

Unstrukturierte Daten für KI aufbereiten: ByteIT Startup Story

Agenda



Einführung RAG

Wie Sprachagenten Ihr Wissen aufnehmen



Daten für KI aufbereiten

Aktuelle Probleme und Lösungen



Startup Gründen

Erfahrungen, Chancen und
Fallstricke

Sprache: Deutsch / Slides: Englisch

Today: AI as a Tool to help us do work we do not want to do



RAG consists of 4 essential Steps



01.

Parsing

Unify all data without losing information



02.

Chunking

Connect related pieces of data



03.

Embedding

Map meanings into a space



04.

Reranking

Smartly search through your data

Bad Inputs lead to bad outputs

Your data defines your AI:

Ingesting it properly is hard

Many Sources of Error



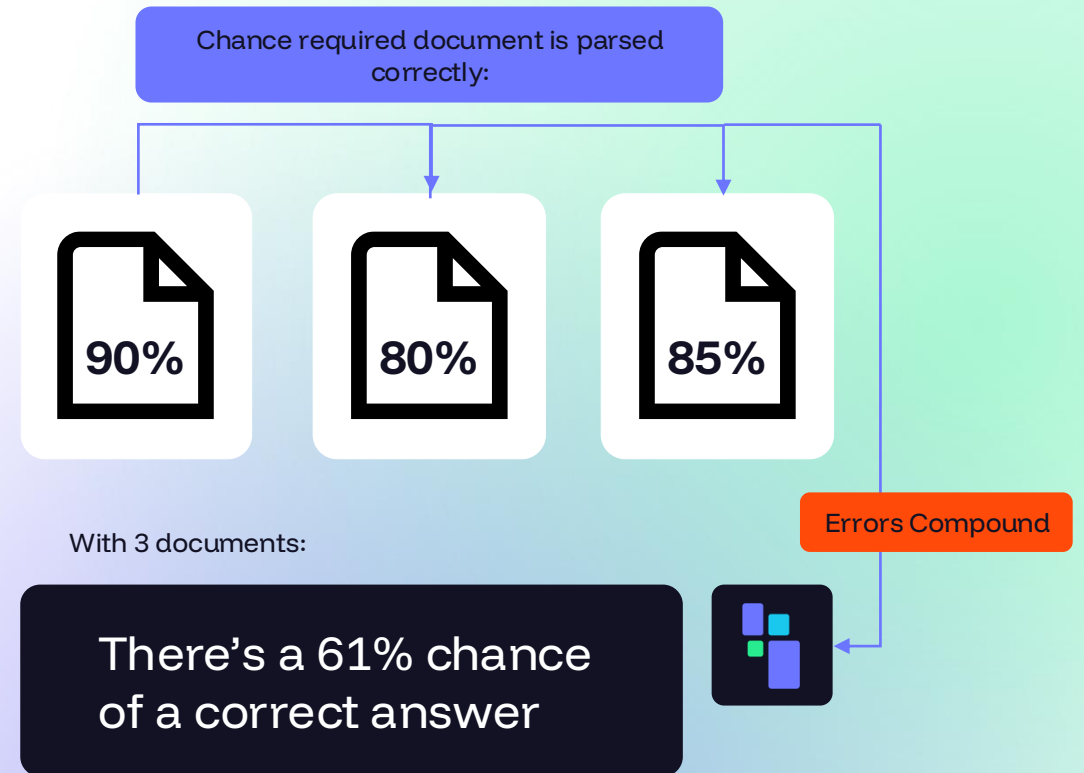
Text as Images



Cross-References



Tables and Forms



RAG consists of 4 essential Steps



01.

Parsing

Unify all data without losing information



02.

Chunking

Connect related pieces of data



03.

Embedding

Map meanings into a space

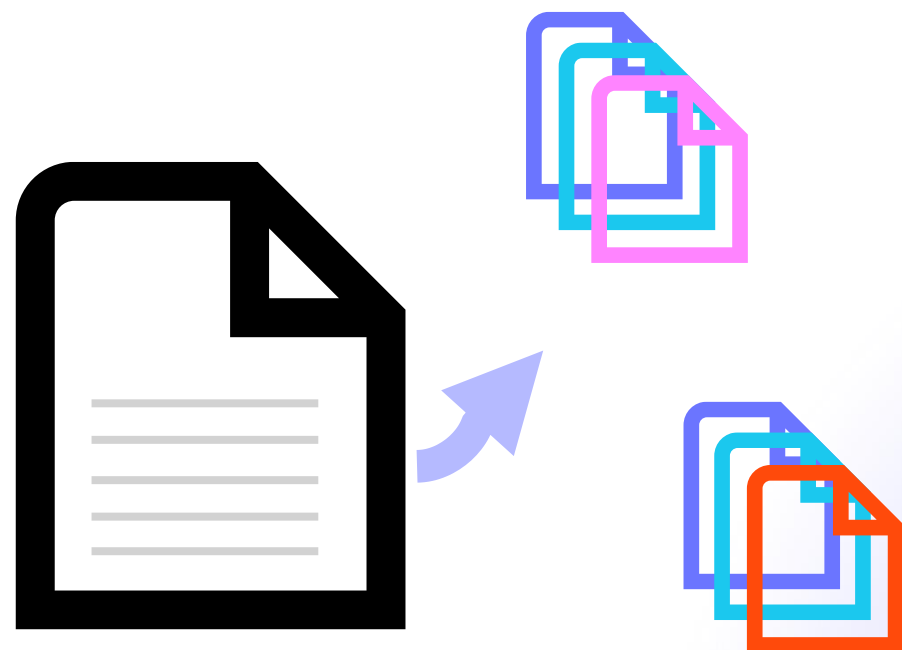


04.

Reranking

Smartly search through your data

Chunking



Shareholders:		
Freefloat	62.60 %	
Trier Asset Mgmt	22.40 %	
Alois Widmann	15.00 %	
p.1		
Liabilities and shareholders' equity		
Subscribed capital		
Capital reserve		
Retained earnings		
Other equity components		
Shareholders' equity		
Minority interest		
Total equity	63.5	61.8
Provisions	2.8	1.2
thereof provisions for pensions and similar obligations	2.6	1.1
Financial liabilities (total)	21.0	27.2
Short-term financial liabilities	3.5	5.5
Accounts payable	7.2	9.3
Other liabilities	8.8	11.0
Liabilities	39.8	48.6
Total liabilities and shareholders' equity	103.4	110.4
p.6		
Net Debt	14.0	21.0
Net Financial Debt	11.4	20.4
Net Gearing	22.0 %	34.8 %
Net Fin. Debt / EBITDA	125.4 %	209.4 %
Book Value / Share	6.9	6.7
Book value per share ex intangibles	0.8	0.6
p.3		

RAG consists of 4 essential Steps



01.

Parsing

Unify all data without losing information



02.

Chunking

Connect related pieces of data



03.

Embedding

Map meanings into a space

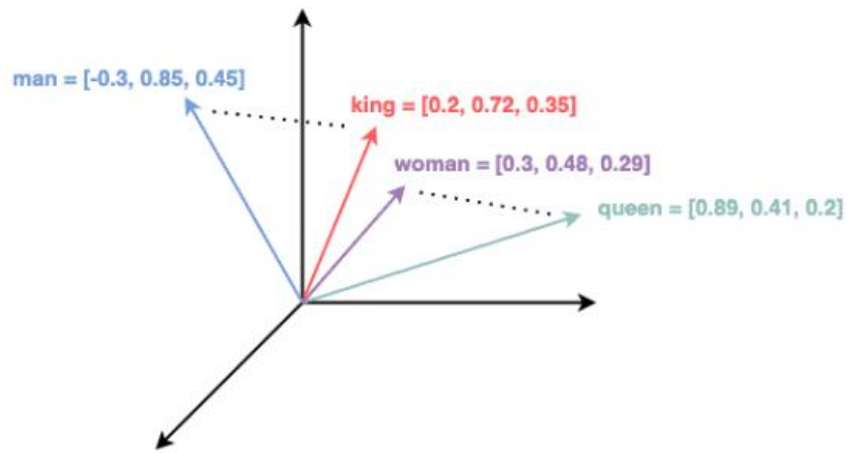


04.

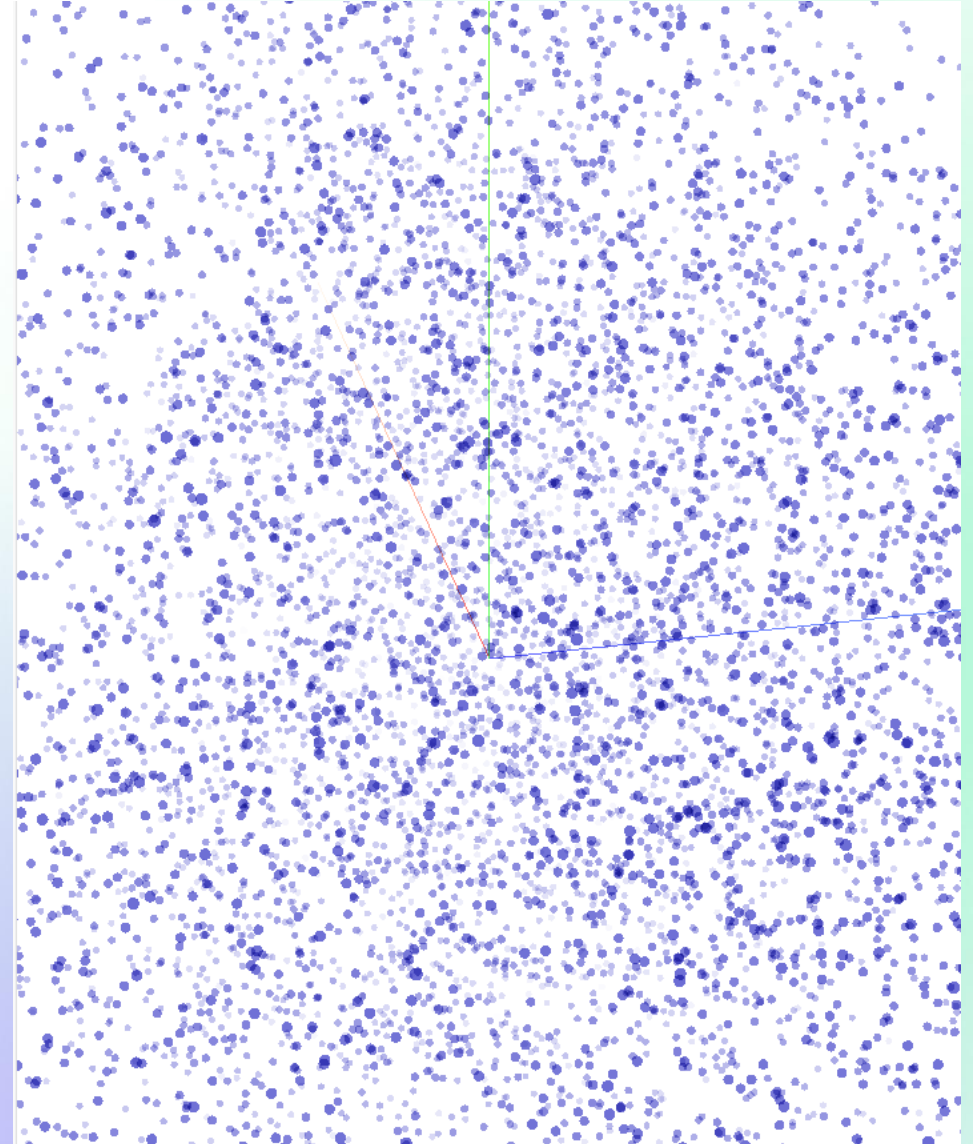
Reranking

Smartly search through your data

Embeddings quantify meaning



This is a beautiful concept! 😊



RAG consists of 4 essential Steps



01.

Parsing

Unify all data without losing information



02.

Chunking

Connect related pieces of data



03.

Embedding

Map meanings into a space

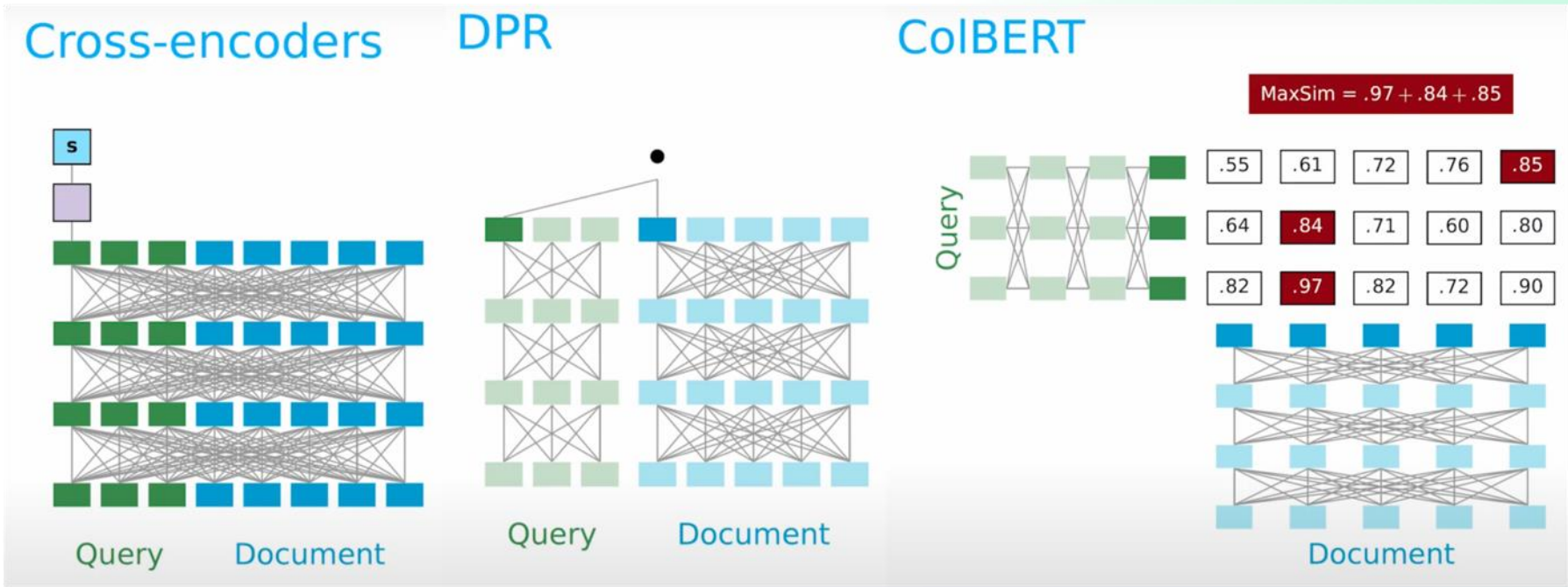


04.

Reranking

Smartly search through your data

Reranking



So many questions

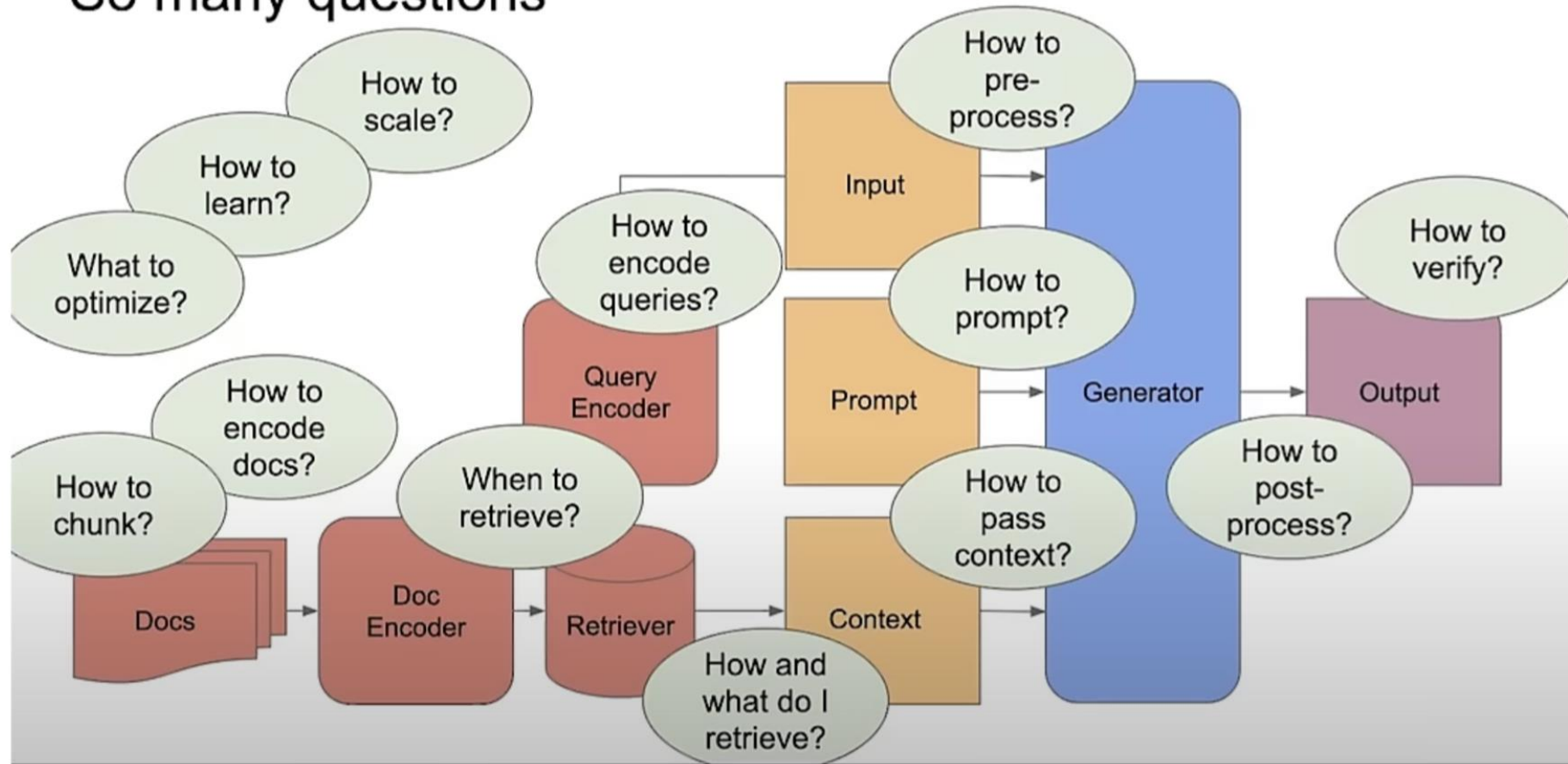


Figure 1: Overview of RAG Research Areas: Figure from Stanford Lecture CS25 V3 I

**ByteIT improves LLM Ingestion
Precision by up to 26% at less than
0.1¢ per page.**

What do we do?

Turning Documents into AI

Enabling the Custom AI Revolution



Any to RAG

We accept nearly any file type and parse it into an LLM readable Retrieval Augmented Generation (RAG) database.



Context for Quality

We markup our parsing results to ensure that structural information is preserved: Headlines, sub-headings, tables, figures, page numbers...nothing gets lost.



Individual Performance

We will offer Chunking, Rerankers and Embedding services to make sure your LLM deeply understands the content of your RAG database.

How do we do it?

Glasses for an LLM: We use **vision models** to see text like you do.



OCR

Reads Images, Cannot Generate



LLM

Cannot read Images, Can Generate



VLM

Reads Images, Can Generate



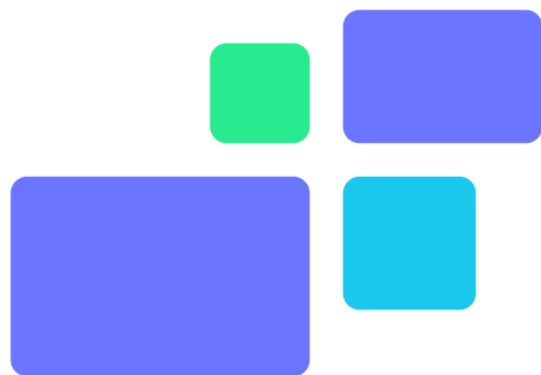
Models that read like an Expert

We trained models to segment, interpret, and structure everything from paragraphs to tables



Improving Accuracy from prototype to production

Your data. Structured. At scale.



Going beyond Text

From dense documents to visual-heavy pages, our vision-models handle diverse formats with ease.



Not just Extraction

ByteIT summarises tables, extracts fields, and chunks content for seamless LLM consumption.

How Customers find ByteIT.

01

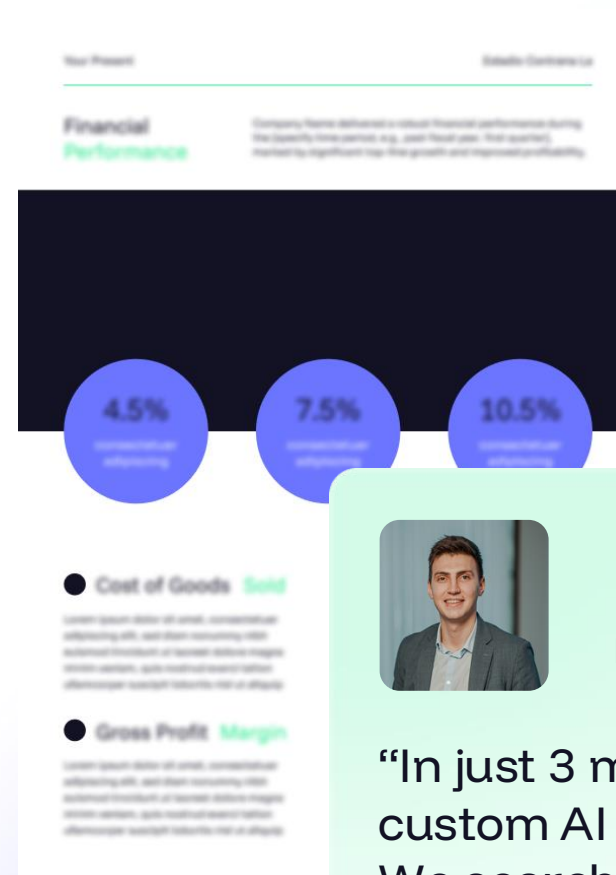
Inbound Sales

Customers are looking for ingestion solutions. We convince them to choose us.

02

Short Sales Cycles

Customer projects are often already approved and have deadlines that speed up purchasing decisions.



Prospective Customer

ML Lead @ Company Name

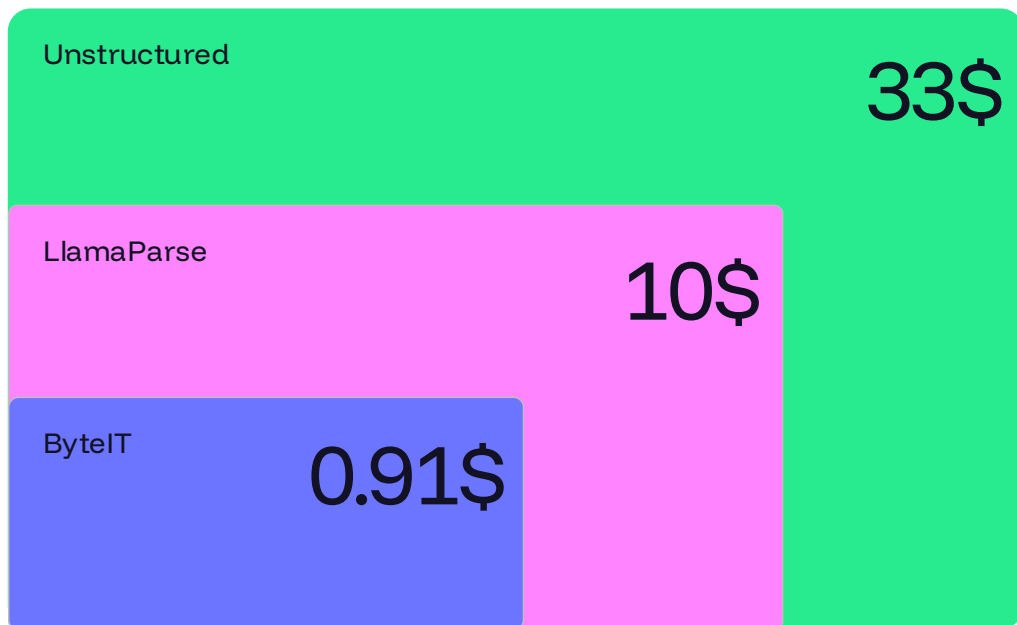
“In just 3 months we are building a custom AI to do our regulatory filings. We searched online for a data ingestion service and chose ByteIT due to its strong technology.”

Our results tell the story.

x33

Lower Costs

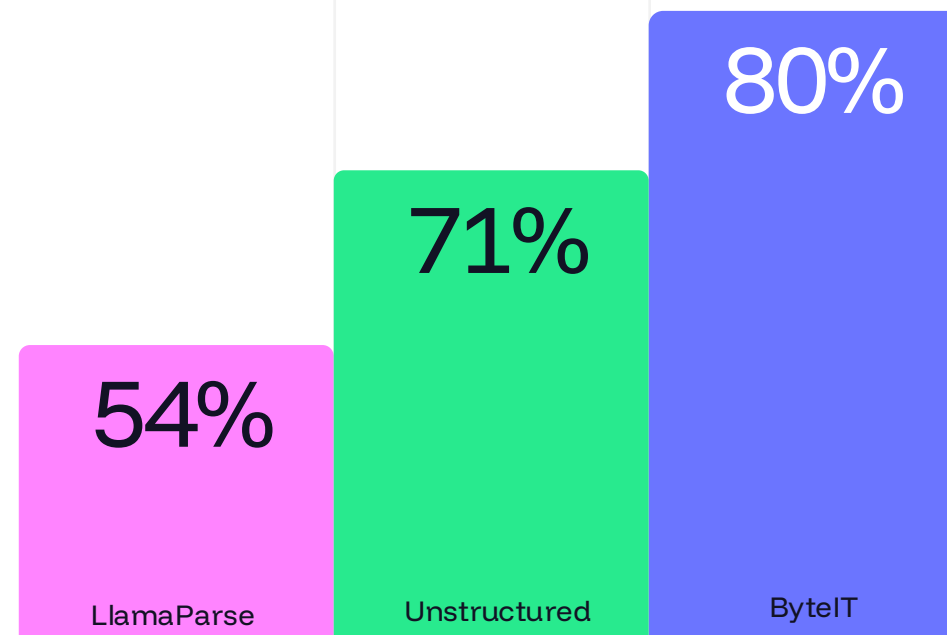
Price in USD for processing 1000 pages
Cost: ByteIT.



26%

Better Precision

On a large open-source benchmark data set.



Ready for the coming Wave

Europe's First Mover in a EUR 20B Opportunity

California-based reducto and unstructured
reached 7-digit Revenue within 1 year.

Lots of room for technical differentiation
and market expansion:



Embeddings

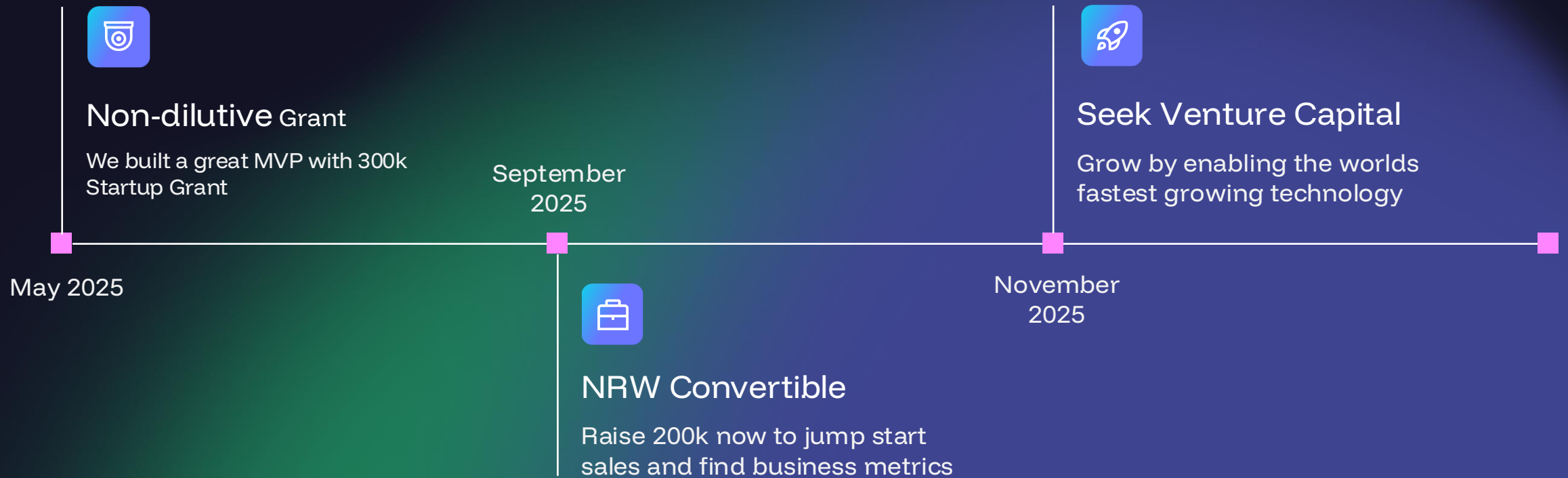


Rerankers



Specialised Agents

Where we are and where we are going



Serial Founders with real dedicated **AI Background**



Dr. Jan Bergerhoff

MSc. Economics (LSE)
PhD. Economics (U Bonn)
MSc. **Artificial Intelligence** (U Maastricht)

Cofounder CASE

Co-CEO 2016-2023

From 0 to 30 employees

>€ 3Mil. mostly priv. Investment

Technical Lead Fair Artificial Intelligence Recruiting

Project (FAIR)



Adolfo Santamónica

B.Sc. Computer Science (U Valladolid)
B.Sc. Statistics (U Valladolid)
MSc. **Artificial Intelligence** (U Maastricht)

WiDS Pioneer Winner 2024

PhysioNet Challenge 2023 top 7

Cajamar UniversityHack

2023 Winner of the local competition

36th AI CCC 2022

National winner and world top 7

Founding an IT Startup

01. Co-Founders

02. Business Model

03. Early Funding

04. Team

05. Development Structure

06. Go-To-Market

THANK YOU!

Phone

+49 174 9945799

Email

jb@byteit.ai
as@byteit.ai

Website

Byteit.ai

Social Media

[@byteit](#)