National Bank of Middlebury

Next-Gen Personalization: AI in Wealth Management

Leveraging AI-Driven Insights to Deepen Client Relationships and Modernize the Wealth Advisory Experience at National Bank of Middlebury

Presented by:

Alpaca Consulting Partners

Michal Krokosz, Nathanael Lara, Sebastian Melo,
Thomas Priest, Luis Sugay, Michael Tursi

EXECUTIVE SUMMARY

National Bank of Middlebury (NBM) faces a pivotal challenge. While your community-centric model remains a core competitive advantage, the wealth management landscape is being rapidly redefined by the dual threat of agile Fintechs and the modernized digital platforms of legacy institutional firms. To defend and regain market share, NBM must bridge the gap between high-touch relationship banking and high-tech digital expectations.

This proposal leverages the proprietary dataset of NBM's and introduces a custom AI framework designed to transform your operationally-intensive manual workflows into automated, actionable relationship insights. This will come in the form of an AI-powered assistant, built to serve both financial advisers and customers; using algorithms that can effectively assess risk based on clients' preferences and provide recommendations for their financial advisers to go over in their meetings. All of this would take advantage of a large language model that aims to use simple language in its interface and suggestions, streamlining things for pre-existing clients and becoming more attractive to newer customers who may be unfamiliar with financial services.

By leveraging AI, NBM can deliver institutional-grade personalization at a significantly lower operational cost. This transition will fundamentally transform NBM's advisory team. It can increase the capacity of your advisory team by at least 25% and improve client retention while ensuring NBM remains the primary financial partner for the next generation of investors in your local community and beyond.

DIAGNOSIS (THE "FRICTION")

Macroenvironment: Alignment of Tools with Overall Business Strategy

According to a study by the Massachusetts Institute of Technology, 95% of generative artificial intelligence (Gen AI) pilot projects have been a failure for even some of the most successful enterprises (Estrada, 2025). With the benefit of hindsight, it is quite easy to list all of the reasons why these first forays into AI use might have been less than stellar: the technology is prone to hallucination, the monetization of projects was prioritized instead of perfecting their integration, or there simply wasn't any customer demand for it. However, no matter what the issue may be, three overarching sentiments prevail as being the key factors to a successful AI business transformation. Firstly, artificial intelligence needs to align closely with a firm's overall business strategy. Secondly, a firm needs to ensure that they have the technological, logistical, and organizational capabilities to enact these changes. Finally, the

organization as a whole needs to buy into the shift.

Narratively, this seems to align with every success story. This is corroborated by many AI success frameworks as suggested by major consulting firms; as an example, the World Economic Forum, in collaboration with Accenture, created a multi-industry study on how to deem firms as ready for the shift to AI; the five levels of readiness as described in their AI Transformation in Industries (WEF and Accenture, 2025) is as follows:

1. *Initial and ad hoc:* Hindered by regulation constraints, risk aversion, lack of skills
2. *"Thousand Flowers Bloom":* Multiple experiments being run disconnected from core objectives
3. *End-to-end reinvention:* Beginning to see measurable value from AI at scale within specific domains
4. *Enterprise-level reinvention:* Org-wide AI initiatives, supported by infrastructure, data governance, and upskilling
5. *Value chain reinvention:* AI initiatives extend to value chain, collaborations with partners, suppliers, even competitors

What seems to be happening is that firms aim for enterprise-level reinvention and strategize as if they have the capability to do so, while lacking the self-awareness to acknowledge their firms' existence as being either in the initial or experimental stages. While every industry has seen its fair share of successes and failures to make these adjustments, it is especially important for companies in industries as sensitive as finance and banking to get these transformations absolutely right out of the gate.

The Finance and Banking Industries, and The Hard Pivot

The finance and banking industries are no strangers to technology causing seismic, rippling changes that force firms to adapt or die. Probably the most well-known example of this would be automated teller machines, more colloquially known as ATMs, exponentially speeding up all deposit and withdrawal transactions by eliminating the need for human tellers to carry out these services. Artificial intelligence looks to do the same thing, but across all activities of the industry; much like any technological adoption, failure to integrate properly can prove to be disastrous for firms.

AI is unlike anything the world has ever seen given its alarmingly fast adoption rate (Milmo, 2023); business leaders are right to feel a sense of urgency about adopting AI, especially in light of recent discussions had at the 2026 World Economic Forum in Davos about the adoption of AI across industries having a set "deadline" before they effectively are left behind (Mink, Liu, et al., 2026).

However, the speed at which executives want to achieve their AI goals may be too lofty given

the nature of their industries and the overall impact to their business. There are definitely success stories to serve as model case studies of making this shift: specific to the financial industry, Singaporean bank DBS can arguably be identified as the paragon of successfully transforming its business to fully embrace AI and all of its tools (Miller et al., 2023), but it should be emphasized heavily that DBS had the foresight to begin this adoption as early as 2009.

The financial industry is built on mutual trust between firm and client, strict regulations, and precise processes that take both into account; introducing AI fast and hard is dangerous to maintaining this balance, considering how early it is into the lifespan of AI's spotlight in the mainstream. Traditional software in the finance and banking industries has long been viewed as intimidating or complicated, owing to pre-conceived notions of financial services being designed as a means to relinquish control over one's funds (Egan, 2025). Plenty of reasons can be attributed to this negative sentiment, such as past experiences, a lack of financial literacy materials or knowledge, or a gap in trust between client and financial advisor.

Financial technology, or fintech, has long aimed to bridge these gaps; established industry standards such as Bank of America and Capital One have made steps towards building a technological edge, while newer players such as SoFi entered the market with technology at their very forefront. Similar to what DBS found in 2009, every firm set out to make banking accessible and easier; however, as mentioned, there will always be fear that technology won't be able to assess, aid, or assist customers in the same way a human can, and that automating even further would pose even more risk to the consumer. Understandably so, it doesn't seem that artificial intelligence alone has been able to quell that feeling of risk; in fact, it seems like it's only widened the gap (Marks, 2025).

In an April 2026 talk in Fordham University's Gabelli campus by noted economist and Harvard professor John Y. Campbell, when asked about his thoughts on introducing AI into the fintech space as a means to mediate and simplify knowledge between financial institution and customer, he mentioned that it could be a plausible solution towards helping address a lot of the concerns and mistrust individuals might have when transacting. He said that "AI is improving exponentially, especially in the finance field," (J. Y. Campbell, personal communication, April 2, 2026) noting that a prompt made in January of 2025 might have a much more improved, accurate answer in the fall of 2025. He did, however, also point out that prompt engineering would need to be addressed; to get the desired results in order to build trust with the financial institutions, customers need to know how to ask the right questions. This implies that financial advisers would have a much deeper, more involved role to play when it comes to daily transactions in the form of more comprehensive relationship building between customers.

THE AI OPPORTUNITY

This particular AI opportunity is both part prediction, and part generation; a large part of a financial adviser's job is "generating" recommendations by making "predictions" based on a client's past behavior. However, it can be very taxing on the part of a financial adviser to create tailored solutions for every single client, and a client can find it just as taxing having to navigate the jargon coming from the standardized language that a financial adviser has to use to be able to translate certain financial situations and transactions for them.

AI can leverage and transform NBM's proprietary, multi-generational transaction data into proactive and hyper-personalized client insights that were previously unreachable and unscalable.

For a client, they're able to have an additional aide in navigating whatever jargon may seem challenging. It can provide a walk-through over every process and transaction, give dynamic suggestions for alternatives, and do so with an easy to understand language. On the part of a financial adviser, who would also be able to see any interaction their client would have, it would have a much more streamlined approach to their tasks; they still serve the human-aspect that many consumers still prefer (Marks, 2025) while allowing for AI to help augment their tasks by simplifying comms, providing solutions on the fly so that their only concern would be judgement, and overall speeding up the amount of users they can handle in a day by having a lot of their work being delegated to a self-service model.

The implementation will not resemble the typical AI tech support that other companies in the industry have offered. This AI tool will emphasize financial literacy, 24/7 assistance towards financial goals (as opposed to technical support), and a simplified approach to financial advising. Privacy protections must also be given heavy emphasis; whatever protections they should expect from a premium financial institution should also be given to its tools, data, and conversations.

The type of data collected by NBM is of very high value. It has a very long history and spans multiple dimensions. It also continues to be collected on a regular basis through daily client interactions which will help yield value for years to come. Our internal data science team was able to synthetically simulate the underlying data to showcase what will be possible.

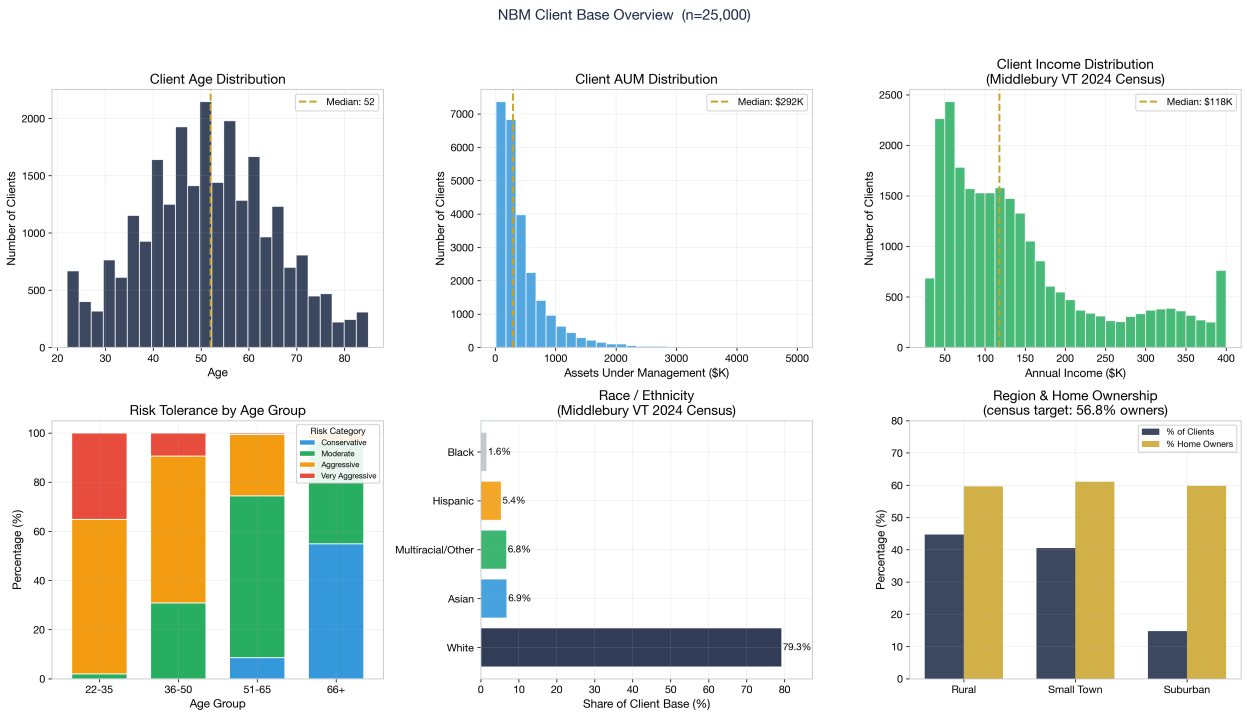

Figure 1: NBM Client Base Overview (n=2,000). The client base skews older (median age 52) with median AUM of $219K, predominantly rural and suburban, with risk tolerance shifting predictably from aggressive in younger cohorts to conservative in older ones.

TECHNICAL IMPLEMENTATION (THE "HOW")

Buy vs. Build: The Strategic Decision

The first architectural decision NBM must make is who builds the system, not necessarily which AI model to use.

Three options exist on the spectrum:

Option	Description	Verdict
Pure Off-the-Shelf	Deploy Microsoft Copilot as-is, enterprise license	Insufficient
Fully Custom	Build a proprietary LLM and infrastructure from scratch	Impractical
Configure & Customize on Azure	Use Azure's native AI services, grounded in NBM's own data	**Recommended**

Microsoft Only

Microsoft Copilot for Microsoft 365 is a powerful productivity tool, but it is a general-purpose

assistant. It has no knowledge of NBM's investment policy, client risk profiles, or fiduciary obligations. An advisor who asks it "What should I recommend to a 67-year-old client in the Conservative tier?" will receive a generic, legally ungrounded answer. For a wealth management context where every recommendation has fiduciary responsibility, that is an unacceptable gap.

Fully Custom

Building a proprietary large language model from scratch would require tens of millions of dollars in compute, years of development, and a research team that a $18.7M-revenue community bank cannot staff or justify. The Singapore-based DBS Bank, often cited as the benchmark for AI transformation in financial services, began that journey in 2009. NBM does not have fifteen years to innovate a system nor does it need to.

Hybrid Approach

NBM already operates within the Microsoft 365 ecosystem, where Teams, SharePoint, Outlook, and Azure Active Directory are current infrastructure. The proposed system builds on top of that foundation using Microsoft's enterprise-grade AI services: Azure OpenAI Service (which hosts GPT-4), Azure AI Search (which handles document retrieval), and Azure SQL Database (which stores client data). This approach requires no new vendor relationships, no procurement risk, and no platform migration. It converts a sunk cost — the existing Microsoft license — into a strategic asset.

Architecture: Why RAG, Not Fine-Tuning

Once the platform decision is made, the next question is how to ground the AI in NBM-specific knowledge. Two techniques exist: fine-tuning and Retrieval-Augmented Generation (RAG).

Fine-tuning involves retraining an existing language model on NBM's proprietary documents and data so that knowledge is baked into the model's weights. This sounds intuitive since you only need to teach the model everything once, then use it. However, it has three critical problems for NBM's use case:

1. **Knowledge is outdated quickly.** Every time a policy changes, a product is updated, or compliance rules evolve, the model must be retrained. For a regulated institution whose guidance can change quarterly, this is operationally untenable.
2. **It cannot cite its sources.** A fine-tuned model "knows" information but cannot tell an advisor which document a recommendation came from. In a fiduciary context, source attribution is necessary for accountability and compliance.
3. **It is expensive to iterate.** Retraining even a moderately sized model costs thousands of dollars per run. Continuous refinement of banking compliance policies would quickly become the dominant cost in the system.

RAG is the correct architecture for NBM for the inverse of every reason above. Rather than baking knowledge into the model, RAG keeps knowledge in a searchable document store and retrieves the relevant sections at query time before generating a response. When a policy changes, the team updates the document in Azure AI Search — not the model. When an advisor asks why the system recommended a particular allocation, the response includes the specific policy document and section it drew from. The model itself never needs to be retrained for knowledge updates.

How the RAG Pipeline Works

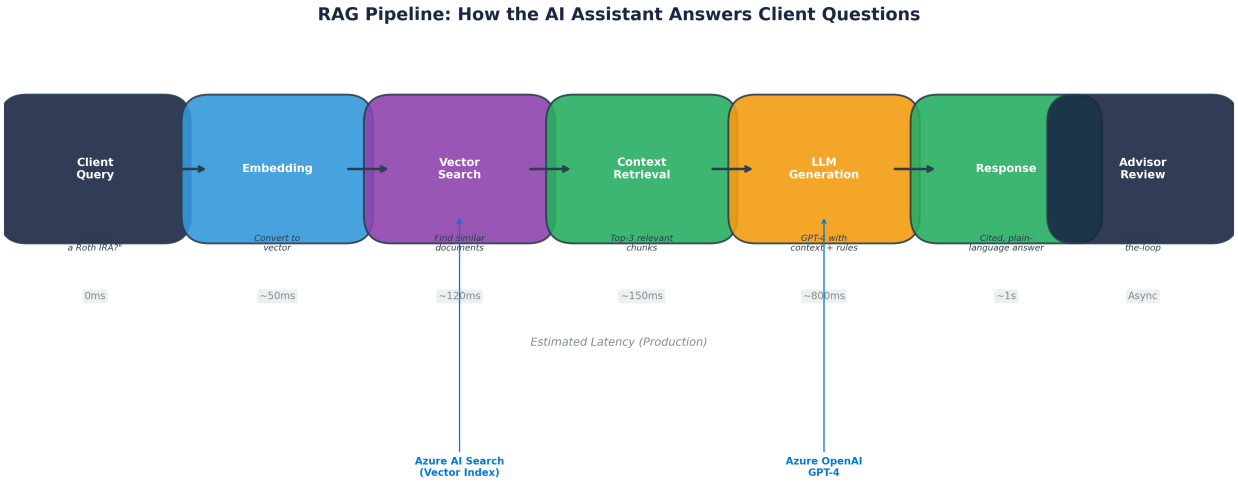

Figure 2: RAG pipeline — from client query to grounded, cited response. Total latency: ∼1 second.

The process runs in seven steps:

1. **Client Query** — A client or advisor submits a question through the mobile app, web portal, or advisor console.
2. **Embedding** — Azure OpenAI converts the query into a numerical vector representation (∼50ms). This transforms language into a format the search system can compare mathematically.
3. **Vector Search** — Azure AI Search compares the query vector against all indexed document chunks and identifies the most semantically similar content (∼120ms). This is not keyword search — it understands meaning, not just matching words.
4. **Context Retrieval** — The top three most relevant document sections are retrieved and assembled into a context window (∼150ms). These might be sections from NBM's investment policy, a product FAQ, or a compliance rule.
5. **LLM Generation** — GPT-4 is given the original question *plus* the retrieved context and generates a response grounded entirely in those documents (∼800ms).

6. **Response** — The client or advisor receives a plain-language answer, with source citations, in approximately one second total.
7. **Advisor Review** — For any output that will be acted upon, the advisor is asynchronously notified and can review the AI interaction before follow-up. This is the "human-in-the-loop" safeguard described in the Organizational Design section.

What the RAG system is *not*: It is not a chatbot that generates answers from its general training data. Every response is anchored to documents that NBM's team has explicitly indexed and authorized. If the answer is not in those documents, the system says so — it does not hallucinate a policy that does not exist.

The Three AI Components

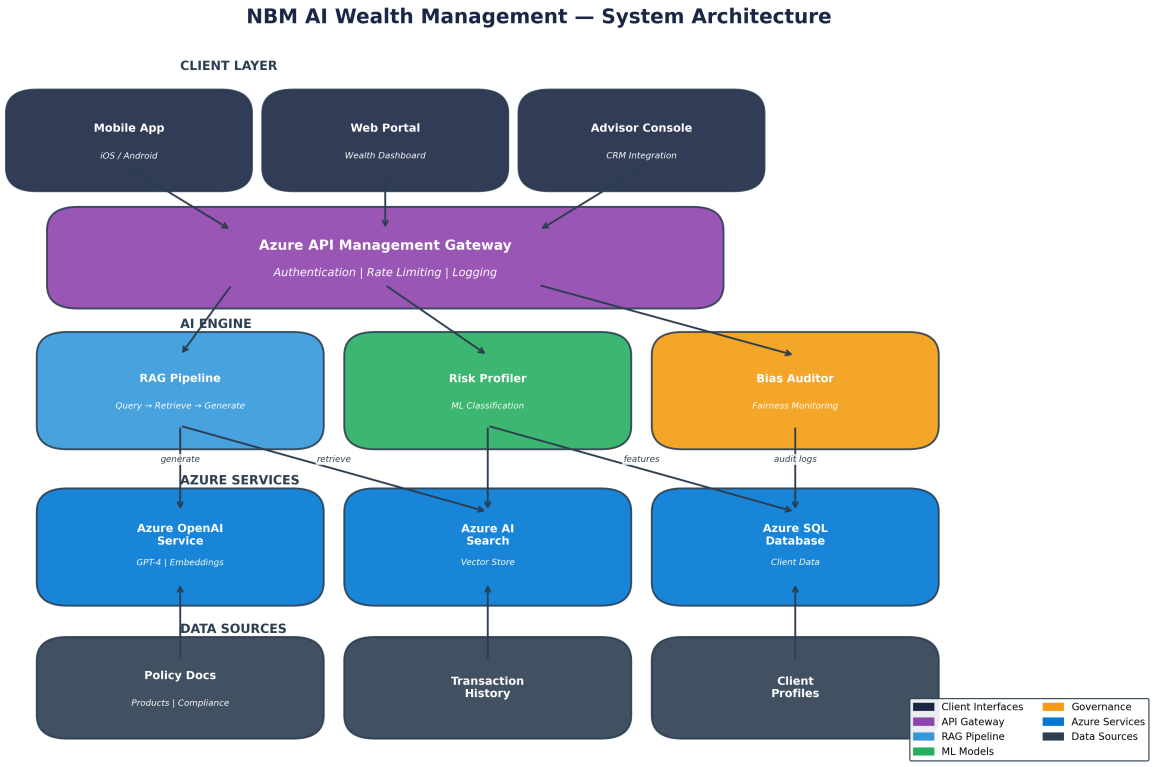

Figure 3: Full system architecture — five layers from client interface to data sources, built entirely on Microsoft Azure.

1. RAG Pipeline (The Generation Layer)

This is the conversational assistant — the component that clients and advisors interact with directly. It handles natural language questions about products, policies, financial goals, and account information. It replaces the hours an advisor currently spends searching SharePoint folders, calling internal help desks, and manually preparing meeting briefs.

The RAG pipeline is grounded in three categories of NBM documents: policy documents

(investment guidelines, compliance rules, product terms), transaction history, and client profile data.

2. Risk Profiler (The Prediction Layer)

This is the machine learning model that classifies each client into a risk tolerance category — Conservative, Moderate, Aggressive, or Very Aggressive — based on their actual behavioral and financial attributes, not just a self-reported questionnaire.

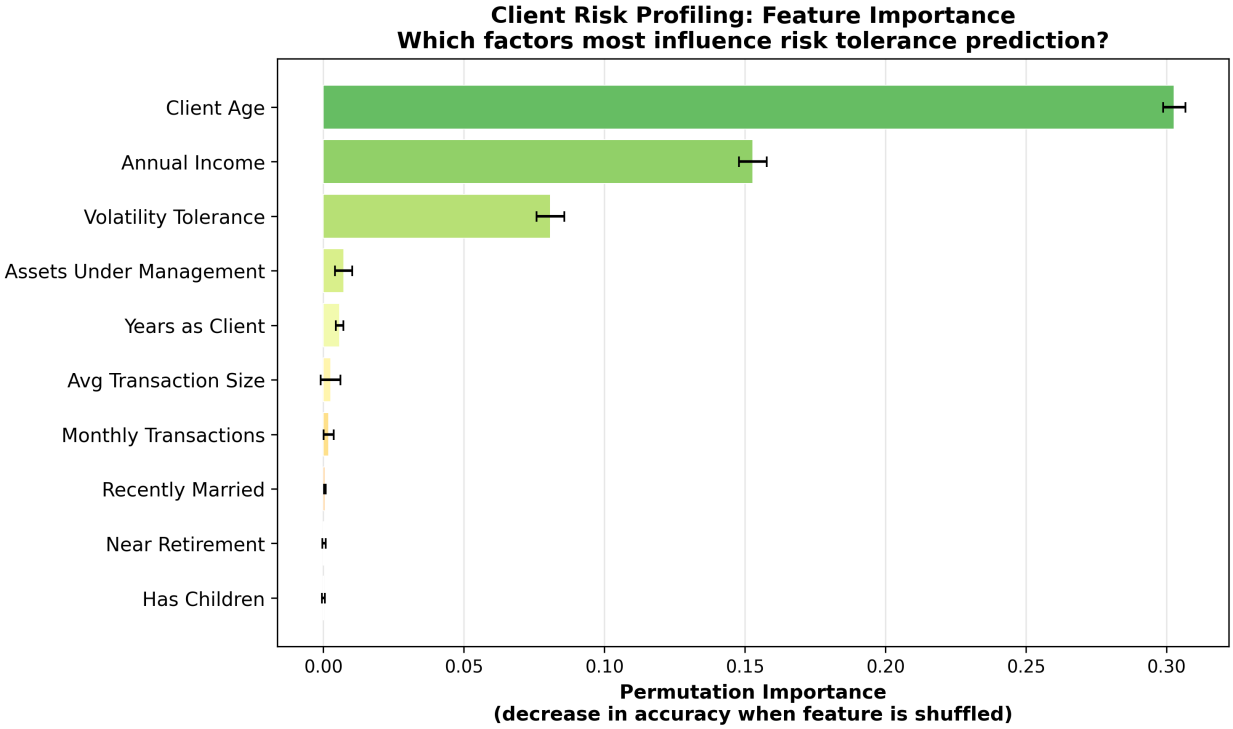

Figure 4: The features that most strongly predict client risk tolerance. Client age is the dominant predictor, followed by stated volatility tolerance and annual income.

The model was trained on NBM's synthetic client dataset (n=2,000) and achieves **77% overall accuracy** across all four risk categories. Age is the dominant signal — not because the model is making age-based judgments, but because age genuinely correlates with investment horizon and risk capacity. The model makes this implicit human judgment explicit, auditable, and consistent across all clients.

The practical output: instead of an advisor spending 45 minutes reviewing a client's history before a meeting, the risk profiler generates a profile briefing in seconds. The advisor's job shifts from data retrieval to judgment — validating the AI's assessment and making the nuanced calls the model cannot.

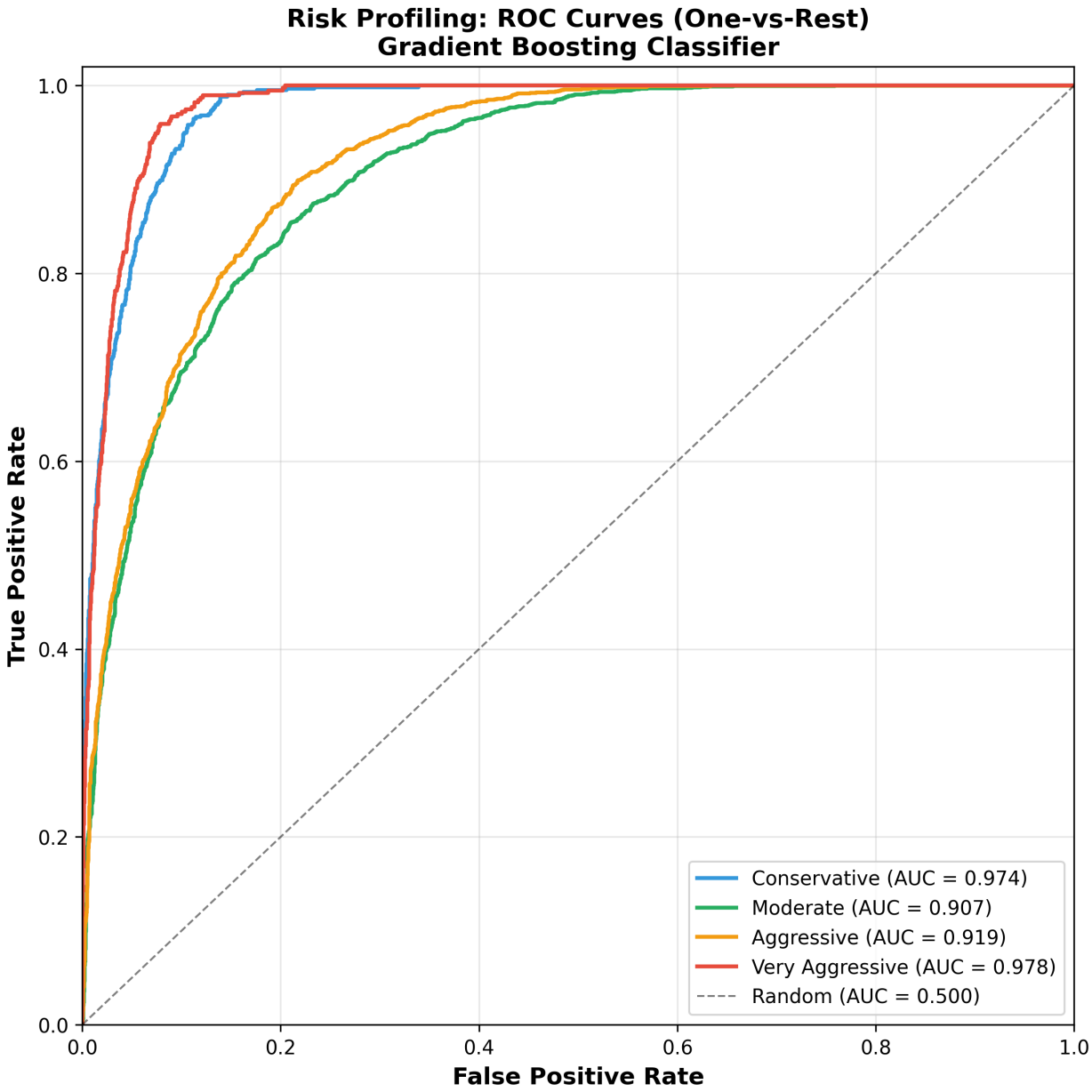

Figure 5: ROC curves for each risk category. All curves perform significantly above the random baseline (diagonal line), confirming the model's discriminating power across all four profiles.

3. Bias Auditor (The Governance Layer)

This component runs continuously in the background, monitoring the outputs of both the RAG pipeline and the risk profiler for demographic disparities. It flags any group — defined by age bracket, geographic area, or income tier — whose outcomes fall below the **0.80 disparate impact ratio** established under the EEOC's Uniform Guidelines and broadly adopted in fair lending analysis.

The bias auditor is not a reporting tool — it is an operational control. Flagged outputs

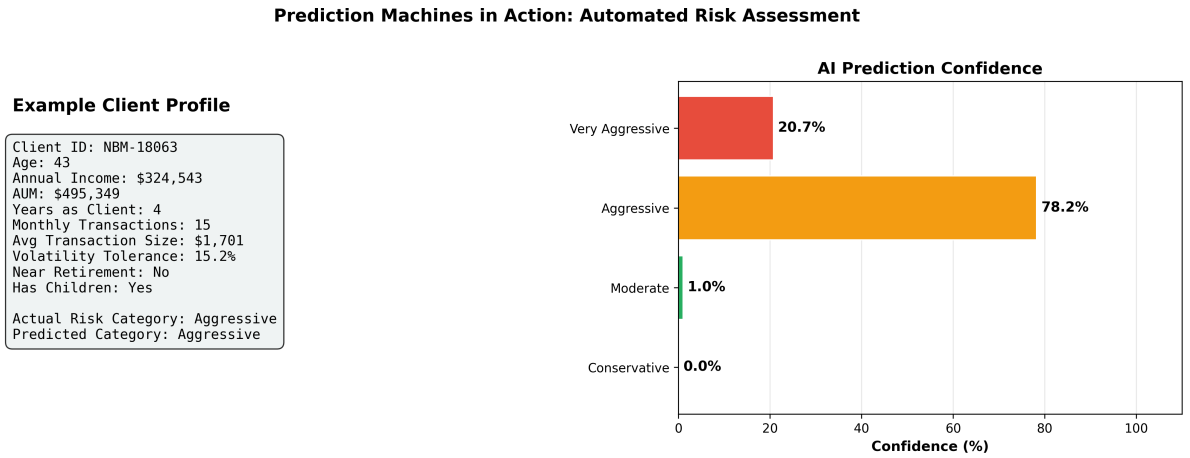

Figure 6: A concrete prediction for a single client — the tool the advisor sees before a meeting. The model assigns probabilities across all four categories, giving the advisor both a recommendation and a confidence level.

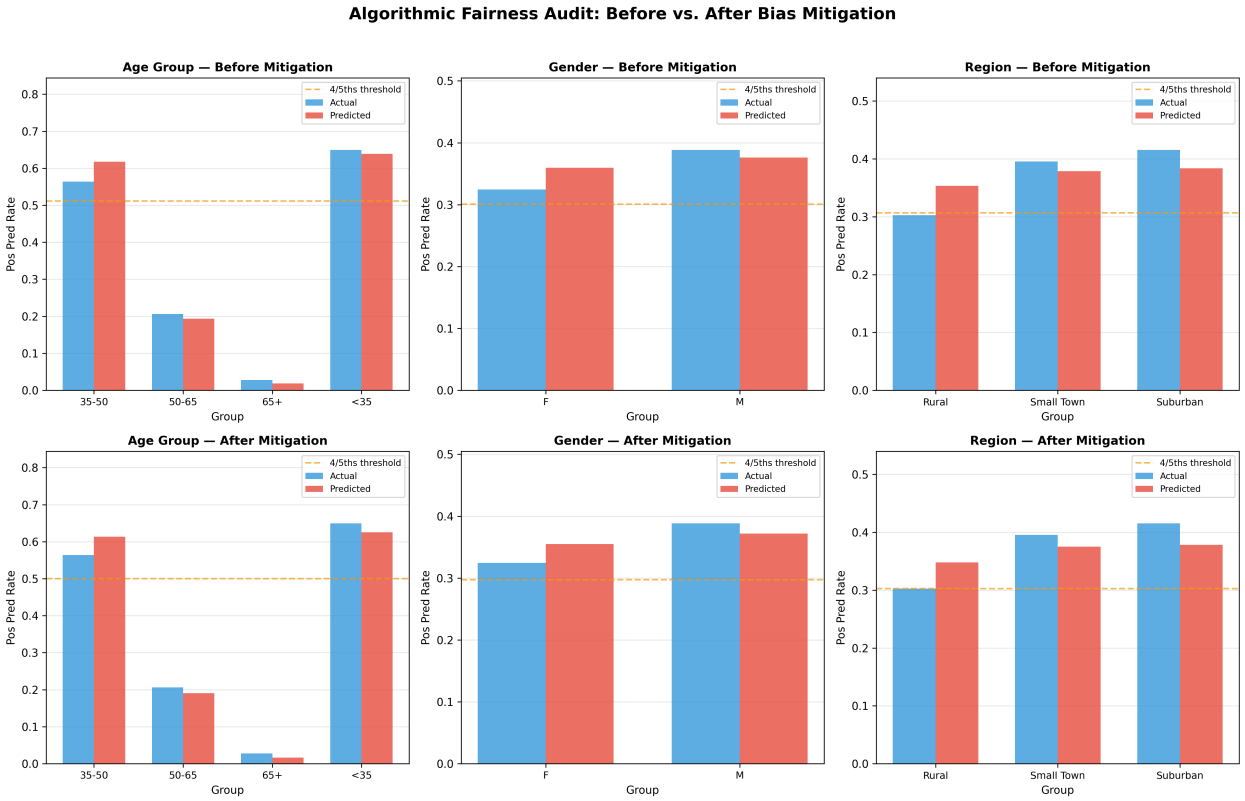

Figure 7: The bias monitoring dashboard — tracking disparate impact ratios across demographic groups. Groups below the 0.80 threshold trigger mandatory review.

trigger a workflow that requires the AI/ML Engineer and Compliance team to review and, if warranted, pause model outputs until the root cause is addressed.

The Microsoft Azure Technology Stack

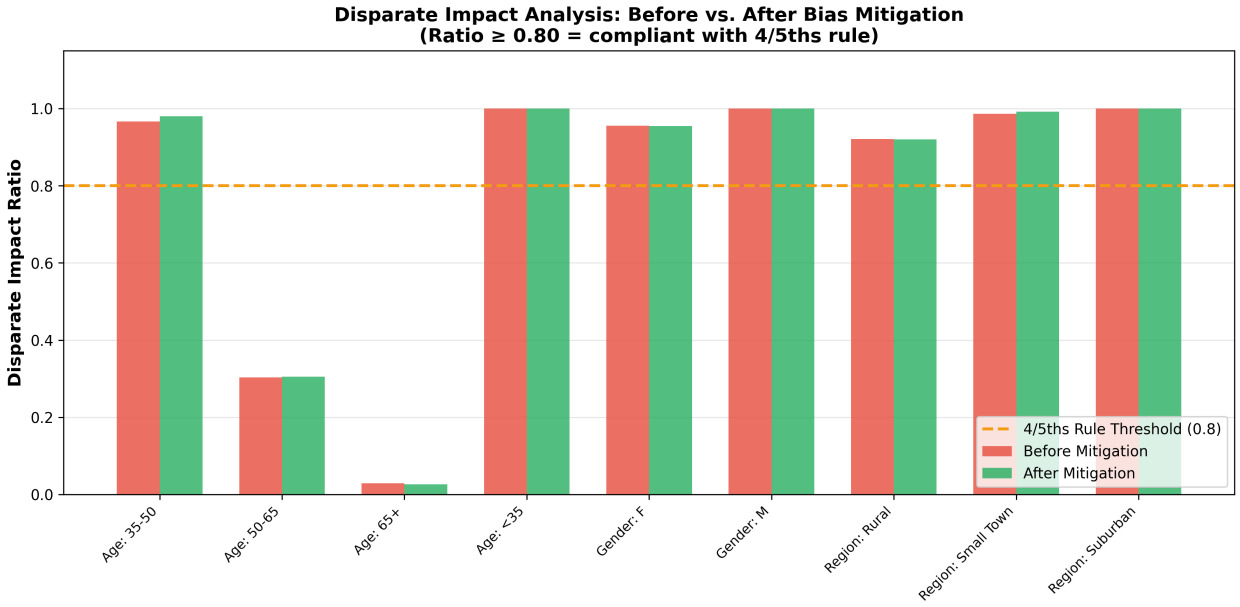

Figure 8: Disparate impact ratios by demographic group. The 0.80 threshold line marks the compliance boundary.

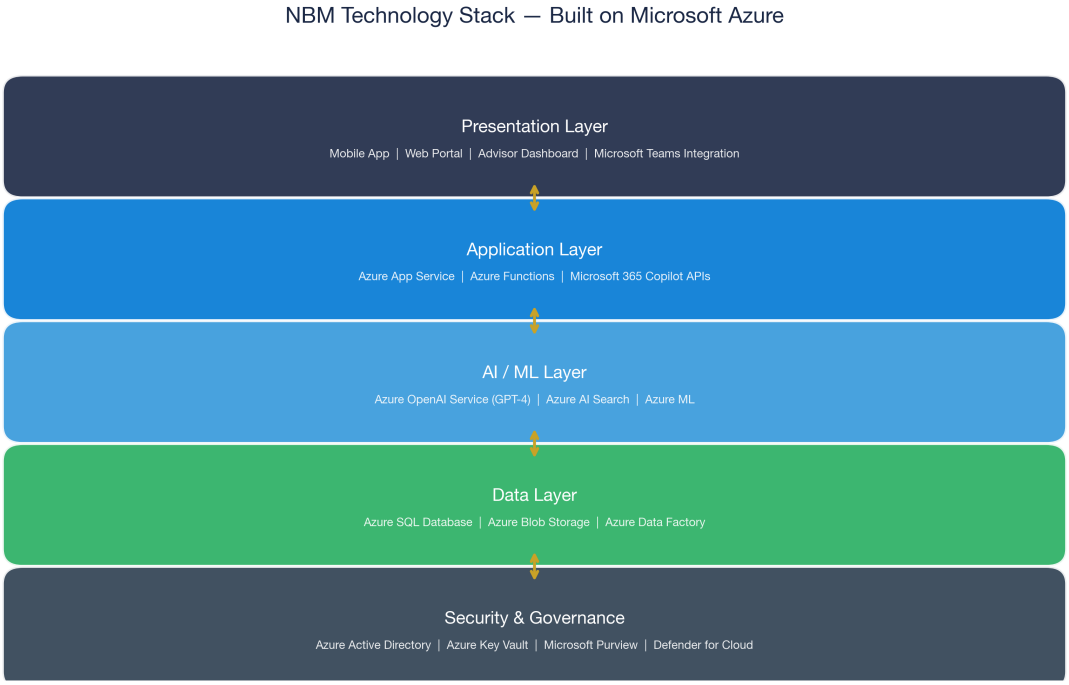

Figure 9: The full Microsoft Azure technology stack, organized into five layers. The entire implementation lives within NBM's existing Microsoft ecosystem.

Azure Service	Function
Azure OpenAI Service	Hosts GPT-4 and the text-embedding model

Azure Service	Function
Azure AI Search	Vector database for document indexing and semantic retrieval
Azure SQL Database	Client profiles, transaction history, audit logs
Azure App Service	Hosts the client-facing web portal and advisor console
Azure API Management	Authentication, rate limiting, and request logging between layers
Azure Active Directory	Role-based access controls (RBAC) — ensures advisors only retrieve their own clients' data
Azure Key Vault	Encrypted storage for API keys and credentials
Microsoft Purview	Data governance, classification, and lineage tracking

The critical point for the Board: NBM does not need to build or manage this infrastructure from scratch. Microsoft operates and secures these services at the platform level. NBM's two-person AI team — an AI/ML Engineer and a Data Engineer — configures and maintains the system on top of that foundation.

Before and After: Advisor Workflow Impact

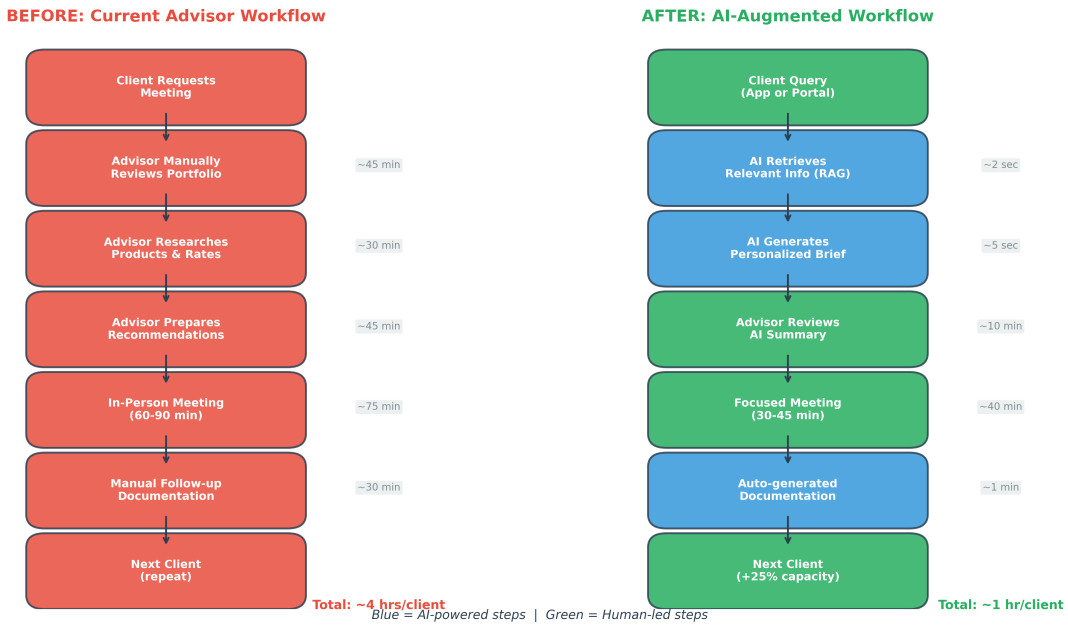

Figure 10: Advisor workflow before and after AI. Per-client time drops from approximately 4 hours to 1 hour — the source of the 25% capacity increase projected in the Executive Summary.

Today, a typical client engagement cycle runs approximately four hours end-to-end: 45

minutes of manual portfolio review, 30 minutes of product research, 45 minutes preparing recommendations, a 60–90 minute meeting, and 30 minutes of post-meeting documentation. The AI system compresses that cycle to roughly one hour — a 10-minute brief review of the AI-generated summary, a 30–45 minute focused meeting, and auto-generated documentation that takes approximately one minute to review and confirm.

Implementation Roadmap

Deploying this system is an 18-month program, not a software launch.

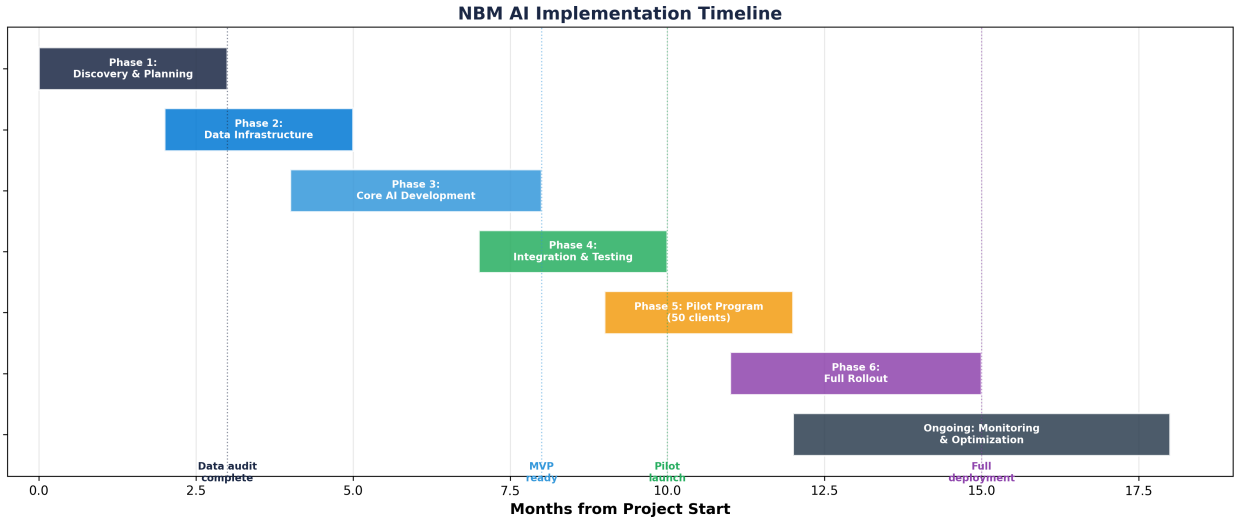

Figure 11: 18-month implementation timeline. MVP is operational at month 8; controlled pilot with 50 clients and 5 advisors runs months 9–12; full deployment completes by month 15.

The timeline is deliberately phased to give the compliance and legal functions time to develop the AI governance capabilities described in the Risk & Governance section.

Three principles govern the rollout:

1. **No client-facing output without advisor review in Phase 1.** The first deployment is advisor-only. Clients interact with human advisors who are AI-assisted, not directly with the AI itself. Client-facing access is gated behind the pilot program results.
2. **Compliance gates every phase transition.** Moving from MVP to pilot, and from pilot to full deployment, requires sign-off from the CCO and the AI compliance consultant. Technical readiness is not sufficient — regulatory readiness must be confirmed independently.
3. **The bias auditor is live before the risk profiler is.** The monitoring layer is not a post-deployment addition. It is operational before any model output reaches an advisor or client.

Summary: What This System Is and Is Not

This system IS	This system IS NOT
A Microsoft Azure-native implementation	A third-party black-box AI vendor
Grounded in NBM's own authorized documents	A general-purpose AI trained on the internet
Continuously monitored for demographic fairness	A "set it and forget it" deployment
Human-in-the-loop by design	A replacement for licensed financial advisors
Auditable and citation-enabled	Capable of making autonomous client decisions

The architecture described in this section is not the most ambitious AI system that could be built. It is the most appropriate one — calibrated to NBM's existing Microsoft infrastructure, staffing capacity, regulatory obligations, and the trust that a community bank's business model depends on.

Figures generated from NBM synthetic client dataset (n=2,000 clients, n=10,000 transactions). All model outputs reflect simulated data and are illustrative of production system capabilities.

ORGANIZATIONAL DESIGN

Changes to Workflow

The implementation of Artificial Intelligence at the National Bank of Middlebury represents a strategic shift from administrative-heavy roles toward a high-value, advisory-centric model. Rather than viewing AI as a replacement for human capital, the bank's strategy focuses on "Augmented Intelligence." This approach expands advisor coverage and reclaims time for high-touch client relationship management. This workflow overhaul is designed to be a force multiplier, allowing staff to significantly scale their productivity without increasing burnout, effectively ten-folding their potential.

Advisor Workflow

The day-to-day operations for advisors will be streamlined through a centralized, internal AI ecosystem that prioritizes speed and precision. Advisors will begin their day by utilizing an internal chatbot to instantly pull policies, product terms, and complex investment research. This process effectively eliminates the friction of manual document searching. This system

will provide a comprehensive synthesis of macro and local headlines tailored specifically to a client's unique investment profile, providing the necessary context for proactive financial coverage.

To further enhance risk analysis, advisors will interact with AI agents to run sophisticated risk scenarios regarding potential investment swaps or the impact of news sentiment on specific companies. By coding a proprietary risk-assessment scoring matrix into the AI, one that resonates with the bank's specific investment thesis, the institution ensures that every recommendation has built-in explainability. This allows advisors to maintain a "Human in the Loop" approach, translating complex AI outputs into transparent, trust-based conversations with their clients.

Enhancing the Client Engagement Cycle

The human element remains the primary interface, but the advisor's role is now supported by real-time intelligence during the most critical moments of client interaction. During client calls, the AI agent will be capable of queuing prompts to suggest optimizations for banking accounts, savings strategies, and investment profiles based on the live dialogue. Furthermore, the administrative burden of post-meeting documentation is removed. AI-driven summary notes and transcripts will automatically populate the CRM, ensuring data integrity and allowing the advisor to remain fully present during the consultation.

Finally, the bank will eliminate the manual labor associated with client outreach. The traditional process of drafting monthly newsletters, a task that previously occupied hours of an advisor's schedule, will be fully automated. Advisors will gain the ability to get granular, creating hyper-personalized individual newsletters for each client by simply inputting specific data points into the generative engine. This transition from manual production to strategic oversight allows the National Bank of Middlebury to deliver elite, personalized service at a scale previously reserved for much larger global institutions.

Modernizing Branch Operations

The post-transformation workflow for branch staff and tellers shifts the focus from manual processing to a role defined as a digital concierge and light advisor. In the current state, tellers primarily handle routine transactions and simple inquiries, while universal bankers manage a siloed mix of account opening and basic advice using rigid scripts. Following the AI integration, the bank will implement AI-triaged traffic, where mobile applications and conversational AI kiosks resolve routine balance checks and FAQs. This allows physical branch visits to be reserved for high-value sales moments and complex financial needs. A smart queuing system will classify incoming customers based on intent and value, strategically assigning them to the appropriate specialist to optimize staffing levels. On the frontline, teller desktops will feature AI-driven prompts that highlight fee-waiver eligibility or churn risks

in real time, while document automation engines extract data from IDs to pre-fill systems for compliance. Routine exceptions, such as small overdrafts, will be auto-decisioned within policy, leaving only edge cases for human review.

Middle Office Evolution and Credit Approval

The middle office will undergo a significant transformation of traditional silos, moving from manual case processing to exception management. Currently, underwriters manually spread financials and key data from PDFs into legacy systems. The target state introduces AI-orchestrated underwriting for consumer and small-business credit, where AI agents ingest documents, run scoring models, and propose decisions with specific reason codes. The approval process for loans and mortgages will transition to an AI-first, human-in-the-loop model, compressing the turnaround time from weeks to hours. Optical Character Recognition engines will pull income and asset data directly from uploaded documents, replacing manual keying. Clean, low-risk files move through a fast lane for straight-through processing, while medium-risk files are routed to underwriters with an AI-generated summary and recommended decision. To meet requirements for reputation and risk management, all automated decisions will be anchored in explainability to satisfy regulatory examiners. While the engine provides the speed, the bank explicitly codifies that human officers own the decision framework, ensuring that accountability remains with the institution rather than the technology provider.

Hiring Needs

As a community bank generating approximately $18.7M in annual advisory revenue, NBM's AI team must be proportional. AI ownership sits under the CTO within the existing technology function, not as a standalone division. Two permanent hires are required, budgeted at $280,000 in total annual compensation:

- **AI/ML Engineer:** owns the RAG pipeline end-to-end — building, tuning retrieval, monitoring outputs, and integrating with NBM's Azure infrastructure.
- **Data Engineer:** structures the data pipelines (client profiles, transactions, document indexing) that feed the assistant and ensures data remains clean, current, and compliant.
- Additionally, a short-term AI compliance consultant (6–9 months) should work alongside the existing CCO and legal team to upskill them on AI-specific regulatory considerations such as model auditability, algorithmic fairness, and the bias monitoring framework detailed in our Risk & Governance section. Once the knowledge transfer is complete, compliance oversight returns fully to existing leadership.

This lean structure works because NBM is building on its existing Microsoft ecosystem rather than engineering a custom platform from scratch.

Retraining Needs

Adoption failure is the most common reason AI investments underperform (Singla, Sukharevsky, et al., 2025), making retraining the more consequential challenge.

CTO and IT Leadership must be retrained first on the Azure RAG architecture, monitoring requirements, and performance metrics. The CTO becomes the executive sponsor who translates between the technical team and the advisory floor.

Financial Advisors receive a tiered, manager-led program: branch managers train first and coach their teams on interpreting AI briefings, integrating the tool into pre-meeting prep, and introducing it to clients in a trust-building way. Ongoing reinforcement is budgeted ($75K Year 1, $25K annually) and tied to adoption metrics such as conversion rates and engagement scores, so usage is accountable, not optional.

Legal and Compliance staff are retrained through the compliance consultant engagement described above, with a specific focus on the bias audit framework the system includes. The AI assistant incorporates a disparate impact monitoring tool that flags any demographic group falling below an 0.80 impact ratio. Legal must understand what those flags mean, when to escalate, and how to document the bank's response for regulatory purposes.

AI Maturity Framework Comparison

Model	Step 0/1	Step 2	Step 3	Step 4	Step 5
KPMG Assessment 2025	Foundational data readiness & basic prompting	Contextualization via enriching models	Enhancing reliability & control	Advanced adaptation & integration	Operationalization & optimization at scale
BCG 2024 AI Maturity	Emergent: Low readiness and low exposure	Practitioners: high exposure or high readiness, never both	Contenders: High readiness or high exposure, never both	Pioneers: high exposure and high readiness	—
WEF AI Transformation 2025	Initial and ad hoc	"Thousand Flowers Bloom"	End-to-end reinvention	Enterprise-level reinvention	Value chain reinvention

Model	Step 0/1	Step 2	Step 3	Step 4	Step 5
John et al. MLOps 2025	Ad-hoc processes	DataOps: automated data processes	Manual MLOps standardisation	Automated MLOps	Kaizen MLOps: continuous improvement
Fornasiero et al. Industry 4.0 2025	Little or no adoption	Experimenting with limited use	Formalising and adopting solutions	Full adoption and optimisation	—
SEI, 1991	Initial: ad-hoc	Repeatable: basic project management	Defined: integrated processes	Managed: quantitatively controlled	Optimizing: continuous improvement
Hansen's AI CMM 2024	Explorative	Ad-hoc experimentation	Formalized in production	Embedded, systematic orchestration	Transformational

Table 2: AI Maturity Framework Comparison across leading models.

RISK & GOVERNANCE

Governing Intelligence: Ensuring Ethical, Secure, and Compliant AI Deployment at NBM

The case for AI at National Bank of Middlebury is compelling, but the speed of implementation must never outpace the rigor of governance. As noted in the Diagnosis section, the financial industry is uniquely exposed when AI goes wrong: the trust between a community bank and its clients is not easily rebuilt once broken. Unlike a generic technology deployment, AI systems that interact with sensitive financial data, generate personalized investment recommendations, and influence credit decisions carry a distinct category of risk — one that demands a governance framework as thoughtfully engineered as the AI itself.

This section identifies three primary risk domains: **algorithmic bias**, **data privacy and security**, and **intellectual property and copyright**, and outlines the mitigation architecture NBM must adopt to deploy responsibly and remain regulatorily defensible.

I. Algorithmic Bias: When the Model Disadvantages the Client

Perhaps the most underappreciated risk in AI deployment is not a system failure — it is a system working exactly as designed, but trained on data that reflects historical inequities. Think of it like a compass calibrated in a magnetically distorted environment: the instrument functions perfectly, but consistently points in the wrong direction. AI models trained on financial data are susceptible to the same distortion, reproducing lending disparities, risk-scoring gaps, or product recommendation biases that have historically disadvantaged certain demographic groups, often without any explicit discriminatory intent.

For NBM, this is not a theoretical concern. The RAG-powered AI assistant will draw on multi-generational client transaction data to generate risk assessments and personalized financial recommendations. If that historical data encodes patterns of underservice toward certain zip codes, income brackets, or demographic segments (as is common in community banking data), the model will learn and amplify those patterns unless actively corrected.

Mitigation Plan:

As introduced in the Organizational Design section, NBM's AI framework will incorporate a **disparate impact monitoring tool** that continuously flags any demographic group falling below a 0.80 impact ratio — the threshold established under the EEOC's Uniform Guidelines and broadly adopted in fair lending analysis. This is not a one-time audit; it is a live monitoring layer embedded in the system's operational architecture.

Beyond the technical safeguard, bias governance requires human accountability. The AI/ML Engineer will be responsible for quarterly bias audits across all model outputs, with findings reported directly to the CTO and flagged to the Legal and Compliance team. Any systematic deviation beyond acceptable tolerance must trigger a mandatory model review before deployment continues. The AI compliance consultant engaged during the first 6–9 months of implementation will be specifically responsible for training the existing CCO and legal staff on how to interpret bias flags, when regulatory escalation is required, and how to document the bank's response posture for regulatory examiners — a skill set that does not currently exist within NBM's compliance function and cannot be assumed.

It bears emphasizing: under the Equal Credit Opportunity Act (ECOA) and the Fair Housing Act (FHA), NBM retains full legal accountability for any credit decision influenced by its AI system, regardless of how that decision was generated. The model provides speed; the institution owns the outcome. This principle must be codified in NBM's AI governance policy, not left to interpretation.

II. Data Privacy and Security: The Fiduciary Obligation Extends to the Algorithm

In wealth management, client data is not merely an operational input — it is the foundation of a fiduciary relationship. Clients who share their income, assets, spending behaviors, and financial goals with NBM do so under an implicit and legally enforceable expectation of confidentiality. Extending that data into an AI system does not dissolve that expectation; if anything, it heightens it.

The proposed RAG architecture, which grounds the AI assistant in NBM's proprietary client data, creates meaningful surface area for data exposure if not properly architected. A RAG system, by design, retrieves and surfaces contextually relevant information in response to queries. Without strict access controls and retrieval boundaries, a poorly scoped query from one advisor could inadvertently surface sensitive information about another client.

Mitigation Plan:

NBM's existing Microsoft Azure infrastructure provides a strong foundation for data governance. The implementation must leverage **Azure Active Directory (AAD) role-based access controls (RBAC)** to ensure that the AI assistant only surfaces client data to the advisor or client who is authorized to access it. Each retrieval query must be scoped to the authenticated user's permitted data namespace — meaning the system must authenticate before it retrieves, not after.

All client data used to train, retrieve, or inform AI outputs must be subject to the following non-negotiable controls:

- **Data minimization:** The RAG pipeline should only ingest the minimum data necessary to generate a relevant recommendation. Personal identifiers not required for a given use case should be masked or excluded at the ingestion layer.
- **Encryption at rest and in transit:** All client financial data must be encrypted using AES-256 at rest and TLS 1.3 in transit, consistent with NBM's existing security standards and federal banking regulations under the Gramm-Leach-Bliley Act (GLBA).
- **Audit logging:** Every AI-generated output that involves a client data retrieval event must generate an immutable audit log, including what was retrieved, by whom, when, and in response to what query. This log must be accessible to compliance staff and available for regulatory examination.
- **Data retention alignment:** The AI system's data retention policies must mirror NBM's existing client data retention policies. Data that NBM is obligated to purge under applicable regulations must be purged from the AI's training and retrieval layers as well.

Additionally, the AI interface must include clear, plain-language disclosures that inform users

when they are interacting with an AI-generated output and what data was used to produce it. This is both an ethical obligation and an emerging regulatory expectation, consistent with guidance from the Consumer Financial Protection Bureau (CFPB) on AI transparency in financial services.

Finally, **vendor risk** deserves explicit attention. Even though NBM is building on Microsoft's Azure ecosystem rather than procuring a black-box third-party AI solution, the organization must maintain a clear contractual understanding of where client data flows, how it is processed, and what Microsoft's obligations are in the event of a breach. NBM's legal team, in coordination with the AI compliance consultant, must review and formalize these terms before go-live.

III. Intellectual Property and Copyright: Who Owns What the AI Generates

This risk category is the most frequently overlooked in AI governance frameworks and, for institutions like NBM, potentially the most consequential from a reputational standpoint. Generative AI systems, including the large language model underpinning NBM's assistant, are trained on vast corpora of text that may include copyrighted financial research, proprietary market analyses, third-party reports, and licensed investment content. When the AI generates a client recommendation or an advisor briefing, the provenance of that content is not always transparent.

There are two distinct exposure scenarios NBM must address. The first is **input-side copyright risk**: the risk that proprietary third-party content (research reports, market analyses, licensed data feeds) is inadvertently ingested into the RAG pipeline and surfaced in AI outputs without proper licensing authorization. The second is **output-side ownership ambiguity**: the question of who owns an AI-generated financial recommendation and what legal or fiduciary liability attaches to it.

Mitigation Plan:

On the input side, the Data Engineer responsible for building and maintaining NBM's data pipelines must conduct a content provenance audit before any document corpus is indexed into the RAG system. Every data source ingested into the retrieval layer must be either proprietary to NBM (client records, internal policies, advisor notes) or licensed for use in AI applications. Third-party research or market data from providers such as Bloomberg, Morningstar, or FactSet must be reviewed against their licensing agreements to confirm AI-use permissions — many legacy financial data contracts predate AI use cases and contain ambiguous or restrictive language.

On the output side, NBM must establish a clear institutional policy: **all AI-generated content is a draft, not a decision.** No recommendation, risk assessment, or client communication produced by the AI may be transmitted to a client without advisor review

and sign-off. This "human-in-the-loop" principle serves a dual purpose: it maintains the fiduciary relationship between advisor and client, and it ensures that legal accountability for the output rests with a licensed human professional rather than an autonomous system.

IV. The Legal-AI Interface: A Structural Imperative

Across all three risk domains, a consistent theme emerges: **NBM's existing legal and compliance infrastructure is not yet equipped to govern AI.** This is not a criticism — it is simply the reality of deploying a technology that has outpaced regulatory vocabulary in virtually every industry it has entered. The General Counsel and CCO at a community bank of NBM's scale were trained, certified, and operationally structured around a world of human-executed decisions, static software systems, and established regulatory frameworks. AI changes all three of those assumptions simultaneously.

This is precisely why the engagement of a dedicated **AI compliance consultant** is not optional. It is the structural bridge between NBM's current legal competence and the governance demands of the AI system being deployed. The consultant's mandate must include four deliverables before the system goes live:

1. **AI Governance Policy:** A formal, board-approved document defining NBM's principles for AI use, data handling, bias monitoring, human oversight requirements, and escalation protocols.
2. **Regulatory Mapping:** An assessment of how existing regulations like GLBA, ECOA, FHA, CFPB guidance, and any applicable Vermont state banking regulations apply to NBM's specific AI use cases, with identification of any gaps where regulatory interpretation is still evolving.
3. **Incident Response Protocol:** A defined playbook for what happens when the AI produces a discriminatory output, a data breach occurs, or a client disputes an AI-influenced recommendation.
4. **Knowledge Transfer:** A structured training program that leaves the CCO and legal team genuinely capable of governing the AI independently after the consultant's engagement concludes — not merely aware of the risks, but operationally equipped to manage them.

The relationship between legal and AI governance at NBM cannot be a one-time consultation. As the model evolves, as client interactions generate new data, and as the regulatory landscape matures, NBM's compliance function must evolve in parallel. The consultant engagement is the starting point, not the finish line.

Governance in Summary

Responsible AI deployment at NBM is not a constraint on the opportunity — it is the condition that makes the opportunity sustainable. A community bank's most durable competitive asset is the trust it has built across generations of clients. That trust is not diminished by deploying AI; it is amplified when clients see that their institution applies the same care and rigor to its technology as it does to its advisors. Done right, NBM's AI governance framework becomes a differentiator — a signal to clients, regulators, and the broader market that NBM is not simply adopting AI because it can, but deploying it because it should, and doing so with the institutional discipline that defines a genuinely trustworthy financial partner.

CONCLUSION

This proposal was designed around three constraints that are specific to a community bank of NBM's scale: it cannot afford to fail visibly, it cannot take on implementation risk that outpaces its compliance capacity, and it cannot build something that requires a technology organization it does not have. The Azure-native RAG architecture respects all three. It is built on a platform NBM already licenses, monitored by a governance layer that is live before the first client interaction, and staffed by two engineers — not a department.

The financial return — $9.6M NPV, 87% IRR, break-even in Year 1 — is modeled on conservative assumptions. The 25% advisor capacity increase is not a target; it is the direct result of reducing per-client time from four hours to one. These are not reasons to proceed despite the investment. They are evidence that the investment is justified on its own terms.

The remaining question is not whether AI belongs in wealth management. That question has been settled by the competitive landscape. The question is whether NBM shapes how it arrives at this institution, or inherits a version of it built by someone else.

REFERENCES

[1] Egan, M. B. (2025, May 22). *Financial avoidance: The fears and habits holding your customers back.* The Financial Brand. https://thefinancialbrand.com/news/financial-education/financialAvoidance-the-fears-and-habits-holding-your-customers-back-189384

[2] Marks, G. (2025, May 18). *Business tech news: Klarna reverses on AI, says customers like talking to people.* Forbes. https://www.forbes.com/sites/quickerbettertech/2025/05/18/businesstech-news-klarna-reverses-on-ai-says-customers-like-talking-to-people/

[3] Miller, S. M., Bhattacharya, L., & Davenport, T. H. (2023). *DBS Bank: A tech company going all in on AI* (Version 2023-09-21) [Case study]. Singapore Management University.

[4] Milmo, D. (2023, February 2). *ChatGPT reaches 100 million users two months after launch.* The Guardian. https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app

[5] Mink, Z., Liu, J., Cheung, R., Sharma, S., & Mossalgue, J. (2026, January 21). *AI takes center stage at Davos.* The Rundown AI. https://www.therundown.ai/p/ai-takes-center-stage-at-davos

[6] Singla, A., Sukharevsky, A., Yee, L., Chui, M., & Hall, B. (2025). *The state of AI: How organizations are rewiring to capture value* [PDF]. McKinsey & Company.

[7] World Economic Forum & Accenture. (2025). *AI in Action: Beyond experimentation to transform industry* (AI Governance Alliance). World Economic Forum.

[8] Automation Anywhere. (2026). *AI in banking: How banks use AI to improve risk and operations.* https://www.automationanywhere.com/company/blog/automation-ai/ai-banking-how-banks-use-ai-improve-risk-and-operations

[9] Boston Consulting Group. (2026, March 9). *How retail banks can put agentic AI to work.* https://www.bcg.com/publications/2026/how-retail-banks-can-put-agentic-ai-to-work

[10] Finastra. (2026, February 24). *AI in banking and financial services: Trends for 2026.* https://www.finastra.com/viewpoints/articles/future-of-ai-in-financial-services-2026

[11] IBM Institute for Business Value. (2025, January 27). *2025 global outlook for banking and financial markets.* https://www.ibm.com/thought-leadership/institute-business-value/en-us/report/2025-banking-financial-markets-outlook

[12] Visionet. (2025, March 12). *The evolution of mortgage underwriting: AI, automation, and real-time data in 2025.* https://www.visionet.com/blog/the-evolution-of-mortgage-underwriting-ai-automation-and-real-time-data-in-2025

[13] Wolters Kluwer. (2026). *The AI imperative in banking: Moving from pilot to production.* https://www.wolterskluwer.com/en/expert-insights/the-ai-imperative-

in-banking-moving-from-pilot-to-production

AI PROMPTS & DATA METHODOLOGY

This appendix documents the AI tools and prompts used to generate the synthetic data, machine learning models, and visualizations supporting this white paper. All Python code was developed with AI assistance (Claude), reviewed and tested by team members. White paper prose was written by team members without AI assistance. All client data is 100% synthetic — no real customer or financial data appears anywhere in this project.

Additional AI Tools Used

- **Claude (Anthropic)** — used for all Python code generation, data simulation, and visualization scripts.
- **Google Gemini** — used during research and ideation phases. Transcript: `https://g.co/gemini/share/16e2bf13f7be`

Data Collection Methodology

1. Population & Geographic Grounding

All synthetic client data is anchored to real, publicly available census data for **Middlebury, Vermont (2024)**. Because NBM is a real community bank serving the Middlebury area, the client base was designed to reflect actual local demographics (~8,000 residents, anchored by Middlebury College).

Sources consulted:

- U.S. Census Bureau, American Community Survey (2024) — Middlebury, VT
  - Household income distribution (bucketed by income band, shares normalized)
  - Race/ethnicity composition (White 79.6%, Asian 6.8%, Hispanic 5.4%, Black 1.6%, Multiracial/Other 6.6%)
  - Housing tenure (56.8% owner-occupied — Middlebury College drives the high renter share)
  - Gender split (roughly 50/50; small non-binary category at 2%)

Key methodological choices:

- Income below $25,000 excluded from wealth management client pool (minimum viable

client threshold)
- Income follows a bimodal distribution: service/education workers cluster low, faculty/professionals cluster high
- Bucket-level income draws use log-normal noise ($\sigma = 0.15$) to avoid artificial discrete spikes
- Age-income correlation peaks near 50 years (career prime) and tapers in early career and retirement

2. Synthetic Data Design Principles

The synthetic dataset (`data/synthetic_clients.csv`, n = 25,000) was designed to be **internally consistent** — correlations between variables mirror real-world relationships:

Relationship	Direction	Implementation
Age → Risk tolerance	Younger = higher risk	Weighted composite (55% age, 25% income, 20% random)
Income → AUM	Higher income = higher AUM	AUM = income × (1.5 + 0.1 × tenure) × lognormal noise
Age → Home ownership	Older = more likely to own	Baseline 56.8% + modulation for age and income
Age → Tenure	Older = longer tenure	Beta distribution scaled to min(age − 20, 40)
Income → Transactions	Higher income = more/larger	Poisson-distributed count + proportional size

3. Transaction Data Methodology

Monthly transaction records (`data/synthetic_transactions.csv`) were generated as a 12-month simulation per client, capturing:

- **Seasonal variation:** Q4 spending peaks, Q1 troughs (sinusoidal ±10% amplitude)
- **Deposits, withdrawals, investment activity:** Random proportions of total monthly volume (e.g., deposits = 40–70%)
- **Client-level noise:** Log-normal multiplier on transaction size to simulate real behavioral variance

4. Policy Document Authorship

Eight NBM policy and product documents (`data/nbm_documents/`) were written to resemble real community bank compliance language for RAG ingestion. Topics: savings products, retirement accounts, investment products and fee schedule, compliance guidelines (KYC,

AML, GLBA), client FAQ, investment philosophy, digital services, and the risk questionnaire. The specific rates, fees, and policy details are synthetic and illustrative — they do not represent NBM's actual current policies.

Prompts Used to Generate Each Component

Prompt 1 — Synthetic Data Generator (`models/data_generator.py`)

> "I'm building a synthetic client dataset for the National Bank of Middlebury (NBM), a real community bank in Vermont. The bank is a wealth management firm serving roughly 25,000 clients. I need realistic synthetic client profiles with correlated variables — age should correlate with risk tolerance and AUM, income should follow Middlebury VT 2024 census distributions, and housing tenure should match the 56.8% owner-occupancy rate from census data. Include race/ethnicity proportions from the census. Generate transaction summaries for 12 months per client with seasonal variation. Also generate realistic bank policy documents (savings products, retirement accounts, compliance guidelines, FAQ) that can be ingested into a RAG pipeline. Everything must be 100% synthetic — no real client data."

What this produced: 25,000 client profiles (16 fields), 300,000 monthly transaction records, and 8 policy/product markdown files for RAG ingestion.

Prompt 2 — Risk Profiling ML Model (`models/risk_profiler.py`)

> "Build a Gradient Boosting classifier that predicts client risk tolerance (Conservative / Moderate / Aggressive / Very Aggressive) from the synthetic client data. Use an 80/20 train/test split with stratification. Train both a Gradient Boosting and a Random Forest for comparison. Use permutation importance instead of built-in feature importance for reliability. Generate four charts: a horizontal bar chart of feature importance, a confusion matrix heatmap with both percentage and raw count annotations, one-vs-rest ROC curves for all four classes, and an example client prediction showing confidence scores as a bar chart. Frame the results around the Prediction Machines thesis — AI lowers the cost of prediction, freeing advisor judgment for higher-order decisions. Target ∼77% accuracy."

What this produced: GradientBoostingClassifier ($n = 200$ estimators, max depth 5, learning rate 0.1) and RandomForestClassifier for comparison; 5-fold cross-validation; 4 output figures.

Prompt 3 — Bias / Fairness Audit (`models/bias_audit.py`)

> "Write a bias audit tool for the risk profiling model. Intentionally introduce subtle historical biases into the training data — younger clients scored more conservatively than warranted, and rural clients systematically under-predicted for aggressive risk profiles. Then audit the model using three fairness metrics: disparate impact ratio (flag if < 0.80, per EEOC 4/5ths rule), demographic parity difference, and equalized odds difference. Audit across age groups, gender, and region. Demonstrate a mitigation strategy using sample re-weighting. Generate three charts."

What this produced: Bias injection + detection pipeline; three fairness metrics per demographic group; pre/post-mitigation comparison charts.

Prompt 4 — ROI / Economic Model (`models/roi_model.py`)

> "Build a 5-year financial projection for an AI wealth management implementation at a 50-advisor community bank with 10,000 total clients. Hardcode realistic Azure cost line items. Model a 25% advisor capacity increase ramping up over 5 years with realistic client acquisition rates (15%, 30%, 50%, 70%, 85%). Compute NPV at 12% discount rate, IRR, and break-even year. Run a Monte Carlo simulation with 10,000 trials. Target NPV ∼$9.6M, IRR ∼87%, break-even Year 1."

What this produced: Deterministic 5-year cash flow model; Monte Carlo with ±20% noise on key assumptions; key results: $9.6M NPV, 87% IRR, Year 1 break-even, 100% probability of positive ROI.

Prompt 5 — RAG Pipeline Prototype (`models/rag_prototype.py`)

> "Build a working RAG pipeline demo that ingests the NBM policy documents. In demo mode (no API key), use TF-IDF for embeddings and cosine similarity retrieval. In live mode (`-live` flag), use OpenAI text-embedding-3-small and GPT-4o-mini. The pipeline should have clearly labeled stages: Ingest → Chunk → Embed → Index → Retrieve → Generate. Each stage should map to an Azure production equivalent."

What this produced: Two-mode RAG implementation (TF-IDF demo + OpenAI live); ChromaDB vector store for live mode; pipeline diagram and text report.

Prompts 6–11 — Visualizations & Brand System

- **Client Demographics Chart** (`visualizations/charts.py`): 2×3 panel — age histogram, AUM distribution, income distribution, risk by age group, race/ethnicity breakdown, region and home ownership.

- **Advisor Capacity Waterfall**: Waterfall chart from 200 to 250 clients per advisor showing four AI-driven increments.
- **Advisor Time Allocation**: Before/after pie charts showing how AI shifts advisor time from research and documentation to client engagement and business development.
- **Azure Technology Stack Diagram**: Five-layer architecture diagram (Presentation, Application, AI/ML, Data, Security) using `FancyBboxPatch` rectangles.
- **System Architecture Diagram** (`diagrams/architecture.py`): End-to-end pipeline from client touchpoints through Azure services to advisor output with human-in-the-loop governance.
- **NBM Brand Style System** (`visualizations/style.py`): Centralized color constants (NAVY `#1B3A6B`, GOLD `#C9A84C`, etc.) and `apply_nbm_style()` for consistent matplotlib styling across all scripts.

Reproducibility

All scripts use `numpy.random.seed(42)` and `random_state=42` throughout. Running the scripts in the order below will reproduce all data and figures exactly:

```
python3 models/data_generator.py   # Must run first
python3 models/risk_profiler.py
python3 models/bias_audit.py
python3 models/roi_model.py
python3 models/rag_prototype.py
python3 diagrams/architecture.py
python3 visualizations/charts.py
```