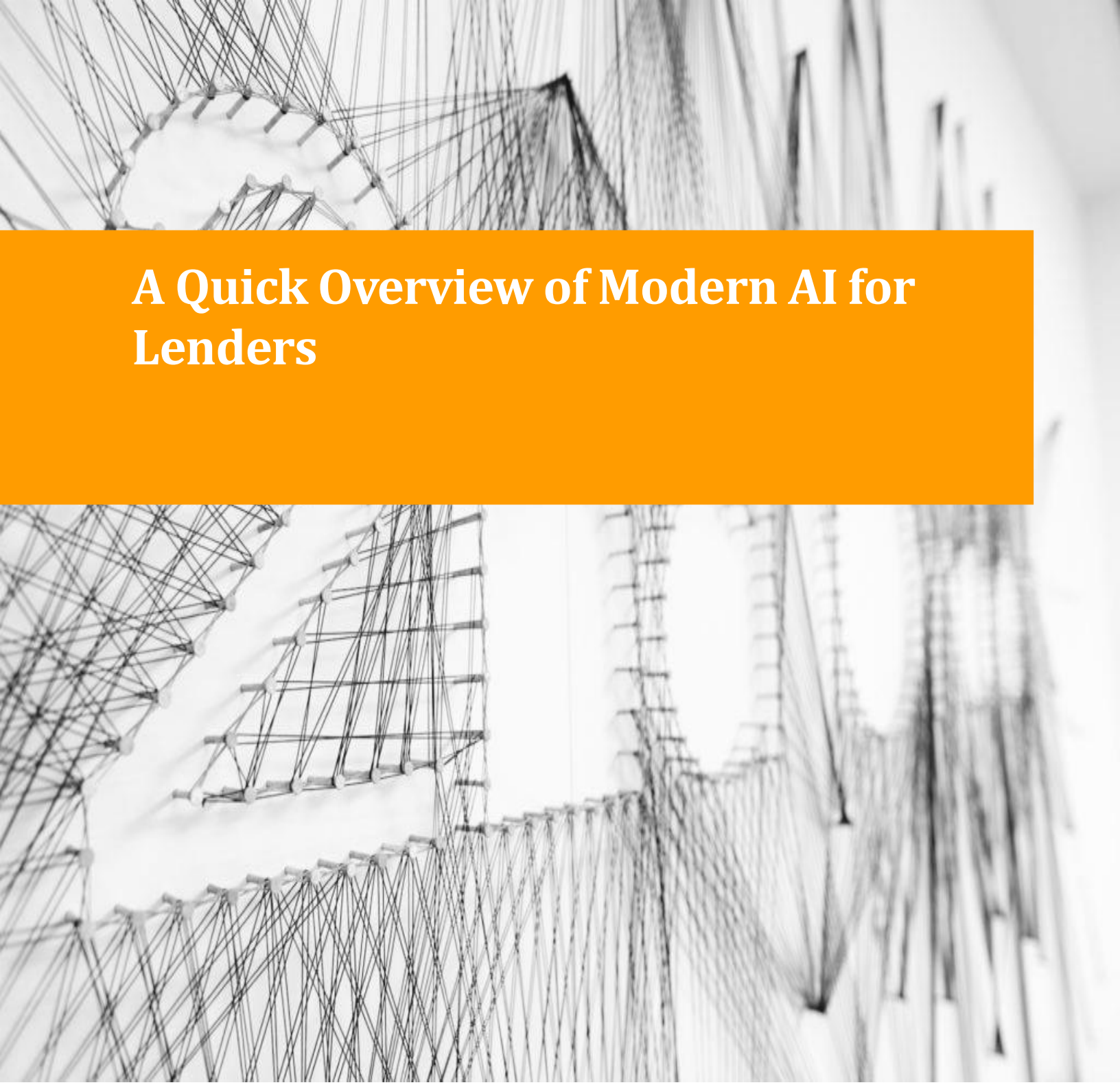

A Quick Overview of Modern AI for Lenders

Aaron McGuire, Syed Raza

March 2024

2nd Order solutions

Executive Summary

This fact sheet provides insights gleaned from research encompassing various sources, public research, and conference attendance by 2OS analysts in the preceding quarter. It offers a snapshot of the current state of Artificial Intelligence (AI) and aims to address key questions and topics including:

- Nuances in Large Language Models (LLMs)
- The intersection of automated decisioning and AI
- Potential risks of AI in transaction servicing

This document delves into each point, providing context, caveats, and references to support the analysis. It is intended to be a resource for understanding the landscape of AI today, with a focus on ensuring regulatory alignment.

Contents

Background

Defining AI: Artificial Intelligence vs. Machine Learning

Before delving into specific topics, it's crucial to clarify the distinction between artificial intelligence (AI) and machine learning (ML). While we acknowledge that many people include ML under the umbrella for AI, we believe that keeping them distinct is beneficial for understanding AI and all its benefits and risks. Over the past 5-10 years, numerous lenders have adopted machine learning models to enhance their ability to forecast consumer risk, both for new customer acquisitions and existing consumers. It's common for these lenders to unsuitably label these prediction models as "AI", even though they don't exhibit the characteristics typically associated with artificial intelligence in contemporary understanding. We contend that within lending, machine learning and artificial intelligence represent two discernibly distinct types of models – while ML models are widely adopted by leading issuers today, AI models, including but not exclusively Gen AI, are still in the early stages within lending.

Machine learning represents a model where it is difficult but possible to untangle the exact decision process of the model. The predominant ML models in lending are tree-based models such as Gradient Boosted Machines or Random Forests, which trace back decisions to intricate, yet specific decision trees constructed from real consumer data. The generated outputs of these ML models manifest as numeric values, which require the use of Partial Dependence Plots or Shapley values to understand model reasoning at scale. When built correctly within a responsible modeling framework, these ML models can be trusted to make sensible decisions that adhere to fair lending requirements. While ML models can be self-updating, when used in lending, these models typically operate as static entities that stay consistent over time.

Artificial intelligence, often referred to as Generative AI (GenAI) represents a highly sophisticated branch of machine learning that transcends conventional ML models. It leverages neural networks and deep learning as foundational structures to create new data from intricate, evolving prompts. Unlike ML models, which typically yield numeric outputs, GenAI creates *new data.* Outputs can vary across multiple formats including textual, video, audio, and image outputs, alongside semi-numeric summaries of textual categorization. Unlike the static nature of ML, modern GenAI solutions constitute dynamic "model systems" subject to constant evolution. These innovative tools find application in diverse areas such as synthetic data generation, the development of chatbots/virtual assistants, document summarization, and the personalization of creative content. A summary of enterprise projects using GenAI today can be found in ***Figure 1*** below. There has been a lot of hype around Gen AI use cases in financial services for the last year or so but we are starting to see measurable impacts of the new technology as shown in some examples in Figure 2.

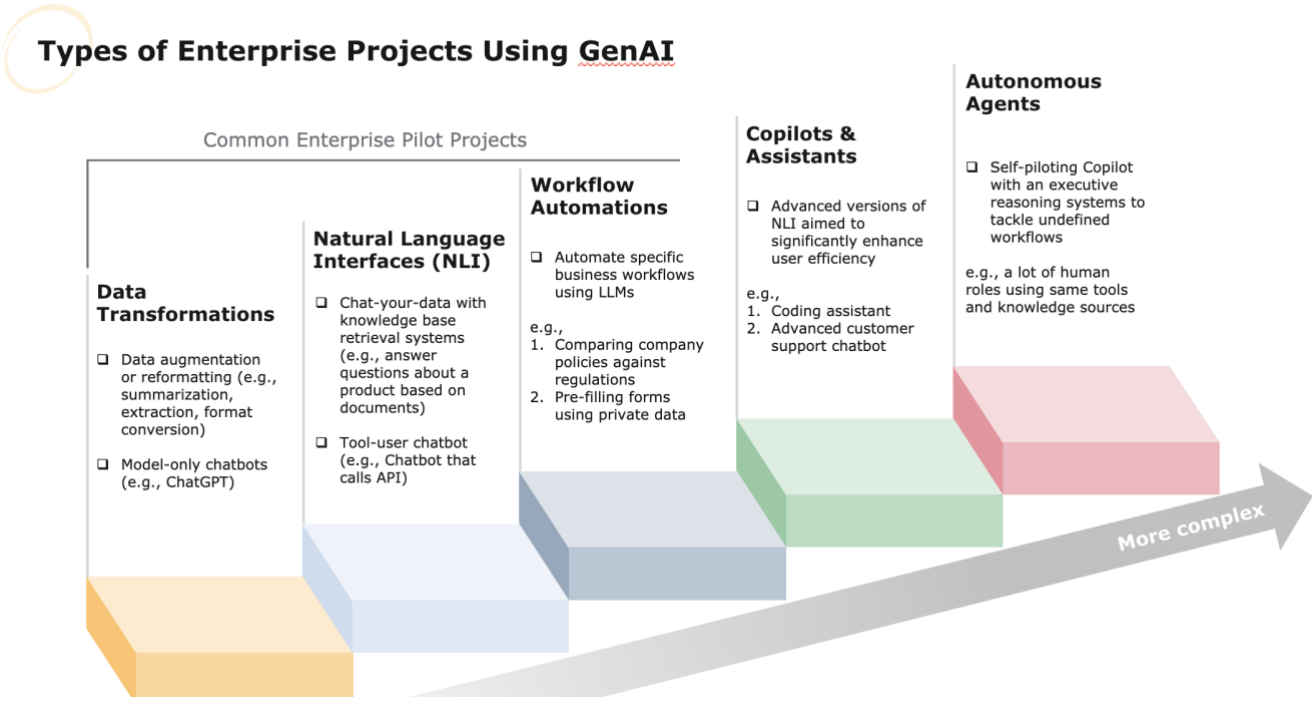

***Figure 1**: Common uses of GenAI by order of complexity.*

Most modern discussion about Generative AI focuses on **Large Language Models (LLMs)**, a series of models within the Generative AI umbrella designed to achieve general-purpose language generation and complex pattern generation by pre-training complex network models on massive datasets. Large language models are generally thought of to have language outputs, but some modern models also have multimodal output, allowing users to generate things like audio, images, videos, and other media from user prompts.

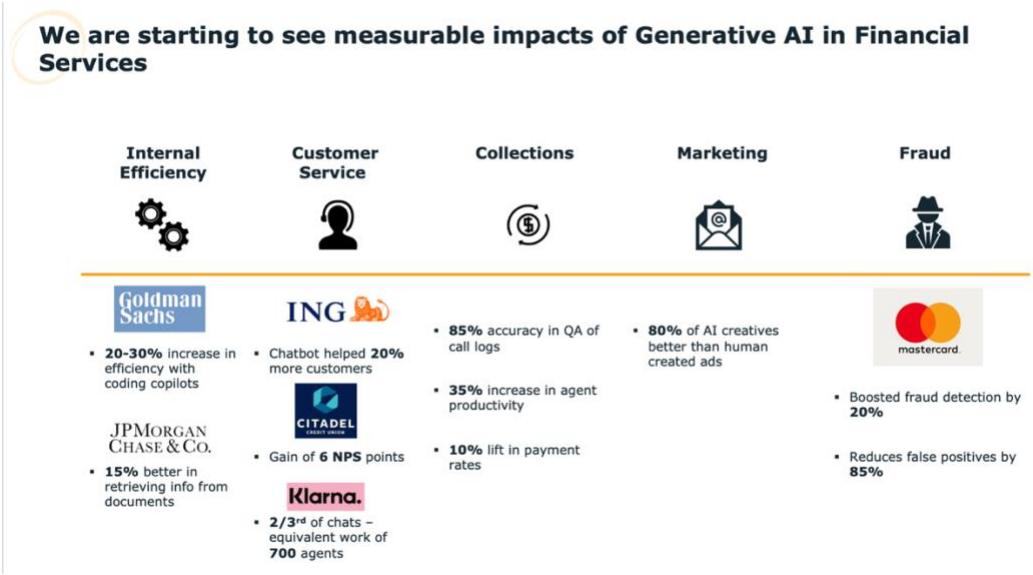

Figure 2: Initial Impacts of Generative AI in Financial Services

Hot Topics in Artificial Intelligence

I. Nuances of Large Language Models

As lenders begin to price out and decide their approach for their personal Generative AI journey, one of the questions we've constantly fielded is one of model type. It's often difficult to figure out the ideal LLM for your specific use case, and the architecture surrounding the different models can be both highly distinct and opaque. Which begs the question – **what nuances should one be mindful of when distinguishing between various LLMs such as Llama2 and ChatGPT?**

The most common method of distinguishing between different LLMs is by their parameter count – the more parameters a model uses, the more complex its architecture. More parameters indicate that the underlying network structure is utilizing a greater number of "neurons" to make connections and build answers. Noteworthy public LLMs and their respective parameter counts include:

- ChatGPT 3.5 (350 billion parameters)
- Llama2 (70 billion parameters)
- ChatGPT 4.0 (rumored to be around 1 trillion parameters)

It's worth noting that parameter count doesn't always directly correlate with effectiveness – Llama2, for instance, outperformed ChatGPT on multiple tests in late 2023. Other factors influencing an LLM's effectiveness include architecture, the quality of training data (whether raw or curated), and the crucial role of "context length" in model performance.

When evaluating different foundational models against each other, it's important to categorize the competing models and deciding which categories are "must-haves" for your organization. In our experience, the categorizations we have found most useful for the broadest group are:

- Open-source versus closed-source (i.e., "can others access your model's build code?" … in most cases, the training data isn't accessible to others, which can be a major limiter depending on the internal banking use case)
- Language versus multimodal (i.e., "can the model generate responses beyond text, such as images, audio, video, or the necessary system responses for our basic operational software?")
- Foundational models versus derived models (i.e., "is the model built upon another existing model, or does it constitute its own design?")

Common metrics used to differentiate these models performatively include:

- Size (parameters): indicates the number of parameters used to train the model, and how many "neurons" the model fires (as aforementioned)
- Pretraining length (tokens): denotes the length of data used for model training, with Llama2 featuring a pretraining length of 2,000 tokens, equivalent to approximately 750 words for 1,000 tokens
- Context length: represents the maximum number of tokens a model can retain when generating new text, with Llama2 boasting a context length of 32,000 tokens.

Of all these metrics and categorizations, **open-source vs closed-source models** represent arguably the biggest decision point when evaluating models against each other.

Proprietary models like GPT 4 (OpenAI) and Claude 2 (Anthropic) operate as a complete black box for consumers. Users input data and receive output without any visibility into the inner workings of the model. This lack of transparency raises significant concerns regarding the applicability of these models in lending contexts, as any assessment of model results would need to be fully post-hoc.

It's worth clarifying that the term "open-source" is often used loosely in this context. As previously noted, none of these models are fully open-source; rather, they are more accurately described as "open-weight." In the context of LLMs, weights represent the numerical values associated with the connections between neurons within the model network. Each connection between neurons is assigned a weight, indicating the strength and direction of influence one neuron exerts on another. While the weights of the models are openly available, the underlying algorithm and training data remain proprietary. An example of a model with open-weight architecture is the Llama model from Meta, which disclosed a significant amount about its model weights. Several noteworthy AI models and their accessibility are detailed in ***Figure 33*** below.

Open Source (Open Weight)	Closed Source (Proprietary)
Llama 70B	ChatGPT (GPT 3.5). GPT 4
Mixtral	Claude 2.0
Flacon 180B	Gopher

***Figure 33**: Noteworthy AI models classified by open vs closed source.*

Historically, closed-source models have outperformed open-source models by significant margins; however, the release of the Llama models in 2023 marked a pivotal moment, as it was the first instance where open-source (open-weight) models demonstrated comparable or superior performance to closed-source models.

The rise in sophistication of open-source models is significant as it may enable lenders to adapt and fine-tune the weights of existing open-source models for internal use, thereby enhancing flexibility in addressing specific lending challenges. Open-weight models offer the advantage of being hostable on-premises, allowing consumers to isolate them from the internet and ensuring that the companies running the black-box model are not being granted access to data that is legally protected or sensitive for the lender in question. This capability proves valuable for use cases where data confidentiality is paramount, a scenario prevalent in virtually every lending-centric application.

While training the models initially demands significant computational resources, the process of fine-tuning is considerably more cost-effective. Consequently, this accessibility has catalyzed a proliferation of derived models tailored for specific use cases, exemplified by creations such as vicuna, koala, and LawGPT. Our expectation is that most if not all lending use cases will be in the fine-tuning or Retreival-Augmented Generation (RAG) space rather than training entirely new foundational models; most lenders do not have the specific technological infrastructure required without massive tech investment, which eats into the possible gains these tools could provide in customer servicing and profiling.

For example, imagine a lender wants to use a Generative AI model for a collections agent copilot. The lender can use an off-the-shelf proprietary model that is quick to deploy and may do reasonably well, but any off-the-shelf AI model is trained on generic data and may not be as impactful. The lender can improve the model performance by specializing it for a particular use-case and through exposing it to its own proprietary data, like their own transcripts of exceptionally good internal collections agent calls. There are multiple approaches for specialization, with RAG and fine-tuning being the top two options.

With RAG atop the framework model, the lender gives the model access to query from an external database of good historic calls. In this case, the model object and model weights remain unchanged, but are given extra access to an external knowledge database to augment the model's capabilities. The other popular option is fine-tuning, where most of the model weights remain static, but some weights are retrained based on the lender's proprietary data, specializing the model for a particular use case. We have observed that a smaller fine-tuned model can often outperform a larger off-the-shelf model for specialized use cases. This makes it particularly useful in use cases like lending where low latency is an important constraint, as smaller models can be significantly faster to run.

For more on LLMs, we encourage readers to dig in to this excellent survey of large language models, last updated late last year.

II. Automated Decisioning in Lending & AI

We've received questions from big banks trying to figure out how Generative AI and LLMs might make their way into the core lending lifecycle, especially in their core profitability vectors. To that end, many have asked us to help them assess **how new AI methods might impact automated decisioning in consumer lending.**

We believe that automated decisioning processes, prevalent at the application-stage of consumer lending, are unlikely to be significantly impacted by AI. The current best-in-class automated decisioning can be summarized as follows:

1. A consumer submits an application for a lending product such as a credit card, personal loan, or auto loan.
2. The lender gathers data from the consumer reporting bureaus to acquire basic applicant information.
    - Leading lenders enhance this data collection by integrating internal performance data seamlessly via an internal API
3. Utilizing internally developed models, the lender analyzes the gathered data to generate both a credit risk score and a fraud likeliness score.
    - The credit risk score evaluates the likelihood of adverse outcomes like default, with occasional sub-models addressing nuances such as short-term versus long-term default. In support of #7, elements of the risk score aligning to the consumer profile's riskiest attributes are saved in order to support potential turndown reasons.
    - The fraud likeliness score evaluates the likelihood that an applicant is fraudulent – if the score exceeds a threshold, the applicant is then subject to manual fraud processing and

additional process steps, often including identity verification checks which they must complete before proceeding in the application process.

4. The bank utilizes its internal credit risk scores along with supplementary applicant data to classify applicants into homogenous cohorts sharing similar behaviors and characteristics.
5. The bank then predicts the performance of each cohort through the use of a valuation forecasting tool that generates an approve/decline decision. Decisioning tends to be more accounting-based than model-based, leveraging actual cardholder behavior and testing data over the prior 5 years to predict financial impact of both decisions.
6. The process branches and terminates here, per approve/decline:
    - In the event of an approval, the valuation forecasting tool generates a projection of the optimal credit line to give to the applicant. The applicant is offered the product and issued their credit if the terms are agreeable to them.
    - In the event of a decline, banks will use tools to ascertain the reasons behind the decision, enabling them to categorize turndown reasons and promptly communicate them to the rejected applicant.

As previously stated, we maintain that most of the broader application and approval process detailed above is unlikely to undergo significant disruption from AI, as defined in the background section. Nevertheless, we do anticipate the potential for one or two steps in the approval automation process to evolve in the medium-term due to AI advancements.

We do not anticipate that steps 5 and 6 of the approval automation process (predicting performance using a valuation forecasting tool and calculating optimal credit lines) will be impacted by AI advancements. Currently, the modern AI landscape is centered around language modeling, systems capable of generating customized responses to complex questions. Language models struggle with mathematical and accounting processes – their engines do not have the operational and axiomatic semantics to succeed in the reasoning necessary to do reliable math. Because steps 5 and 6 of the approval process heavily rely on precise mathematical staging for accurate outcomes, it is difficult for modern AI models to outperform the current best-in-class methods in this regard. While we have seen coding copilots speed up the build process for these models and augment the capabilities of lending analysts, the technology is simply not ready for churning out fully automated end-to-end modeling solutions – you still need a human-in-the-loop.

Furthermore, we don't anticipate that GenAI models will disrupt the adverse action stage of the current approval process (step 7 outlined above). The Equal Credit Opportunity Act (ECOA) mandates banks to inform applicants of the reasons for denial, ensuring transparency in the credit underwriting process and guarding against any potential credit discrimination. Banks have honed effective methods for staging ML models to facilitate adverse action notifications. AI models, however, are notorious for their inability to provide comprehensive explanations for their decisions, posing a challenge to compliance with ECOA requirements. Even if lenders could fully rely on the underlying mathematical principles of an AI system,

it remains difficult to imagine that they would employ it for tasks necessitating adverse action notifications. [1]

There are two stages in the automated decisioning process where it is feasible for a bank to experiment with AI usage. The first instance arises within the sub-bullet of step 2, where the lender augments bureau data with internal applicant information. Should AI models prove capable of summarizing additional data effectively, some lenders might be inclined to experiment with "persona" based variables that use AI to summarize and categorize applicant information. As previously noted, it is difficult to precisely pinpoint why an AI model makes the decisions it does – this complexity complicates its role as a customer segmentation tool, especially in cases involving interaction with automated decisioning. There are nonetheless some applied decisions outside the realm of adverse action notifications where such persona variables could be useful; for instance, in decisioning around things like proactive credit line increase programs and marketing targeting models. Despite this, in discussions with lending executives, we've observed considerable reluctance towards this concept in the immediate future, and for valid reasons. Concerns regarding fair lending practices outweigh most experimentation efforts, as few banks wish to risk violating UDAP guidelines.

Although Gen AI may not be as useful in most of the automated lending workflows mentioned above, it can still be leveraged for manual underwriting workflows. Gen AI tools excel at sifting through large reams of multi-type documents and retrieving relevant information for human consumption. This workflow can act as a copilot for a manual underwriter, like with small business underwriting, where a significant amount of document verification and reviewing is involved to approve the loan. Tools powered by Gen AI can prefill the information with citations based on lender-defined criteria. As these models can hallucinate and are not completely accurate, the human underwriter still must make the final call and review the work of the Gen AI tool, but these tools can provide efficiency gains by helping take the first step in aggregating and populating the necessary data for a decision.

The second area in automating decisioning where AI may disrupt current protocol is in step 3, where banks ascertain the likelihood of an account being fraudulent. AI is poised to potentially disrupt existing fraud detection practices more than other automation aspects, given that fraud models prioritize customer protection, operate somewhat independently of adverse action notification requirements, and have unique unstructured elements unlike underwriting models. It also is (traditionally) the area where U.S. lenders are most keen on novel analytic solutions; many of the model typologies that sit at the bedrock of Generative AI methods (i.e., neural networks) are already used by best-in-class lenders in the fraud space, albeit not pre-trained on the same volume of language data. LLMs could prove to be useful tools to help process and assess sentiment and same-user traits through analysis of the language tokens users are sending to their in-app chatbot, and the foundational architecture of LLMs could be used to expand customer device profiling information over swaths of time to help detect new use patterns.

We've engaged in conversations with several banks that appear to be exploring this avenue. Gen AI models are great at extracting structured data from unstructured data, and there are usually significant

[1] There is a small nuance here – it is possible that Generative AI models could be used to ingest, process, and describe the total picture of adverse actions or model audit actions being offered over time, in order to give compliance teams additional automated pattern recognition support to spot issues and emergent insights in a lender's turndown reasons over a variable time window. However, we would not consider this a part of the automated decisioning process, as this would be an overlay support system for monitoring and compliance teams and not a core foundational lever to the decision itself.

swaths of unstructured data in the fraud use cases. In a recent study, Mastercard claimed Generative AI helped in boosting their fraud detection rates by 20% and reduced the false positives by 85%. [2]

III. Transaction Servicing

Will AI models find a role in transaction servicing in the future?

For straightforward servicing tasks, such as routine point-of-sale interactions, we don't anticipate direct utilization of LLM-type models by lenders in the near-term. However, this assessment only offers part of the picture – tools for transaction fraud detection and anti-money laundering (AML) could potentially leverage AI, albeit in a gradual manner, based on our research findings and conversations with issuers.

As discussed in our response to whether AI will disrupt application automation, we believe there's a small opportunity to leverage AI in fraud detection and AML analysis. Nevertheless, the application of AI in this domain is expected to be nuanced. Most lenders are reluctant to introduce significant customer friction unless there is a high probability of fraud, considering fraud is inherently a numbers game. We would deem current solutions as roughly at par with existing solutions, with hallucinations in modern AI due to data volume representing a major stumbling block, and the runtime of modern AI models representing a potential slowdown of a lender's transaction servicing arm.

Ergo, before lenders fully embrace AI in transaction-level fraud detection, we believe there will need to be a significant increase in the effectiveness of fraud detection in both transactional activities and card takeover scenarios. Consequently, we anticipate an approach to AI adoption in this context that is nuanced and focused on building out data insights and customer behavior explication, in contrast to a simplistic "adopt some sort of generative framework and run all fraud through it" structure, barring a massive step forward in the efficacy of these models.

In most cases, effective fraud models focus on predicting the probability of fraud and take action only when this likelihood exceeds a certain threshold. ML models in this space that direct predicting fraud state tend to be more powerful than amalgam models producing unstable LLM responses. However, leveraging LLM tools to generate persona-like variables that both capture patterns of customer activity and identify clear breakage in pattern in a way that defeats human processing could enhance fraud detection capabilities. In areas like BIN fraud and account takeover, this could represent a strong vector for LLM enhancement of relatively nascent current solutions. Many lenders use a much wider variety of information and techniques for fraud models; things like user IP addresses, collated merchant information, outside VISA/MC swipe information, etc. LLM comprehension is more likely to find new patterns when exposed to widely varied tableaus of data than the relatively restrictive list of usable features for approval models, meaning that despite the risks, LLMs are more likely to identify patterns that are not easily human readable in this space, giving it a small opportunity to help build features and highlight concerning trends that can be integrated within a bank's fraud analytic process. While we doubt that this will be immediately applied to servicing, there's a far clearer path to that application than underwriting or risk.

In sum, we anticipate a gradual integration of LLMs and AI structures in transaction servicing. As noted in **Error! Reference source not found.**, there are a host of difficulties in platform and data staging that will

[2] https://www.mastercard.com/news/press/2024/february/mastercard-supercharges-consumer-protection-with-gen-ai/

hinder quick adoption. Nonetheless, we foresee increased experimentation by banks in 2024, particularly in low-risk areas such as enhancing push notification strategies, with LLM-backed variables empowering those tests. Recognizing fraudulent activity patterns and isolating card takeover events through innovative data analysis is crucial for minimizing losses, driving leading banks try to embrace cutting edge approaches when possible. Furthermore, our observations indicate that in most lending institutions, new modeling techniques are typically first embraced by fraud teams before they are adopted by risk modeling teams.Given the inherent difficulties in applying Generative AI in this space, we see a higher likelihood of AI being applied in a "copilot" framework.

Notes, Boilerplate, and Extra Reading

As always, 2nd Order Solutions cannot share the exact lenders, institutions, or executives we interviewed while building these answers. However, we can share a variety of public sources we've found to be useful as we build and augment our own internal responsible AI practice.

Here are a handful of useful sources that are at most one quarter old:

- **Proposed AI Regulation Standards (via the European Union)**
  - This is the EU's proposed AI regulations from earlier this month. It leaked a few weeks ago. We have some general thoughts about it; in general, all banking applications would be considered high risk, and the disclosure thresholds feel a bit high from the U.S. perspective. We think the time-table staggering is a good idea, and are a huge fan of the "regulatory sandbox" concept (although skeptical that many countries will have the funding to establish these at scale).
  - There is a high likelihood that the data fidelity requirements will make all of the current LLM pre-made systems obsolute, and introduce significant downgrades to the language processing capacity of future models (with the trade-off being significantly better controls on what the model is exposed to and the propensity for the model to make discriminatory or offensive statements.)
  - https://www.linkedin.com/posts/dr-laura-caroli-0a96a8a_ai-act-consolidated-version-activity-7155181240751374336-B3Ym/
- **Proposed AI Model Governance (via AI Verify)**
  - This is a proposed framework for AI governance provided by AI Verify, a foundation in Singapore. We have found this a pretty useful proposal to help with framing what governance for these models should look like – unlike many, this doesn't elide major risks, and the "labeling" disclosure approach is a novel way to address some of the issues raised in the draft proposal from the EU on AI governance.
  - https://aiverifyfoundation.sg/downloads/Proposed_MGF_Gen_AI_2024.pdf
- **Adversarial ML versus AI (via NIST)**
  - We found this NIST research very helpful in adding more color to a threat we've been monitoring for the last few months – that is, adversarial model training, where language models (due to their nature of being trained on user responses) are at risk of re-training

from bad actors with specific expertise. This opens up a brand new avenue for intelligent fraudsters, in the case that a bank uses LLMs in a customer-facing way.
    - https://www.nist.gov/news-events/news/2024/01/nist-identifies-types-cyberattacks-manipulate-behavior-ai-systems
- **Live LLM App Stack (via a16z)**
    - Due to the rapidly changing nature of the LLM tech stack, a16z has created a very useful live list of architecture being used by the most recent LLM applications. There is so much volatility in this space that resources like this become essential to ensure alignment with best-in-class solutions; this also helps demonstrate how quickly the current landscape changes, as weekly-to-monthly checks will reveal entirely scrapped or novel solutions for old problems and new architecture.
    - https://github.com/a16z-infra/llm-app-stack
- **MLOps Live #25: GenAI in Financial Services (via McKinsey)**
    - Several of our analysts attended this session; it provides an interesting window into some of the challenges and difficulties lenders have had and will continue to have in applying these techniques.
    - https://www.youtube.com/watch?v=vwfKPw4zv2A
- Machine Learning or Generative AI: What's better for Fraud Prevention
    - https://www.sardine.ai/blog/machine-learning-vs-generative-ai

Acknowledgements

This report was prepared by Aaron McGuire and Syed Raza, with assistance from Perry Keatley

- **Aaron McGuire**
- **Syed Raza**
- **Perry Keatley**

You may contact the authors by email at:

- aaron.mcguire@2os.com
- syed.raza@2os.com
- perry.keatley@2os.com

About 2OS

2nd Order Solutions (2OS) is a boutique credit risk advisory firm that specializes in solving the world's most challenging credit problems. 2OS was founded 12 years ago and consults to a wide range of banks, card issuers, fintechs, and specialty finance companies in the US and abroad.

Acknowledgements

2OS has deep experience with lending businesses across Card, Auto, Small Business, and Personal Loans, at all points in the credit lifecycle. 2OS partners have vast expertise in all aspects of Collections, both as operating executives and as consultants.