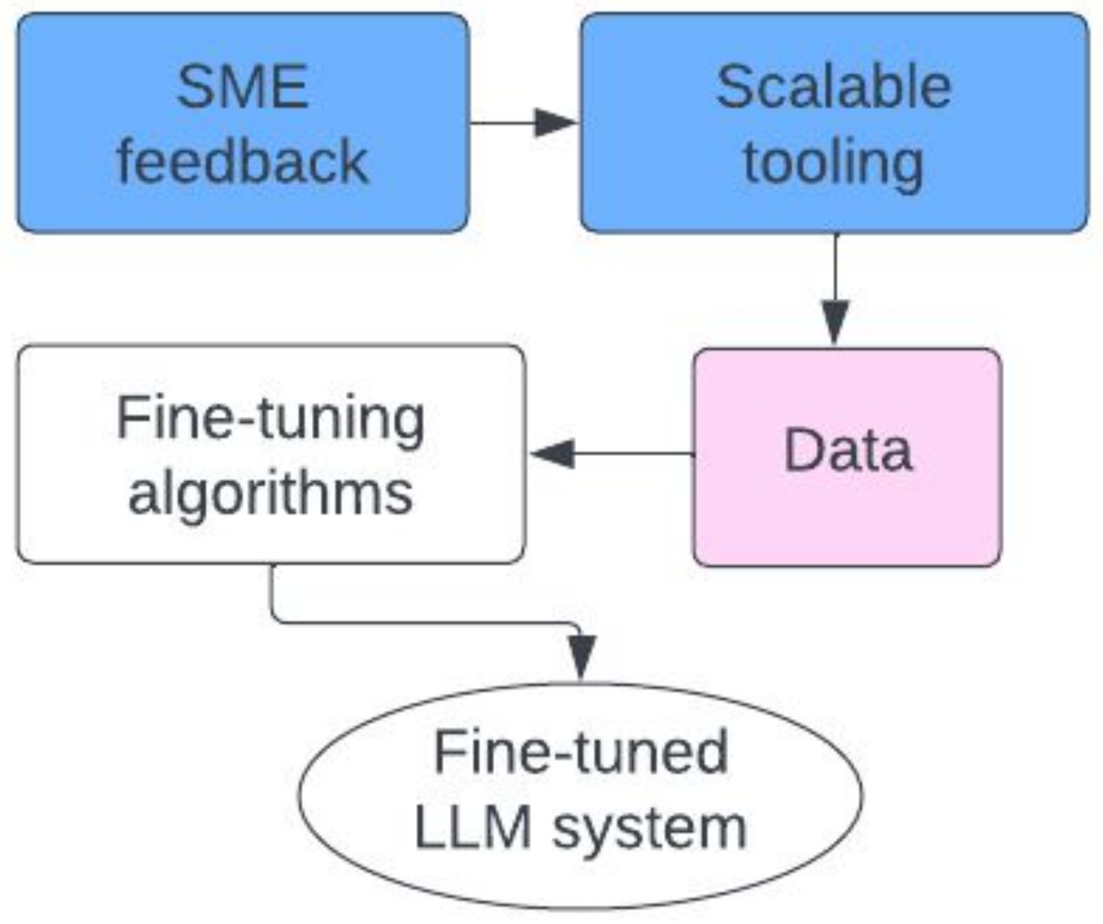
Snorkel thesis:

Data development is key.

Only subject-matter experts know what good looks like.

Snorkel approach:

- Keep subject-matter experts in the loop.
- Maximize value of their time with scalable methods to develop data.



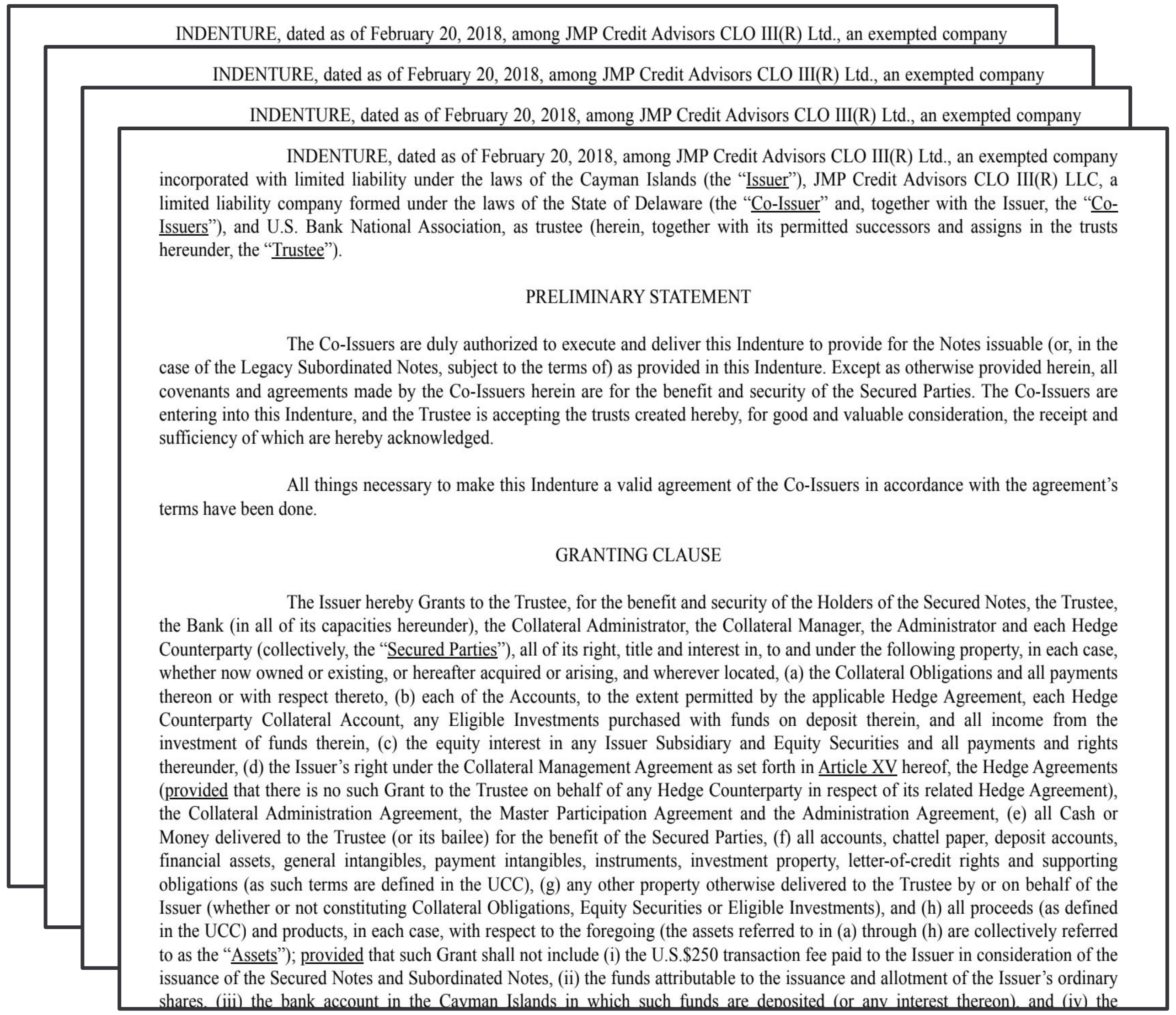
Case Study: Retrieval-augmented generation (RAG) for a top global bank

We fine-tuned portions of an LLM-based question-answering system using programmatic data development techniques, over a 3 week period:

	Baseline LLM (GPT-4) + vector retrieval	**Fine-tuned** LLM Q&A system	***Improvement***
Accuracy	**25%**	**79%**	***+54 pts.***

Case Study: Retrieval-augmented generation (RAG) for a top global bank

Unstructured Financial Documents


INDENTURE, dated as of February 20, 2018, among JMP Credit Advisors CLO III(R) Ltd., an exempted company

INDENTURE, dated as of February 20, 2018, among JMP Credit Advisors CLO III(R) Ltd., an exempted company

INDENTURE, dated as of February 20, 2018, among JMP Credit Advisors CLO III(R) Ltd., an exempted company

INDENTURE, dated as of February 20, 2018, among JMP Credit Advisors CLO III(R) Ltd., an exempted company incorporated with limited liability under the laws of the Cayman Islands (the "Issuer"), JMP Credit Advisors CLO III(R) LLC, a limited liability company formed under the laws of the State of Delaware (the "Co-Issuer" and, together with the Issuer, the "Co-Issuers"), and U.S. Bank National Association, as trustee (herein, together with its permitted successors and assigns in the trusts hereunder, the "Trustee").

PRELIMINARY STATEMENT

The Co-Issuers are duly authorized to execute and deliver this Indenture to provide for the Notes issuable (or, in the case of the Legacy Subordinated Notes, subject to the terms of) as provided in this Indenture. Except as otherwise provided herein, all covenants and agreements made by the Co-Issuers herein are for the benefit and security of the Secured Parties. The Co-Issuers are entering into this Indenture, and the Trustee is accepting the trusts created hereby, for good and valuable consideration, the receipt and sufficiency of which are hereby acknowledged.

All things necessary to make this Indenture a valid agreement of the Co-Issuers in accordance with the agreement's terms have been done.

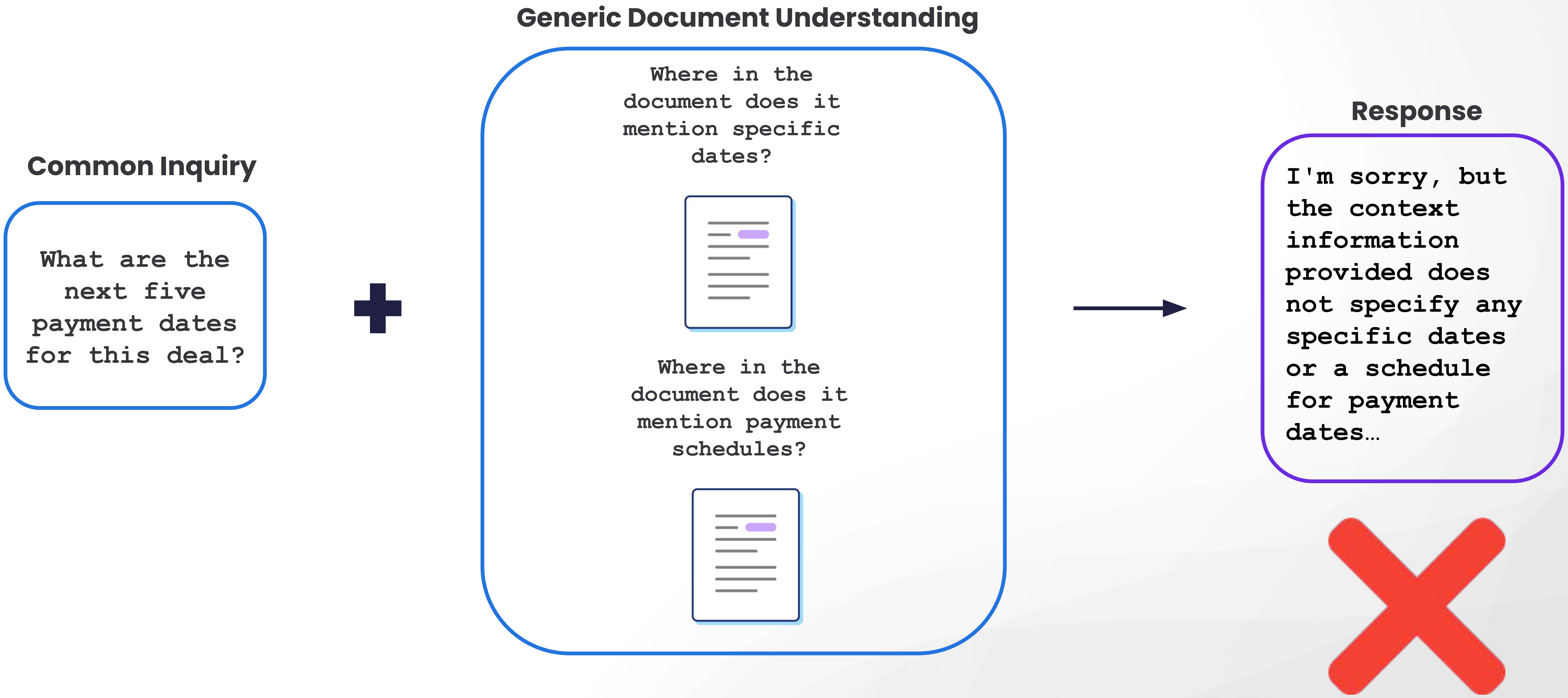
GRANTING CLAUSE

The Issuer hereby Grants to the Trustee, for the benefit and security of the Holders of the Secured Notes, the Trustee, the Bank (in all of its capacities hereunder), the Collateral Administrator, the Collateral Manager, the Administrator and each Hedge Counterparty (collectively, the "Secured Parties"), all of its right, title and interest in, to and under the following property, in each case, whether now owned or existing, or hereafter acquired or arising, and wherever located, (a) the Collateral Obligations and all payments thereon or with respect thereto, (b) each of the Accounts, to the extent permitted by the applicable Hedge Agreement, each Hedge Counterparty Collateral Account, any Eligible Investments purchased with funds on deposit therein, and all income from the investment of funds therein, (c) the equity interest in any Issuer Subsidiary and Equity Securities and all payments and rights thereunder, (d) the Issuer's right under the Collateral Management Agreement as set forth in Article XV hereof, the Hedge Agreements (provided that there is no such Grant to the Trustee on behalf of any Hedge Counterparty in respect of its related Hedge Agreement), the Collateral Administration Agreement, the Master Participation Agreement and the Administration Agreement, (e) all Cash or Money delivered to the Trustee (or its bailee) for the benefit of the Secured Parties, (f) all accounts, chattel paper, deposit accounts, financial assets, general intangibles, payment intangibles, instruments, investment property, letter-of-credit rights and supporting obligations (as such terms are defined in the UCC), (g) any other property otherwise delivered to the Trustee by or on behalf of the Issuer (whether or not constituting Collateral Obligations, Equity Securities or Eligible Investments), and (h) all proceeds (as defined in the UCC) and products, in each case, with respect to the foregoing (the assets referred to in (a) through (h) are collectively referred to as the "Assets"); provided that such Grant shall not include (i) the U.S.$250 transaction fee paid to the Issuer in consideration of the issuance of the Secured Notes and Subordinated Notes, (ii) the funds attributable to the issuance and allotment of the Issuer's ordinary shares, (iii) the bank account in the Cayman Islands in which such funds are deposited (or any interest thereon), and (iv) the

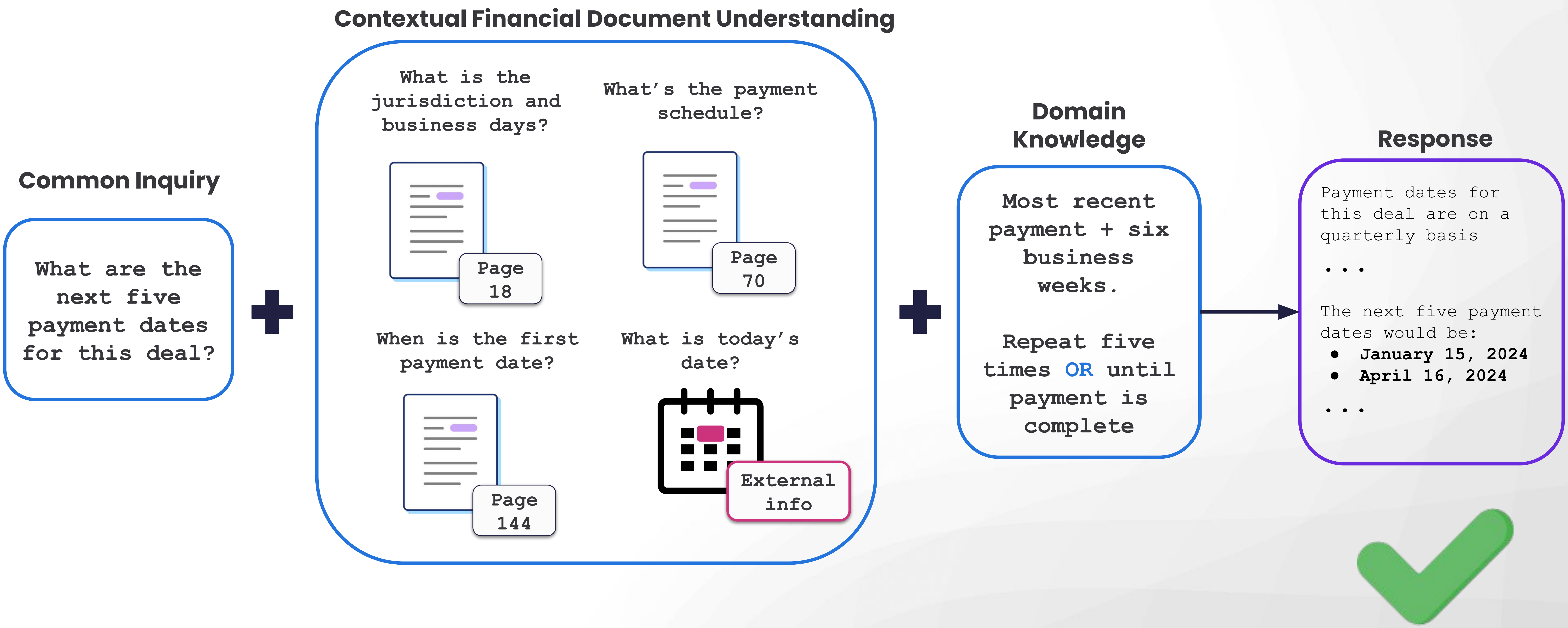
Document characteristics

- **Domain Specific**: Content is specific toniche financial products
- **Dense Documents**: Each document ranges from 200-500+ pages long
- **Subject Matter Expertise**: Bank experts are required to comprehend content and produce accurate results
- **Complex business logic**: to extract relevant information from various parts of the document

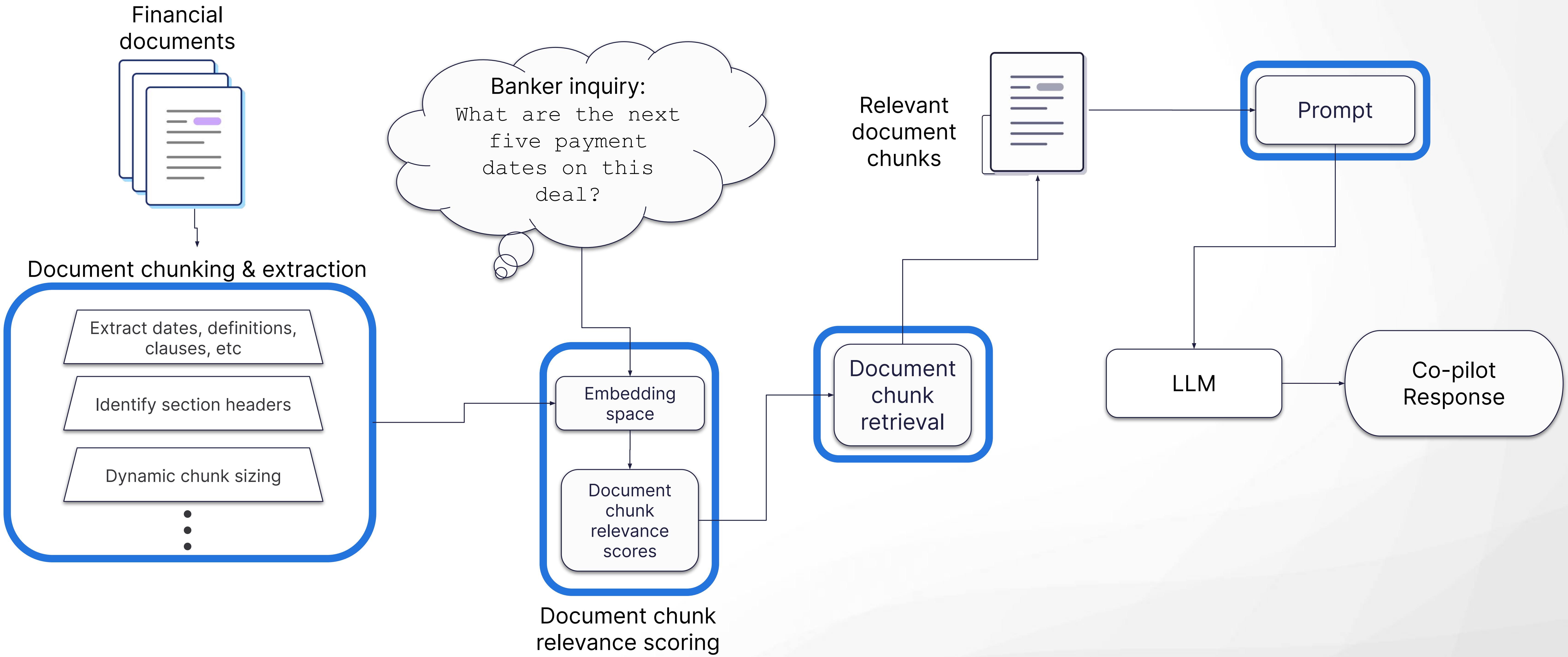
Example: Out-of-the-box RAG Performance



Example: Fine-tuned Performance

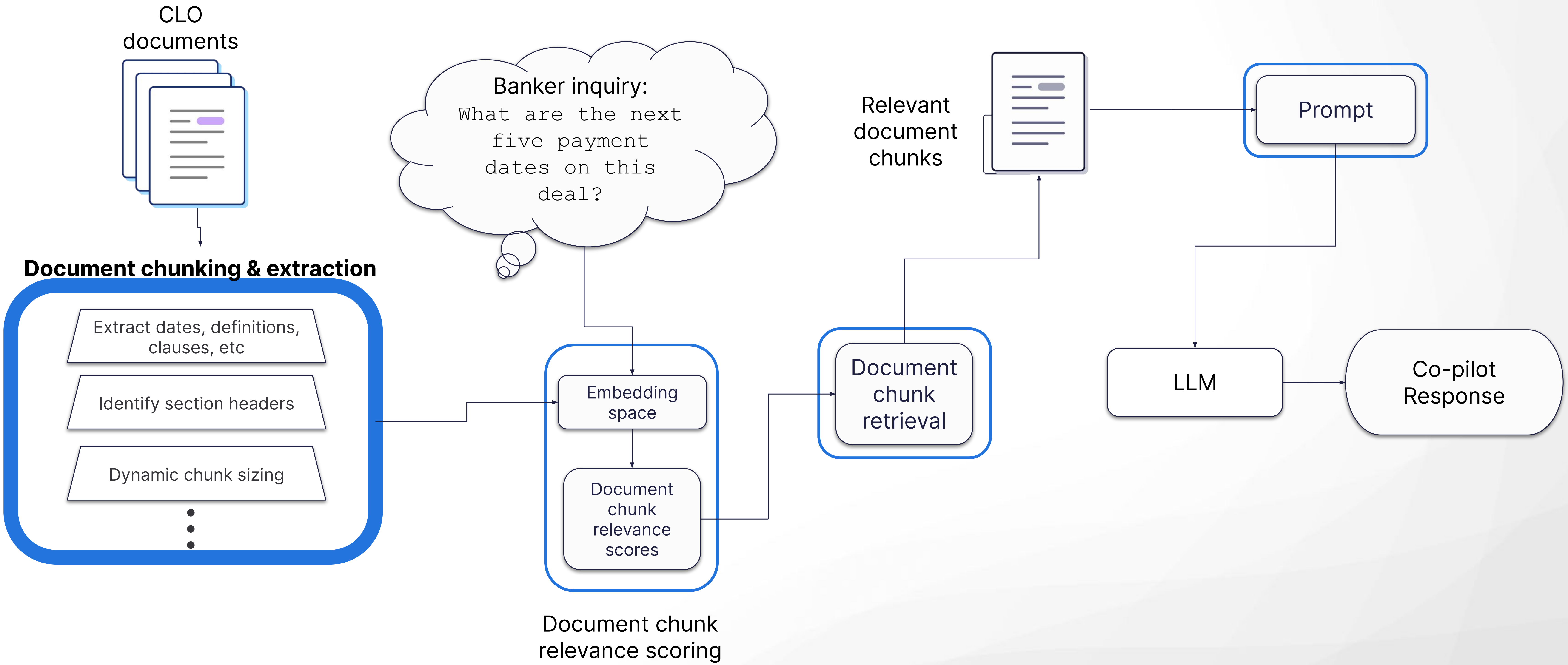


How Snorkel Did It

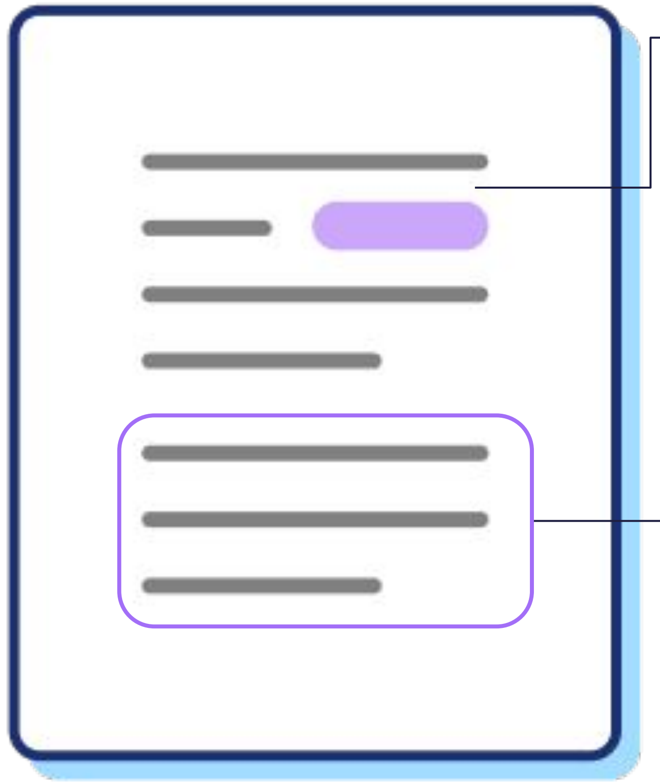


Fine tuned all components with subject-matter expert input → programmatic annotations

Document chunking, tagging, & extraction



Document tagging & extraction



Challenge: Q&A system struggled with key entities like dates, entities, definitions

Solution: Use Snorkel Flow to rapidly label structured entities

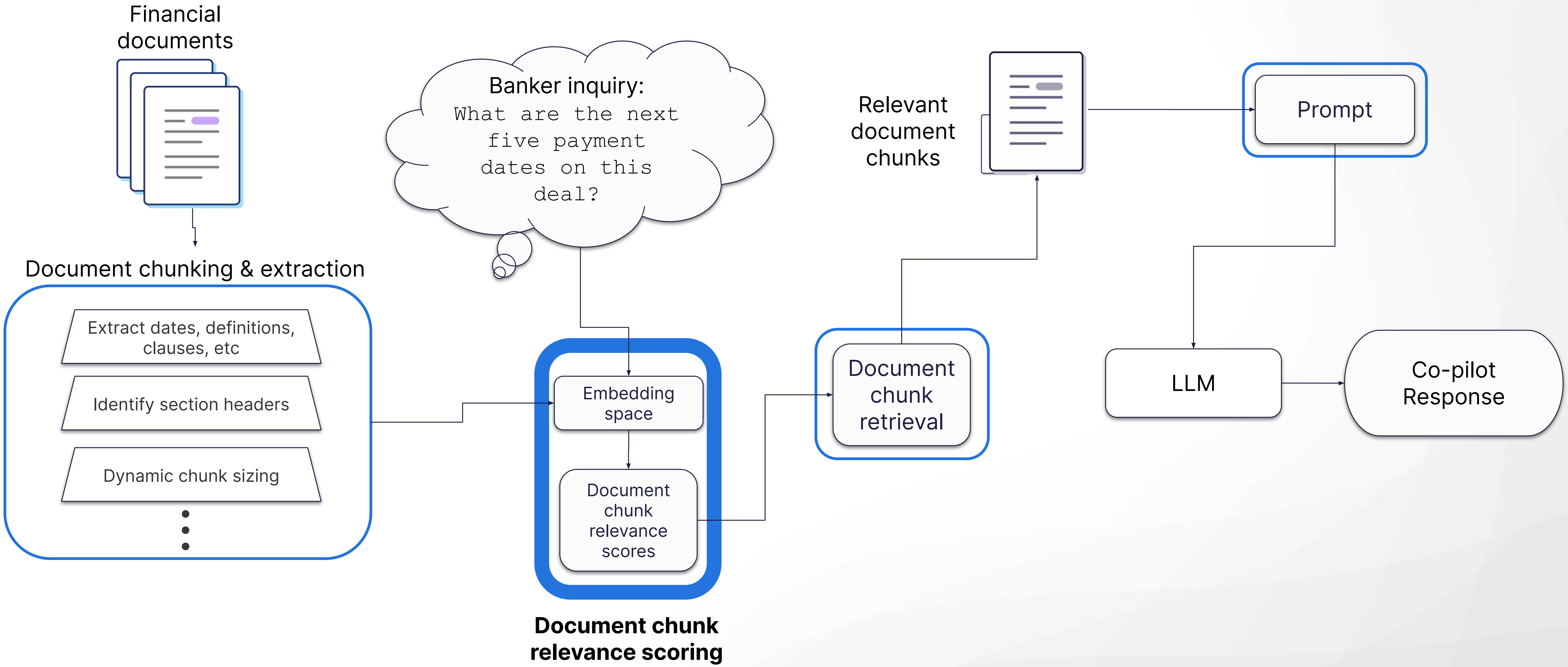
Challenge: Q&A system struggled to parse key sections, lumping them together

Solution: Use Snorkel Flow to parse section headers, tables, etc.

	Time to build training data & ML Model	SF Generated Labels	Model Accuracy
Date Model	4 hours	141,000 spans	99 F1
Definitions Model	4 hours	43,100 spans	93 F1

Subject-matter expert manually annotated relevant sections and gave us logic for those annotations → retrieval of key information

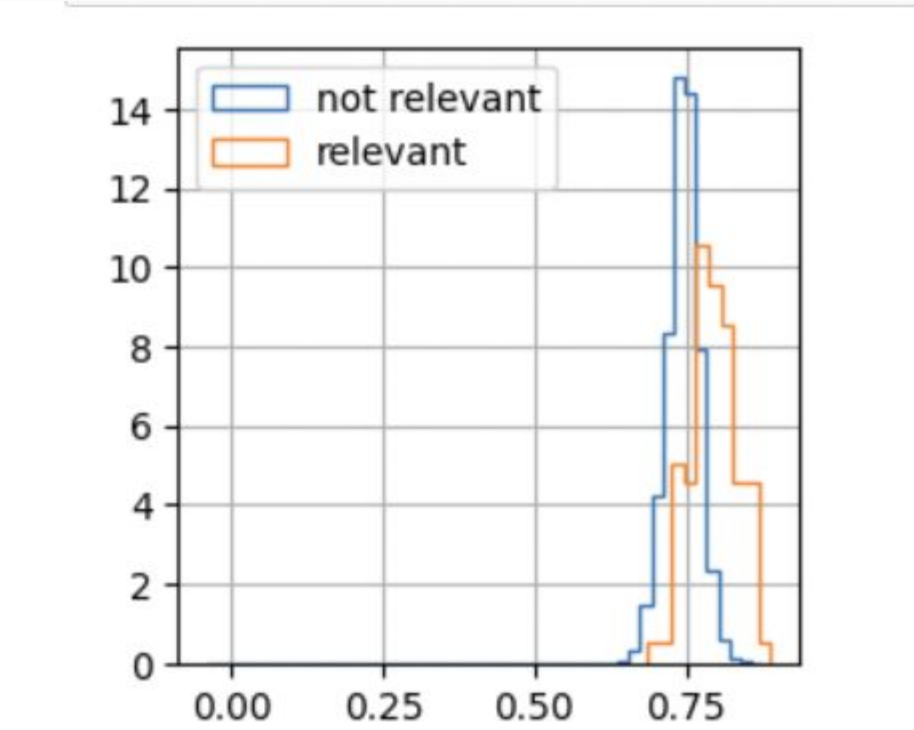
Fine-tuning the relevance scoring model



Fine-tuning the relevance scoring model

Baseline pipeline

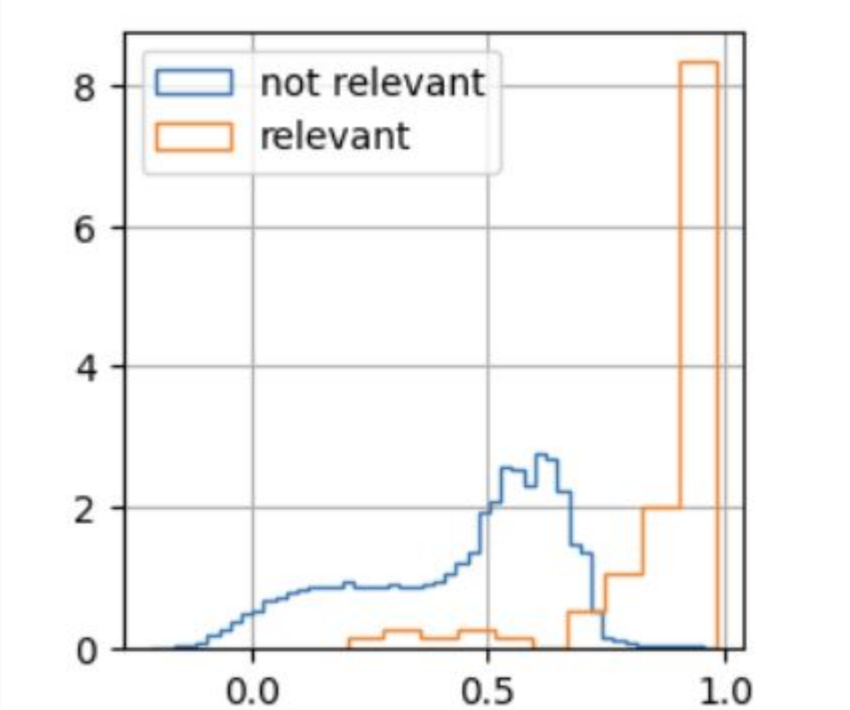
- Pages and chunks ranked with Ada embedding



Relevance scores

Fine tuned pipeline

- Pages and chunks ranked with BGE embedding
- Embedding fine tuned with programmatic data

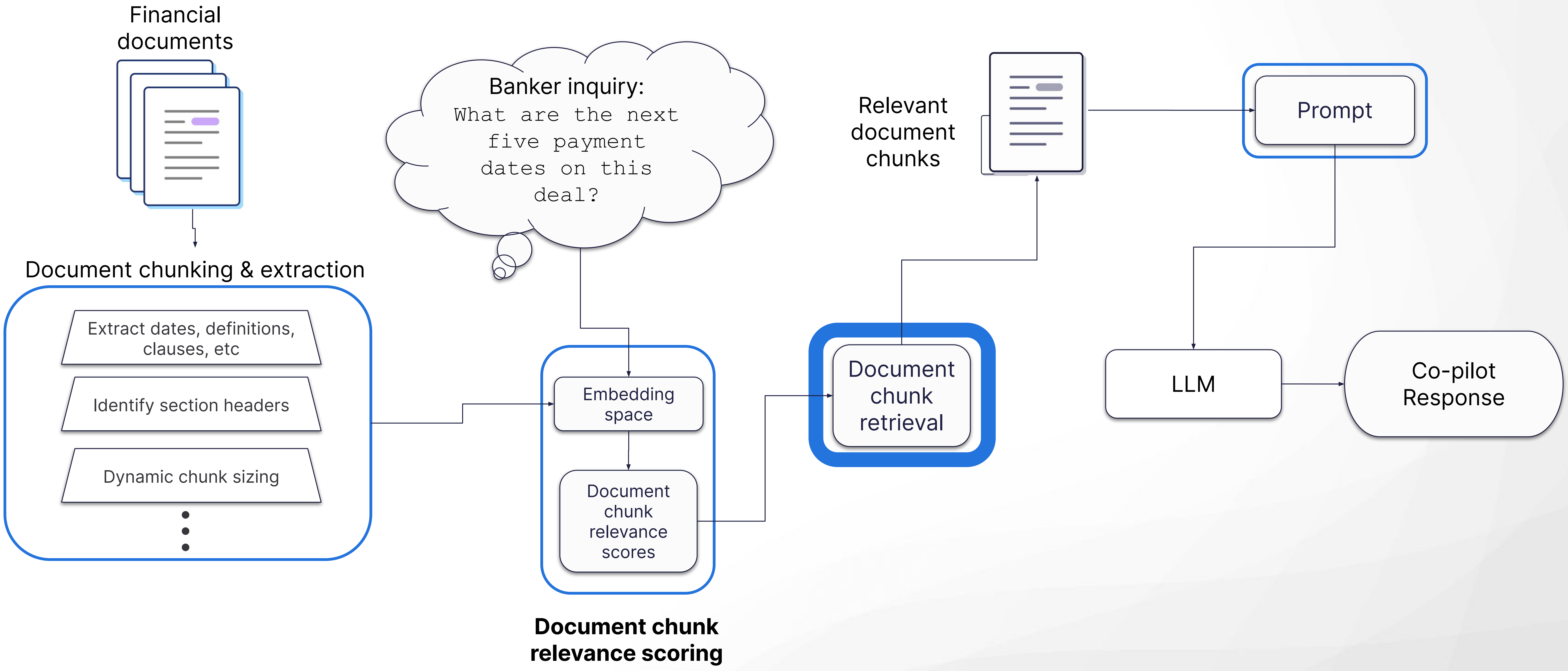


Relevance scores

Challenge: Default embedding lumped together relevant and irrelevant chunks

Solution: Fine-tune embeddings to distinguished relevant and irrelevant chunks
Subject-matter expert annotations and logic used for training set

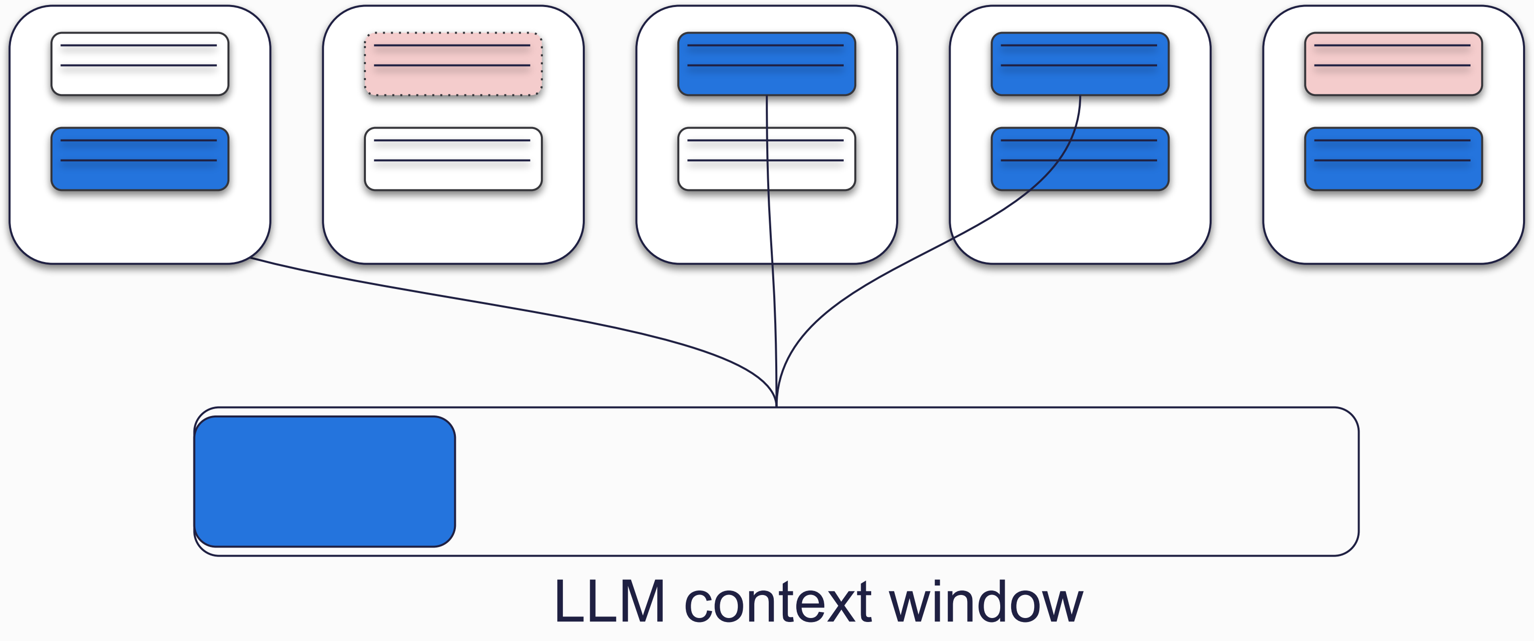
Fine-tuning the chunk retrieval algorithm



Fine-tuning the chunk retrieval algorithm

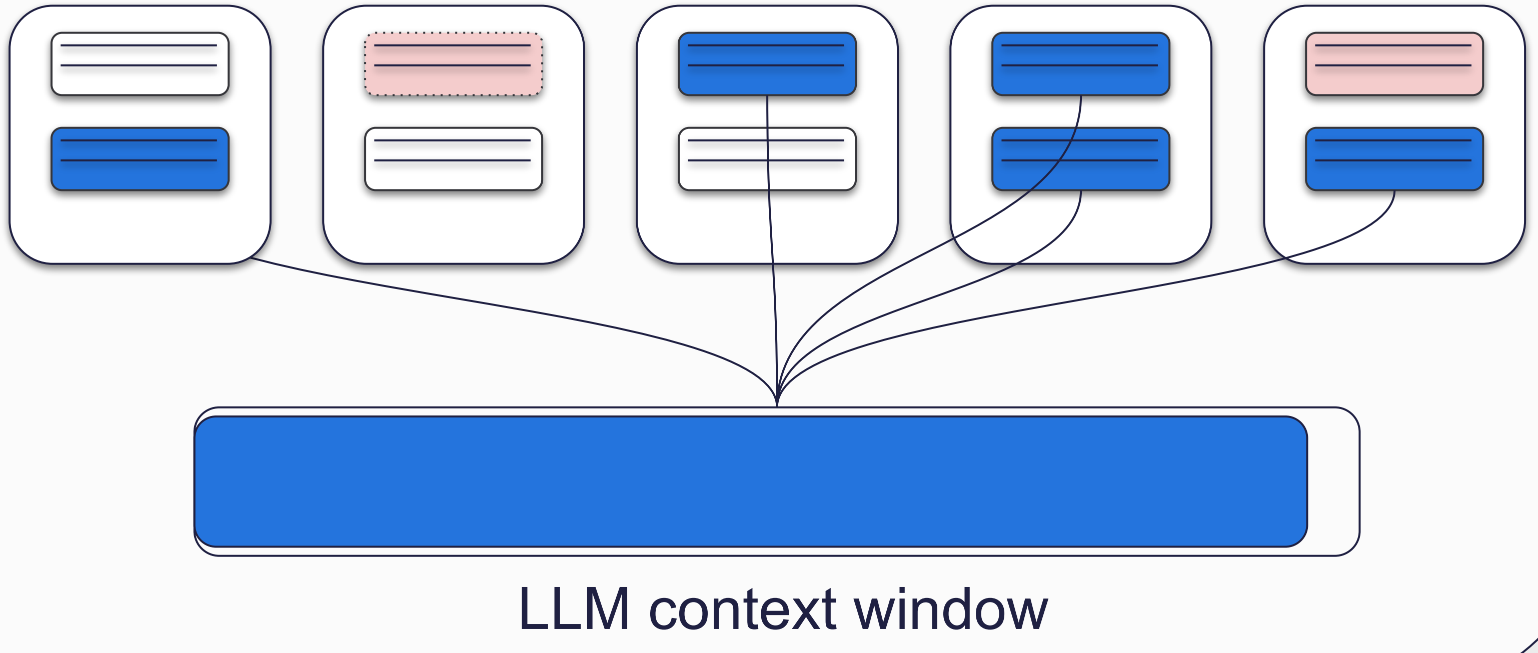
Baseline retrieval

- Fixed number (3) for every question based on ranking by relevance scores



Fine-tuned retrieval

- Number of chunks adaptive, based on relevance scores and context window size



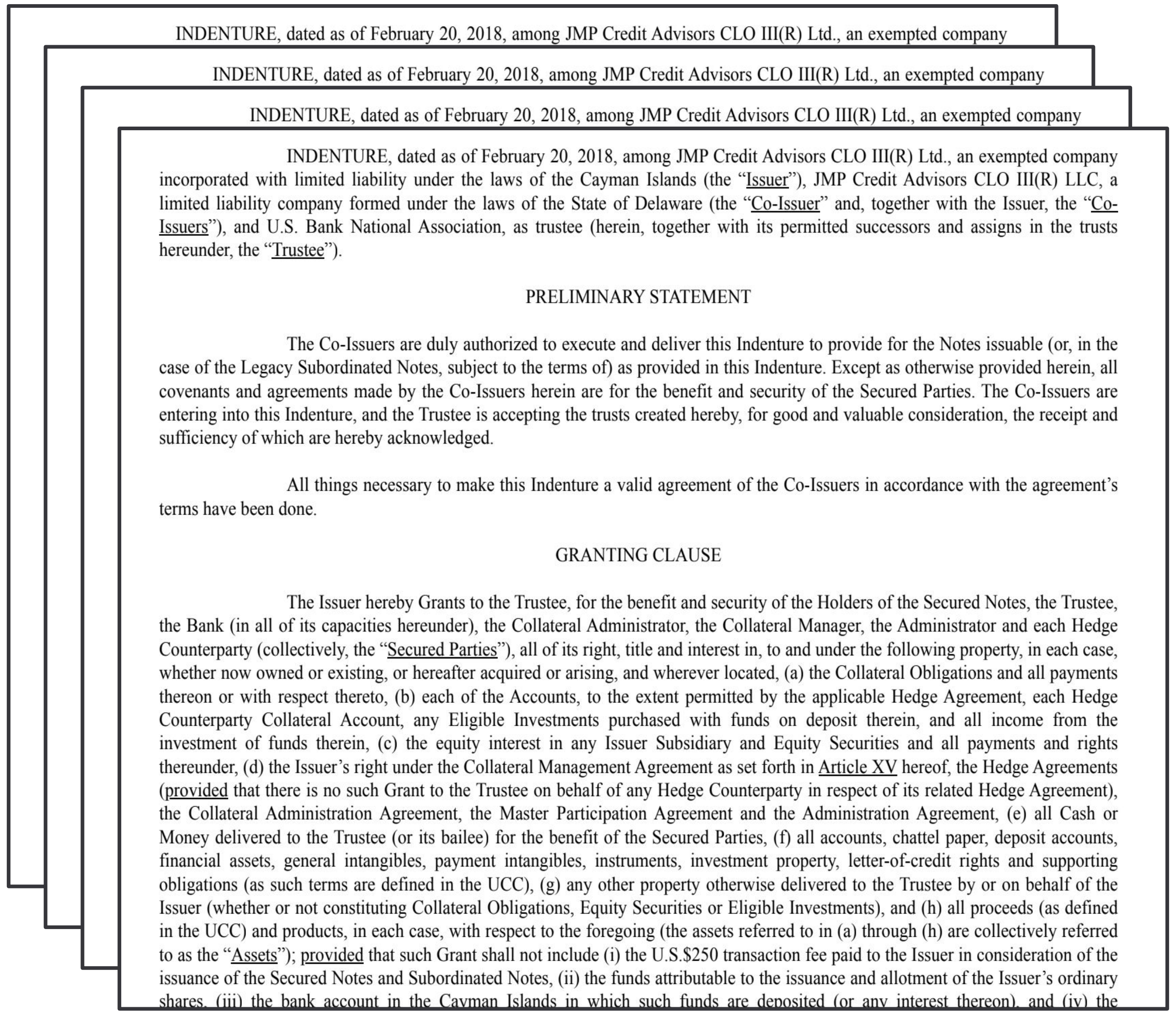
Low scoring chunks   High scoring chunks

Challenge: Multiple, disparate document sections can be relevant to a given question

Solution: Allow for varying number of retrieved chunks based on relevance scores
Subject-matter expert annotations and logic used for training set

Results: 54 point increase in answer accuracy

Unstructured Financial Documents


INDENTURE, dated as of February 20, 2018, among JMP Credit Advisors CLO III(R) Ltd., an exempted company

INDENTURE, dated as of February 20, 2018, among JMP Credit Advisors CLO III(R) Ltd., an exempted company

INDENTURE, dated as of February 20, 2018, among JMP Credit Advisors CLO III(R) Ltd., an exempted company

INDENTURE, dated as of February 20, 2018, among JMP Credit Advisors CLO III(R) Ltd., an exempted company incorporated with limited liability under the laws of the Cayman Islands (the "Issuer"), JMP Credit Advisors CLO III(R) LLC, a limited liability company formed under the laws of the State of Delaware (the "Co-Issuer" and, together with the Issuer, the "Co-Issuers"), and U.S. Bank National Association, as trustee (herein, together with its permitted successors and assigns in the trusts hereunder, the "Trustee").

PRELIMINARY STATEMENT

The Co-Issuers are duly authorized to execute and deliver this Indenture to provide for the Notes issuable (or, in the case of the Legacy Subordinated Notes, subject to the terms of) as provided in this Indenture. Except as otherwise provided herein, all covenants and agreements made by the Co-Issuers herein are for the benefit and security of the Secured Parties. The Co-Issuers are entering into this Indenture, and the Trustee is accepting the trusts created hereby, for good and valuable consideration, the receipt and sufficiency of which are hereby acknowledged.

All things necessary to make this Indenture a valid agreement of the Co-Issuers in accordance with the agreement's terms have been done.

GRANTING CLAUSE

The Issuer hereby Grants to the Trustee, for the benefit and security of the Holders of the Secured Notes, the Trustee, the Bank (in all of its capacities hereunder), the Collateral Administrator, the Collateral Manager, the Administrator and each Hedge Counterparty (collectively, the "Secured Parties"), all of its right, title and interest in, to and under the following property, in each case, whether now owned or existing, or hereafter acquired or arising, and wherever located, (a) the Collateral Obligations and all payments thereon or with respect thereto, (b) each of the Accounts, to the extent permitted by the applicable Hedge Agreement, each Hedge Counterparty Collateral Account, any Eligible Investments purchased with funds on deposit therein, and all income from the investment of funds therein, (c) the equity interest in any Issuer Subsidiary and Equity Securities and all payments and rights thereunder, (d) the Issuer's right under the Collateral Management Agreement as set forth in Article XV hereof, the Hedge Agreements (provided that there is no such Grant to the Trustee on behalf of any Hedge Counterparty in respect of its related Hedge Agreement), the Collateral Administration Agreement, the Master Participation Agreement and the Administration Agreement, (e) all Cash or Money delivered to the Trustee (or its bailee) for the benefit of the Secured Parties, (f) all accounts, chattel paper, deposit accounts, financial assets, general intangibles, payment intangibles, instruments, investment property, letter-of-credit rights and supporting obligations (as such terms are defined in the UCC), (g) any other property otherwise delivered to the Trustee by or on behalf of the Issuer (whether or not constituting Collateral Obligations, Equity Securities or Eligible Investments), and (h) all proceeds (as defined in the UCC) and products, in each case, with respect to the foregoing (the assets referred to in (a) through (h) are collectively referred to as the "Assets"); provided that such Grant shall not include (i) the U.S.$250 transaction fee paid to the Issuer in consideration of the issuance of the Secured Notes and Subordinated Notes, (ii) the funds attributable to the issuance and allotment of the Issuer's ordinary shares, (iii) the bank account in the Cayman Islands in which such funds are deposited (or any interest thereon), and (iv) the

25% Accurate

With out-of-the-box LLM + RAG

Failed to answer hardest questions

in 3 weeks

79% Accurate

With fine-tuning + other data development

Subject-matter expert <10 hours

Ablation studies: ½ of increase came from fine tuning the embeddings, ½ from fine tuning other components

Summary

- ➔ For large language models (LLMs) in targeted business use cases:
  - ◆ Accuracy depends on the larger ecosystem in which it lives
  - ◆ All system components benefit from fine-tuning with SME feedback

- ➔ Snorkel is developing methods that efficiently incorporate SMEs in these development loops
  - ◆ Case Study: RAG for a top global bank, 54 point increase in QA accuracy in 3 weeks

Thank you!

Chris Glaze
christopher.glaze@snorkel.ai