

In the rapidly evolving landscape of AI-assisted workflows, grounding an AI assistant on your company's knowledge base is a foundational step towards delivering reliable, context-aware, and actionable support to your users. Whether you're a Product Manager at a startup or a Lead at an enterprise leveraging AI at scale, understanding the intertwining of retrieval, source checks, and document Q&A is crucial to bridging the gap between "model as research" and "model as product."

This article walks through pragmatic strategies, pitfalls, and tooling options inspired by companies like **PM Toolkit**, **PayPal**, and **H2O.ai**. Alongside, we'll reflect on lightweight eval approaches with spreadsheets, and the critical mindset shift you need to acknowledge that *the model is not the product*.

Why Grounding Matters: The Model Is Not the Product

It's tempting to conflate a powerful language model with a final product. After all, if the model outputs plausible answers, isn't the problem solved? Not quite. Even state-of-the-art models like GPT-4 or open-source variants lack context awareness beyond the training data or the immediate prompt. Without grounding on your 📖 knowledge base:

- Answers risk being hallucinated or outdated
- Consistency suffers across repeated interactions
- User trust can erode quickly

Companies like **PayPal** have faced these issues when integrating AI assistants into customer support workflows, understanding that trust and accuracy require rigorous source verification and linking answers explicitly back to real, curated documents.

Step 1: Define Your Use Case and Eval Before Release

Before diving into model fine-tuning or retrieval architectures, clarify exactly what you want your AI assistant to do. Do users need it for:

- **Document Q&A:** Quickly retrieve specific answers from your knowledge base?
- **Source Checks:** Verify that the AI response references original company policies or data?
- **Workflow Automation:** Execute routine tasks or enable intelligent augmentations?

Take inspiration from how **PM Toolkit** embeds calculators and decision frameworks — such as TAM/SAM/SOM analysis, ROI estimations, retention & churn metrics, RICE prioritization, and A/B test planning — right into spreadsheets and toolkits. These are real-world user stories: "As a PM, I want to estimate my addressable market size quickly, so I can validate a new product idea." Ground your evals similarly.

Creating Your Eval Checklist

Based on your use case, [pmtoolkit](#) create a spreadsheet-based eval framework capturing key metrics and scenarios. For example:

Eval Aspect	Metric/Measure	Data Source	Threshold/Goal	Method	Accuracy	Correct answer vs knowledge base
Sample queries & annotated answers	≥ 90%	Manual reviews & spot checks	Source Transparency	Percentage		
references to original docs	System logs	100%	Automated link extraction	Response Latency	Time to answer	

(seconds) System logs ≤ 3 seconds Load testing User Satisfaction NPS or PMF score User surveys via PM Toolkit tools $\geq 60\%$ Periodic surveys post-interaction

This aligns with principles from **PM Toolkit**'s calculators for velocity, conversion rate, and NPS measurement — all critical for evaluating an AI assistant as a living product, not just a demo.

Step 2: Leverage Retrieval as Your AI Assistant's Backbone

Retrieval-augmented generation (RAG) has become the gold standard for grounding AI assistants on company knowledge. Instead of solely relying on the model's internal weights, the AI performs:

1. **Document retrieval:** Search your indexed knowledge base to surface relevant content.
2. **Context construction:** Concatenate retrieved snippets to feed the model, guiding it towards grounded answers.
3. **Answer generation:** Produce responses based on this combined input.

H2O.ai notably incorporates retrieval mechanisms into their AI platform, allowing data scientists to fine-tune responses based on business datasets and documentation, exemplifying how retrieval enables accuracy and improves source fidelity.

Choosing Your Retrieval Techniques

- **Keyword matching & BM25:** Good for structured text corpora with known vocabulary.
- **Embedding-based similarity:** Use sentence transformers or vector databases to semantically match queries to documents.
- **Hybrid approaches:** Combine fast keyword filters with semantic embeddings for scale and precision.

Top-tier teams often version-tune their retrieval pipelines alongside their models, measuring how retrieval impacts downstream answer quality in real user task scenarios.

Step 3: Implement Rigorous Source Checks and Document Q&A

Grounding is incomplete without transparency. Your AI assistant must cite its sources clearly — this is critical in regulated sectors or when users rely on compliance-related knowledge.

Consider the following best practices:

- **Source highlighting:** Include inline references to document titles, sections, or timestamps.
- **Confidence scoring:** Have the model or pipeline assign confidence to answers based on retrieval strength and consistency.
- **Human-in-the-loop triggers:** For low-confidence queries, route to a human reviewer to prevent hallucination.

PayPal's AI-driven customer support uses such layered approaches, combining retrieval with source checks to mitigate risk. They also utilize spreadsheet-driven eval suites to monitor false positive/negative rates and churn in accuracy over time.

Step 4: Understand Task-Dependent Model Behavior and Failure Modes

Not all AI tasks are created equal. For example, question-answering over a stable company wiki has different failure modes than intent detection or summarization of dynamic transactional data.



It is vital to map out expected model behaviors and possible failures specifically for your use case. E.g.,

- **For document Q&A:** Failure mode is a confident but fabricated response.
- **For workflow automation:** Failure mode can be executing an unintended action.
- **For innovation tasks (ideation, strategy):** Failure may mean missing out on novel directions or hallucinated insights.

Tracking these in your eval examples, written as detailed user stories — rather than synthetic “lab prompt” inputs — helps you build comprehensive mitigate-and-monitor plans.

Step 5: Pattern Match Your Workflow Strategy – Automation, Augmentation, or Innovation?

How you ground your AI assistant also depends on your broader automation strategy:

Pattern Description Knowledge Base Role Example Automation AI performs tasks independently, reducing human effort Strong grounding needed with clear error handling PayPal’s chargeback dispute bot relying on policy docs Augmentation AI supports humans in decision-making, offering speed and scale Grounding with retrieval plus confidence indicators to aid human judgment PM Toolkit’s calculators blending AI suggestions with user inputs Innovation AI generates novel ideas, hypotheses, or exploratory analysis Loose grounding; user validation essential to catch hallucinations H2O.ai’s AI-powered data scientist assistant proposing new models

Knowing your pattern keeps your expectations clear and informs how strict your grounding pipelines and eval thresholds need to be.

Bringing It All Together: Tools and Best Practices

Lightweight Eval Suites in Spreadsheets

Instead of heavy engineering overheads, many teams start with spreadsheet-driven eval stacks. Inspired by **PM Toolkit**, you can integrate multiple calculators into Google Sheets or Excel to track KPIs like:

- Accuracy and precision of document answers over time
- User satisfaction via NPS and PMF score surveys
- Impact versus effort matrices (RICE, ICE, Kano) for prioritization
- Cycle time, velocity, and DAU/MAU metrics for usage tracking

This approach also allows product teams to visualize improvements, spot regression, and manage tradeoffs transparently.

Automated Pipelines for Production-Scale Grounding

Once your eval thresholds are proven, scale by building pipelines that:

- Automate document ingestion & indexing
- Integrate retrieval and model inference with source citation
- Maintain kill switches and fallback paths for failure scenarios
- Continuously monitor performance with alerting on eval criteria

Remember: shipping without a kill switch is the cardinal sin of AI product management!

Conclusion

Grounding an AI assistant on your knowledge base is not a “set-and-forget” task. It requires a thoughtful blend of rigorous retrieval methods, transparent source checks, task-aware failure mode analysis, and continuous evaluation — all underpinned by clear user-centered use cases. Draw inspiration from proven companies like **PM Toolkit**, **PayPal**, and **H2O.ai**, who exemplify balancing automation with augmentation and innovation.

Start small with spreadsheet-driven evals that reflect real user stories, plan your retrieval strategy carefully, and always remember: *the model is not the product*. Your product’s integrity depends on the systems you build around the model.

Equip yourself with patience, curiosity, and a personal checklist of “ship/no-ship thresholds” — the hallmark of successful AI product leadership.

