

As artificial intelligence tools such as **ChatGPT** and **Google Gemini** become increasingly embedded in financial services workflows, the lure of faster preliminary research, automated report parsing, and concise summarization is undeniable. Yet for family offices and multi-family offices managing ultra-high-net-worth (UHNW) clients, the stakes are high. Introducing AI without rigorous pilot testing risks relying on persuasive but inaccurate outputs. This blog post explores best practices around pilot testing AI tools before firm-wide adoption, focusing on mitigating the risk of *hallucinations*, aligning use cases with AI capabilities, and establishing robust error tracking mechanisms — all within a family office context.

Why Caution Is Critical: The Risk of AI Hallucinations

Among AI's greatest challenges is the phenomenon known as "hallucinations" — outputs that are coherent, confident, and plausible but factually incorrect. Both ChatGPT and Google Gemini, despite rapid advancements, are susceptible. Hallucinations pose a unique risk in wealth management environments where decisions based on inaccurate data can significantly impact client trust and portfolio outcomes.

Faithfulness vs Factual Hallucinations

Understanding *faithfulness* in AI output is essential. Faithful responses accurately reflect the underlying data or requested information source. However, AI might generate factually hallucinated statements — information unsupported or contradicted by known facts. This distinction matters hugely for a family office, especially when parsing nuanced private markets data where public data scarcity exacerbates error risk.

Private market data often lacks transparency and standardization, so AI tools may "fill gaps" with plausible but unfounded information. This highlights why pilot testing with domain-specific benchmark questions is a necessity before firm-wide rollout.

Family Office Use Cases for AI Tools

Before testing, clearly identify where AI tools will be applied. Common family office workflows where AI can add value include:

- **Preliminary research:** Scanning news, industry trends, and fund documents to surface relevant insights.
- **Parsing and structuring reports:** Extracting key data points from manager reports, capital calls, and performance summaries.
- **Summaries:** Creating concise overviews of lengthy documents, meetings, or external investment memos.

Each use case has distinct accuracy requirements and risk profiles. For instance, summarization tools might tolerate some degree of paraphrasing, but parsing performance data demands high factual precision.

Pilot Testing: The Cornerstone of Responsible AI Adoption

Rolling out AI tools without a pilot is akin to launching an investment fund without backtesting. Pilot testing acts as the guardrail, enabling teams to understand tool limitations, calibrate expectations, and design mitigation plans.

Step 1: Define Benchmark Questions

Dive deep into your typical use cases and extract representative tasks to build a battery of *benchmark questions*. These questions must mirror canadianfamilyoffices.com the complexity and domain specificity your office



Use pilot feedback to refine prompt strategies, restrict AI usage to lower-risk tasks, or combine AI output with human review. Many family offices adopt a hybrid approach: automating initial document parsing but retaining analyst validation for final reports.

Comparing ChatGPT and Google Gemini for Family Office Pilots

Feature / Aspect	ChatGPT	Google Gemini
Hallucination Rate	Known to hallucinate; success depends heavily on prompt quality and domain specificity.	Emerging tool with advanced context retention; still requires pilot testing to confirm real-world accuracy.
Domain Adaptability	Good with fine-tuning; integrates well with document ingestion APIs.	Designed with multi-modal data fusion, potentially stronger for combining text and quantitative data.
Auditability	Limited traceability; depends on user-driven logging and external monitoring.	Google emphasizes interpretability but real audit trails remain an implementation challenge.
Deployment Options	Cloud-based with API access; on-prem options limited.	Proprietary cloud with planned enterprise features; pilot availability varies.

Practical Tips for Family Offices Piloting AI Tools

1. **Maintain a 'Things AI Got Wrong' Log:** Keep an agile, plain-text running note of hallucination instances or odd outputs to build collective institutional knowledge over time.
2. **Prioritize High-Impact Use Cases:** Start with non-decision-critical functions such as summarization or thematic scanning before moving to data extraction influencing client reports.
3. **Integrate with Data Governance:** Ensure AI outputs are tagged with metadata on version, prompt, and review status for transparency and audit readiness.
4. **Train Staff on Limitations:** Avoid treating chatbot answers as final or cited sources; foster a culture of critical engagement and verification.
5. **Establish Continual Monitoring:** Post-pilot, implement periodic benchmark retesting to monitor drift in AI tool performance as models update.

Conclusion

Pilot testing AI tools like ChatGPT and Google Gemini is not just best practice; it is essential for safeguarding family offices against the seductive but risky traps of AI hallucinations. By defining benchmark questions, tracking errors meticulously, involving cross-functional reviewers, and integrating findings into data governance frameworks, family offices can responsibly adopt AI to enhance preliminary research, report parsing, and summarization tasks.

Ultimately, a disciplined pilot testing approach balances innovation with control, empowering teams to harness AI's efficiency gains while protecting the fidelity and faithfulness of client insights.

