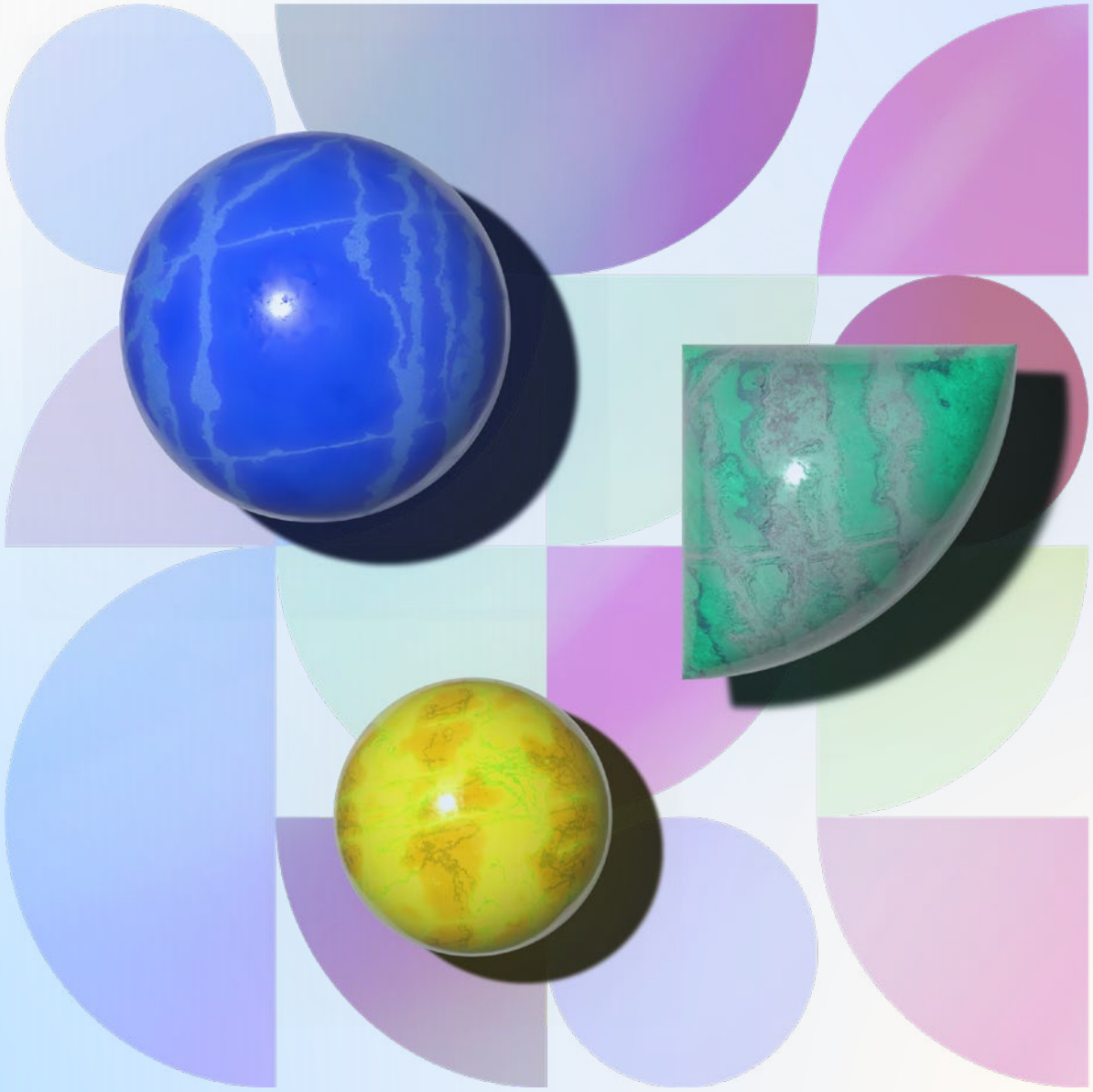
EigenTechnologies



GUIDE

A Guide to Large Language Models:

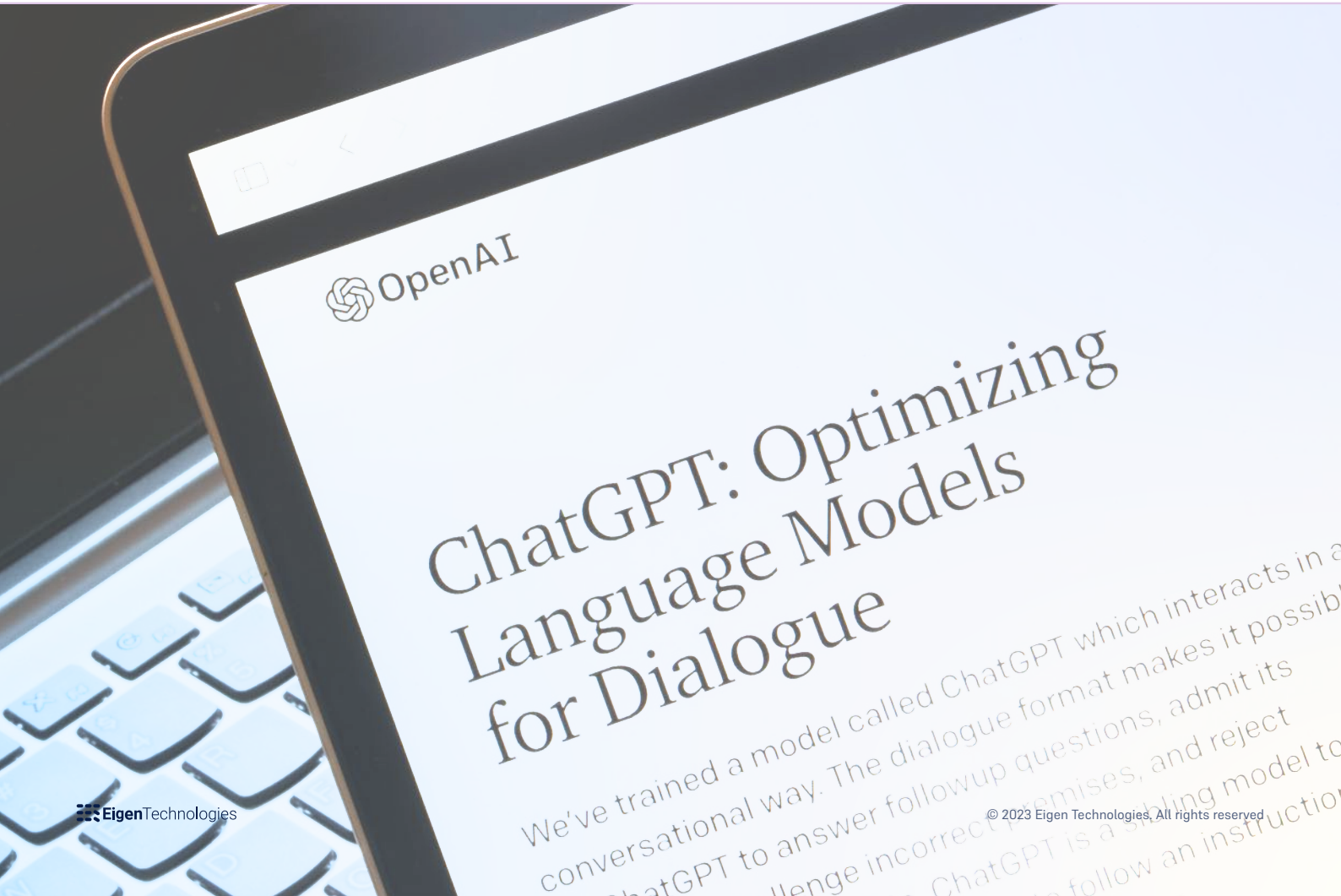
How Intelligent Document Processing Can Leverage the Likes of GPT-X

Table of Contents

1. Executive Summary

It has been impossible to avoid the hype and talk about ChatGPT, with new articles and stories breaking every day since it launched in November 2022. So, we thought it was important to delve into the topic of large language models - the type used to power the latest wave of generative AI solutions like ChatGPT – so we can explain their relevancy and application within our specialist field of intelligent document processing.

In this guide, we explore large language models (LLMs for short) and their potential applications within the world of intelligent document processing (IDP), including the stages in the end-to-end process they can support and some specific use cases within financial services, as well as some of the technical challenges and risks currently associated with using them.



2. What are GPT-4, ChatGPT, BERT and Large Language Models?

Given the topic, what better way to explain this than have ChatGPT write the introduction:

"In recent years, there has been a rapid development in the field of artificial intelligence and machine learning, leading to the creation of advanced large language models (LLMs) such as GPT-4, Chat-GPT, BERT, and others. These models are designed to process and generate natural language, allowing them to perform tasks such as language translation, sentiment analysis, and even creative writing. As the technology behind these language models continues to improve, they are becoming increasingly prevalent in various industries, from marketing and advertising to healthcare and finance. In this blog post, we will delve into the world of large language models, exploring what they are, how they work, and their potential applications."

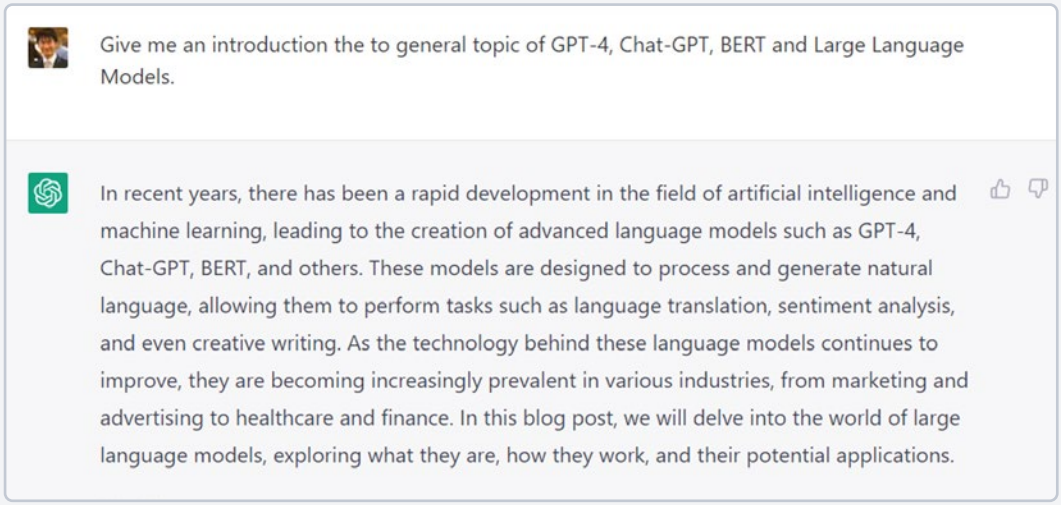


Diagram 1:

ChatGPT-generated introduction to the topic of GPT-4, ChatGPT, BERT and LLMs

This article by Stephen Wolfram also offers a good semi-technical explanation of Chat-GPT.

3. How do Large Language Models Fit into Intelligent Document Processing?

Imagine building a fighter jet. You need to bring together a ton of different components from structural pieces like wings and the fuselage, to more complex components like the engines or avionics. A working fighter jet is only complete when all these pieces come together.

Building an IDP platform is similar. There is an entire infrastructure and workflow pipeline that is analogous to the fuselage and wings. There's a need for model training and model governance just like avionics. And of course, you have the actual interpretive AI models themselves that are the engine that powers your machine. Diagram 2 below, shows all the components that make up the Eigen IDP platform.

IDP Journey from Document to Data

Requires a large number of discreet tasks working together

1. Eigen can leverage either BERT or GPT-X to optimize results; scale of LLMs to scale from OpenAI APIs to smaller on-premise deployments depending on cost/compliance requirements
2. Primarily for tables understanding
3. Regulation, compliance, model risk management policies, or data policies preclude the ability to leverage LLMs

Diagram 2:

Intelligent document processing components

To further complicate matters, in the world of IDP, it is not simply a case of querying the text strings within a document. There are additional complications, such as the visual language of the documents to be processed. These can take the form of tables (a huge problem), checkboxes, or graphs. This introduces the problem that AI model(s) need to be multi-modal – marrying natural language processing (NLP) (such as GPT-X) with machine vision.

So, what does this mean in the context of GPT-4? In the fighter jet analogy, it means that we can swap out ***one core component of the engine for another component to enable the jet to fly faster and be more manoeuvrable under certain conditions.*** Reverting to IDP, this means that GPT-4 can help extract and interpret *certain* questions in *certain* (*mostly text-only*) documents better than previous models. In the world of generative AI and chat, GPT-4/ ChatGPT is revolutionary, but in the world of IDP, it is essentially 'just' an additional tool to be "*potentially*" swapped out under certain conditions.

This is not meant to downplay the potential impact. Going back to the fighter jet analogy, this could mean a cruise speed of Mach 1.5 vs an older version of Mach 1. While it doesn't fundamentally change the rest of the plane, it is significantly more performant for certain tasks. Similarly, in the IDP world, a bigger, more powerful LLM (like GPT-4) can mean better data extraction accuracy and flexibility and the ability to deal with more complex tasks that were previously more rules-based. While LLMs cannot make an impact everywhere in the IDP data flow, more powerful LLMs like GPT-4 may open the ability to do new tasks previously thought impossible like 'chatting with your documents live', an exciting functionality we are releasing in Q2-23.

4. Opportunities with Large Language Models/GPT-4 in End-to-End Intelligent Document Processing

In table 1 below, we list the various stages and components of IDP (as shown in diagram 2) and how LLMs and GPT-X can potentially improve each one. Of the 24 components, we identified six where LLMs/GPT-X have a high potential of improving results, five where there's potential for them to deliver some improvements and 13 where there is no potential or expected improvement from the addition of LLMs/GPT-X.

STAGE	TASK IDP/ COMPONENT	DESCRIPTION	TODAY	GPT/LLM POTENTIAL	2023 EIGEN ROADMAP
Pre-processing	Machine Vision (MV): OCR (Typed)	Transforming scanned documents into machine-readable format (e.g., JPEG to searchable PDF)	Multiple options: Commercial OEM (ABBYY, Kofax), Open Source (Tesseract)	None	Continuous upgrade as new commercial or open-source OCR components improve over time
Pre-processing	MV: OCR (Handwriting)	Transforming handwriting images into searchable text	2023 Roadmap Item	None	2023 Roadmap Item via commercial OEM partner
Pre-processing	Natural Language Processing (NLP): Classification	Classifying documents into categories (e.g., is this a mortgage application vs a passport vs a bank statement?)	Eigen Proprietary	**High – may increase accuracy/speed**	**2023 R&D item**
Pre-processing	NLP: De-blobbing	Splitting up documents that may have been scanned into a single PDF	Eigen Proprietary	**High – Improved classification will result in a knock-on effect on de-blobbing**	**2023 R&D item**
Pre-processing	MV: Check-boxes	Reading check-boxes (usually hand-written) into machine-readable data	Commercial OEM Partner	None	Continuous upgrade as new commercial or open-source check-box detection components improve over time
Pre-processing	MV: Tables Detection & Reconstruction	Detecting and transforming table images (e.g., in scanned PDF) or PDF tables (e.g., messy XML) into clean machine consistent HTML or JSON table data structures	Eigen Proprietary Table Foundational Model	None	Continuous upgrades (including table normalisation) on performance and adding new sources of training data to the foundational model
Post-processing	Platform: Parent/ Child Mapping (post classification and discovery)	Linking documents into groups of related families (e.g., master contract doc with amendment docs and schedule docs	Rules-based approach (post extraction) to link docs	None	Future R&D Item
Single Document Intelligence	NLP: Point Extraction (Regulated)	Extracting specific data points (a value, an entity, a short phrase, a date) from a document	Eigen Proprietary (including Eigen domain specific topic and language models)	None – the approach today serves a very specific data compliance and model governance purpose	Continuous upgrades on performance and feature engineering
Single Document Intelligence	NLP: Section Extraction (Regulated)	Extracting longer phrases, clauses, sections	Eigen Proprietary (including Eigen domain specific topic and language models)	None – the approach today serves a very specific data compliance and model governance purpose	Continuous upgrades on performance and feature engineering
Single Document Intelligence	NLP: Instant Answers or Question-Answering Point/ Section Extraction	Ability to ask questions and get an answer back ('single shot extraction' or 'question answering', sometimes called 'chatting with your docs')	Information Retrieval (IR) with Eigen Modified Domain Specific BERT (an LLM)	**High – GPT-4 may perform significantly better than BERT (or other LLMs) on many documents**	**From Apr-23 clients will have the ability to switch between GPT-X and Eigen Domain Specific BERT (an LLM)**

Single Document Intelligence	Multi-modal: Table Extraction	Extracting a specific table in a document and outputting that into an easily digestible format (CSV/XLS/JSON/HTML)	Eigen Proprietary Table Foundational Model in conjunction with an Eigen Proprietary table extraction model	Potential to improve performance of extraction based on text inside the table (currently using a non-LLM approach); potential multi-modal GPT release	Improvements in model governance and user workflows
Single Document Intelligence	Multi-modal: Cell-level tables extraction	Extracting specific cells out of specific tables even when tables can be heterogenous across documents (e.g., 'extract the Group EBTIDA for 2022 across all annual reports)	Eigen Proprietary Table Foundational Model in conjunction with an Eigen Proprietary table extraction model	Potential to improve performance of extraction based on text inside the table (currently using a non-LLM approach); potential multi-modal GPT release	Improvements in model governance and user workflows
Single Document Intelligence	Multi-modal: pure multi-modal point extraction	Extracting information from highly visual heterogenous documents (e.g., PowerPoint, complex invoices, certificates, etc...)	Eigen Proprietary incorporating both Eigen's MV and NLP capabilities	**High – similar to instant answers, LLMs/GPT can be used to improve results as a base language model; potential multi-modal GPT release**	**Transition from an API-only service to a no-code service**
Single Document Intelligence	Correlated Answers	Correlating different answers to different variables (e.g., interest rate for a loan may differ for different facilities, need to correlate the exact interest rate for each facility)	Eigen Platform's rules-based approach via API plug-ins	Medium – GPT's natural language understanding can potentially speed this up significantly and enable the extraction to move away from rule-based correlations	2023 R&D item
Post-processing	No-Code Business and Legal Logic Builder	Interpreting extracted results based on business rules set by user	Eigen Platform	None	Improvements in user workflow
Post-processing	Data transformation and advanced logic	Leveraging either commercial no-code partners or python plug-in scripts to undertake data transformations or more advanced business logic tasks	Eigen Platform (python plug-ins) or commercial partner (e.g., Unqork, Xceptor, Microsoft)	**High – Potential to automate python scripts using Chat-GPT or Co-Pilot**	**Eigen's own solutions team is experimenting with this today for plug-in writing**
Post-processing	NLP: Clause and Section Comparison	Ability to compare clauses or sections against a 'gold standard clause' or across each other to ascertain risk, deviations, etc..	Eigen Proprietary	**High – Potential for LLMs to improve comparisons in a more organic way**	**Q1 Release**
Post-processing	Platform: Parent/ Child Roll-up Logic	Understanding linkage logic (e.g., if an amendment supersedes the original master contract)	Rules-based approach to determine roll-up logic	Potential for LLM to understand roll-up linkage logic innately without rules (likely not possible with GPT-4 but potentially future LLMs)	Future R&D Item
Post-processing	Human-in-the-loop review	Workflow enabling humans to review machine results	Eigen Platform	None	Improvements in user workflow
Ops: Machine Learning (ML) Ops	Model Governance	Providing automated cross-validation services pre-production; managing model drift in production; managing high/low confidence scoring to determine exception handling; reporting for model risk management	Eigen Platform with Eigen open-sourced cross-validation techniques	None	Improvements in user workflow
Ops: ML Ops	Annotator	UX and backend system to enable human users to easily train ML models	Eigen Platform	None	Improvements in user workflow
Ops: Workflow	Workflow Orchestrator	Platform to orchestrate tasks and document flow	Eigen Integration Pipeline (EIP), Unqork, Xceptor, or Microsoft	None	Addition of more no-code components
Ops: Workflow	Document and Answer Server	UX and API endpoints to serve documents (especially large documents) and answers (in a variety of ways)	Eigen Platform	None	Better API endpoints for partners
Soloutions	Pre-build Eigen or Partner Industry Solutions on Eigen Platform	xx	xx	LLMs/GPT can improve solutions insofar as each task module improves performance	

Table 1: Opportunities for improvement within IDP for LLMs/GPT-X

5. What Does this Mean for Key Financial Services-Specific Intelligent Document Processing Use Cases

We know from experience that customer workflow requirements, and the documents they need to process, are often highly complex and domain specific. In table 2 below, we take a look at some key banking and financial services use cases where LLMs could be applied for document to data transformation.

DOCUMENT TYPE	TYPICAL DOCUMENT FORMAT	EIGEN POINT + SECTION EXTRACTION	BERT (alone)*	EIGEN W/BERT (with Eigen Instant Answers)*	GPT-4 (alone)*	EIGEN W/CHATGPT (with Eigen Anstant Answers)*
ISDA + Schedule	40,000 words Includes semi-heterogeneous tables	90%+ F1 (PE+SE)	N/A - Too long	90%+ F1 (PE+SE)	N/A - Too long	60%+ F1 (PE) 90%+ F1 (SE) **
LSTA Loan Agreement	100,000 words Includes mostly heterogenous tables	90%+ F1 (PE+SE)	N/A - Too long	90%+ F1 (PE+SE)	N/A - Too long	60%+ F1 (PE) 90%+ F1 (SE) **
Collateralized Loan Obligation (CLO)	500,000 words Includes semi-heterogeneous tables across docs	90%+ F1 (PE+SE)	N/A - Too long	90%+ F1 (PE+SE)	N/A - Too long	60%+ F1 (PE) 90%+ F1 (SE) **
Brokerage Account Statement	1000 words Includes completely heterogenous tables across docs	N/A – needs machine vision (e.g., Eigen's tables)	N/A – needs machine vision (e.g., Eigen's tables)	N/A – needs machine vision (e.g., Eigen's tables)	N/A – needs machine vision (e.g., Eigen's tables)	N/A – needs machine vision (e.g., Eigen's tables)
Money Transfer Email	50 words Pure String	80%+ F1 (PE+SE)	90%+ F1 (PE+SE)	N/A – information retrieval (IR) not needed	99%+ F1 (PE+SE)	N/A – IR not needed
Bloomberg Chat	25 words Pure String	N/A – not suitable	80%+ F1 (PE+SE)	N/A – IR not needed	99%+ F1 (PE+SE)	N/A – IR not needed

Table 2: Financial services specific use cases where LLMs/GPT-X could be applied

* F1 score is an evaluation metric that measures a model's accuracy. It combines the precision and recall scores of a model.

** Eigen with ChatGPT excels at replicating results for section extraction but required significant post-processing to replicate point extraction due to the generative and non-deterministic nature of its outputs, giving the current Eigen point extraction model still a distinct advantage, as such the exact F1 score leveraging ChatGPT was much lower for point extraction type tasks vs section extraction type tasks. We plan to improve Eigen GPT point extraction Q&A and re-benchmark.

In conclusion, and as expected, ChatGPT specifically is much better suited than previous information extraction techniques for 'regular human language' like chat or email. However, there is architectural work required to get any LLM (including BERT or GPT-X) to work for longer, more complex, and domain specific document types. Specifically, the information retrieval or section extraction layer must be accurate, which benefits significantly from domain specific fine-tuning.

From an Eigen perspective, we have always embraced the power of LLMs since BERT came out in 2019, and we are excited to offer clients a GPT-X option (as an alternative to our existing domain specific BERT offering) in the coming weeks. I include the example of brokerage account statements, which are heavily heterogenous tabular to serve as a reminder that IDP is a heavily multi-modal and domain specific affair.

Looking at table 2 above, regarding longer, more complex, finance specific documents, from a pure accuracy perspective, as long as there is a domain specific fine-tuning with information retrieval (IR) (such as Eigen's IR model to pre-process for BERT or GPT-X), accuracy rates with Eigen's traditional probabilistic graphical models driven point extraction is in line with Instant Answers (Question Answering) that leverage either BERT or GPT-4. As such, deciding which technology to use will be dependent upon:

1. **Cost**
   (traditional Eigen point extraction and BERT are orders of magnitude cheaper than GPT-4, but GPT-4 is much more accurate for shorter, more 'natural human' text where cost is less of an issue)
2. **Model Governance**
   (LLMs make model governance more complex and therefore significantly more costly)
3. **IP/Privacy**
   (a potentially huge issue with GPT which needs to be mitigated)
4. **Model Output Parameters**
   (BERT is great for data extraction due to its bi-directional nature; GPT-X requires significant pre and post processing due to its generative nature, these issues around GPT need to be better addressed if used in larger scale production)

6. The Challenges of Making Large Language Models Work in the Real Intelligent Document Processing World

Looking at the task of data extraction for documents, it is tempting to upload the entire string of text into GPT-4 with a prompt like '<document text> In the text above, what is the [x]?'. If we do that and try to set a prompt of a document that is around ~100k words long (common length for complex financial documents like loan agreements), it will not work. Even with the new and improved GPT-4, there is still a 25k word limit. Obviously GPT-5 and beyond will provide advancements but even if you find an instance that can handle 100k words, the cost of LLMs scale quadratically so extrapolating OpenAI's cost model, this could cost $14 per query.

To put this into context, a London-based paralegal will charge around $10 per query. Eigen's traditional point extraction has a compute cost of around $0.00014 per query to accomplish the same extraction task, and Eigen's BERT-based Instant Answers has a compute cost of around $0.0042 per query. This cost problem associated with document length and these LLMs, can be mitigated through pre-processing models like information retrieval (IR). Diagram 3 below, shows how information retrieval models mitigate this challenge.

The Role of Information Retrieval Models

In overcoming the cost of using LLMs

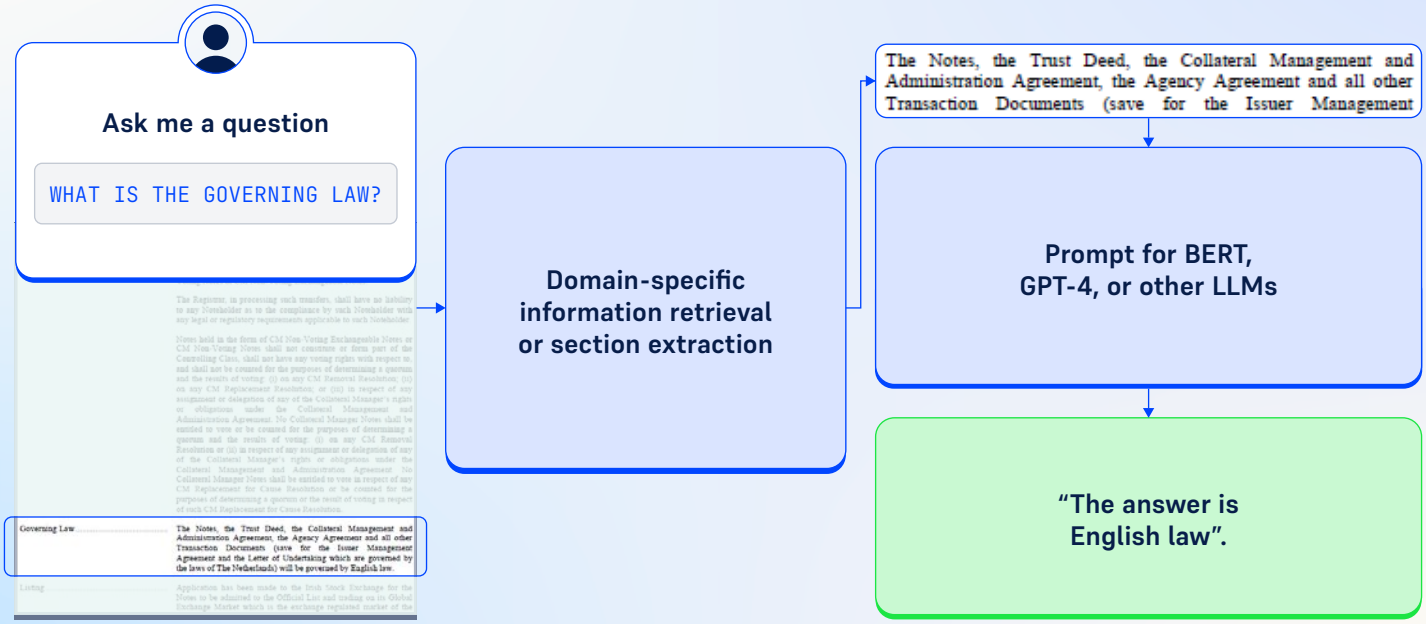


Diagram 3:

How IR models help to solve the document length vs cost challenge of LLMs

So, let's say then that you have some other lightweight (hopefully also domain-specific) model that finds the relevant section of the document via information retrieval, and you provide the prompt for just that section. You potentially solve the length/cost problem, but now you run into the fact that OpenAI's models are general and not domain-specific (or specific to the given use case). Moreover, it tends to pull from 'general internet knowledge' versus taking in the entire context of the document (as we just truncated the document to save on compute costs). In table 3 below, we highlight some of the key performance and accuracy challenges of using LLMs (such as GPT-4) out of the box and how to mitigate for them.

Note: we'll cover privacy and model governance separately in the next section as those risks are harder to mitigate.

CHALLENGE	DETAILS	MITIGATION TECHNIQUE
Word/token limit	GPT-4 has a 32k token limit/25k word limit, which is enough for many use cases but not some of the high value use cases within finance, insurance, legal, etc.	Leverage domain-specific information retrieval (IR) or section extraction (SE) techniques to partition the document into a smaller section to use as prompt.
Compute or API costs	LLMs are extremely computationally expensive (both in terms of time and money) and these costs scale quadratically the longer the prompt.	IR and SE as above or consider using alterative frameworks if cost (money and/or time) is a key consideration.
Domain or Use Case Specificity	Generic LLMs (BERT, GPT, etc) are trained on 'general internet knowledge' and not domain-specific documents, which can lead to substandard results.	Provide domain -pecific fine-tuning to LLMs (using large, annotated datasets) or provide additional domain- specific rules on top of extraction results.
Generative models do not provide provenance	Generative models (like GPT but not BERT) do not out of the box provide provenance of where the information was obtained, thereby increasing risk especially in highly- regulated use cases.	IR can help narrow down the sections to analyse by GPT; specific prompt engineering and string/coordinate matching is needed to enable this functionality.
Hallucination	The LLM generates results outside the context of the prompt, providing false or nonsensical answers.	Tighter prompt engineering and IR and/or SE input generation.

Table 3: Performance and accuracy challenges of LLMs and how to mitigate for them

7. Key Risks Associated with using Large Language Models Blindly

As mentioned in the previous section, two of the key risks that LLMs present when being used for document processing are model governance and IP/data privacy. We cover both below.

MODEL GOVERNANCE

Most regulated companies such as banks, insurance companies, healthcare payors/providers have fairly involved machine learning (ML) model governance and model risk management processes. This is to prevent bias or inaccurate automated decisions or results being generated by their ML/AI models. Moreover, there is typically a stringent process to approve certain models for production.

In the world of IDP, understanding answer confidence, estimating accuracy rates pre-production, and ensuring a smooth flow of model governance metrics are communicated to the user are all key to successful production. Unfortunately, LLMs by themselves do not provide a calibrated confidence score in terms of how accurate an answer might be. Hence, these LLMs need to be placed within the right infrastructure to ensure smooth productization and confidence calibration. At Eigen, we have developed a full model governance and human exception handling framework to ensure the maximization of accuracy and value with the minimization of model risk.

IP & PRIVACY

Regulated entities dealing with extra sensitive information such as banks, insurers, and healthcare companies cannot risk data leaked and integrated into the LLMs generated by OpenAI. JPMorgan Chase recently clamped down on the use of ChatGPT for compliance reasons according to a news report by CNN Business and the Italian government recently banned GPT on the basis of user privacy concerns.

The New York Times covered a story recently about a lawsuit involving the misuse of AI training data. The risk of leakage and misuse of training data for large-regulated entities could be existential for those institutions. Semiconductor engineers at Samsung unwittingly shared confidential data including source code for a new program when they used ChatGPT to help with some tasks.

Coming back to the world of IDP, at Eigen we process some of the most sensitive documents for institutions – both commercially and personally sensitive – including derivative contracts, detailed insurance claim reports, personal healthcare information, personal financial details, etc. As such, our entire company and infrastructure is set up to ensure that AI is leveraged safely; something we have invested significantly in since the foundation of our company.

While it is tempting to throw documents into GPT, there is a minefield of IP, privacy, and safety concerns to consider – and building the infrastructure to handle those is limited to only a small handful of IDP players, such as Eigen.
Game-changing advances in technology always bring about near hysteria about the revolutionary impacts, both positive and negative. Organizations with highly valuable information should always temper their excitement with the potential downside risks. There is a path to success but it's important not to ignore the path to failure.

It's hard to see how ChatGPT complies with current Privacy standards such as GDPR/CCPA and much more research is needed to understand how it meets these requirements. GCHQ advises users not to include sensitive information in queries or anything that could lead to issues if everyone saw their queries, when using AI bots like ChatGPT. The Telegraph reports that City firm Mishcon de Reya has banned its lawyers from typing client data into ChatGPT over security fears, as has Accenture. Organizations with highly valuable data should consider the impact on their organizational reputation of using these tools.

Security decisions need to be made based on understanding the risks and with emergent technology like this it is always a challenge to quantify the risk. It's not the right place to put your most valuable and sensitive data until we know significantly more than we do today. How would you explain to your clients and board the rationale for such a decision?

8. Conclusion

There is no doubt that LLMs and generative AI solutions like ChatGPT drive efficiencies and create opportunities for organizations in many new and exciting ways. To go back to our fighter jet analogy, in the world of IDP, they are a useful component to make the engine faster, more flexible, and more accurate dependent upon the use case(s) in question. But domain-specific fine-tuning, model risk management/governance and security controls are required to make LLMs fit for purpose when it comes to harnessing their power for document processing purposes. IDP specialists like Eigen, able to leverage LLMs in a compliant, cost-effective, and domain-relevant way, offer the best solution for enterprises looking to unlock data from documents using cutting-edge AI.

Learn how the Eigen IDP platform can be applied to a broad spectrum of use cases and document types by visiting our Solutions section or request a demo to see the platform in action for yourself.

ABOUT EIGEN TECHNOLOGIES

We're a team of data scientists, product engineers, solution consultants, and customer services experts with offices in New York, London and Lisbon.

Our mission is to supercharge the way organizations operate by unlocking the value of information and insights trapped in documents. We exist to make data useful and processes seamless for our clients.

The Eigen AI-powered no-code intelligent document processing platform enables firms to extract, classify and interpret virtually any information from any document to make smarter business decisions, eliminate manual processing and optimize the flow of data between systems and people.

Eigen uses natural language processing and cutting-edge machine learning to automate the extraction of answers from documents and can be applied to a wide number of use cases. It understands context and delivers better accuracy on far fewer training documents with market-leading information security.

Our customers include some of the most well-known and respected names in business.

EIGEN CUSTOMERS:

AVIVA · BANK OF AMERICA · BNP PARIBAS · Deloitte. · FORTRESS · Goldman Sachs · HISCOX · ING · orrick

EIGEN INVESTORS:

Goldman Sachs · TEMASEK · LAKE STAR · dawn. · ING · anthemis · Sony Innovation Fund SONY

eigentech.com