

Building AI products without reliable ground truth data is one of the stickiest problems in product management and machine learning engineering today. When you lack labeled data or “gold standard” answers, how do you know if your model is actually helping users—or just creating noise? This challenge is acute in B2B SaaS, where companies like **PayPal** employ AI for fraud detection, **H2O.ai** empowers businesses with automated machine learning platforms, and even **PM Toolkit** offers calculators and frameworks to structure product decisions.

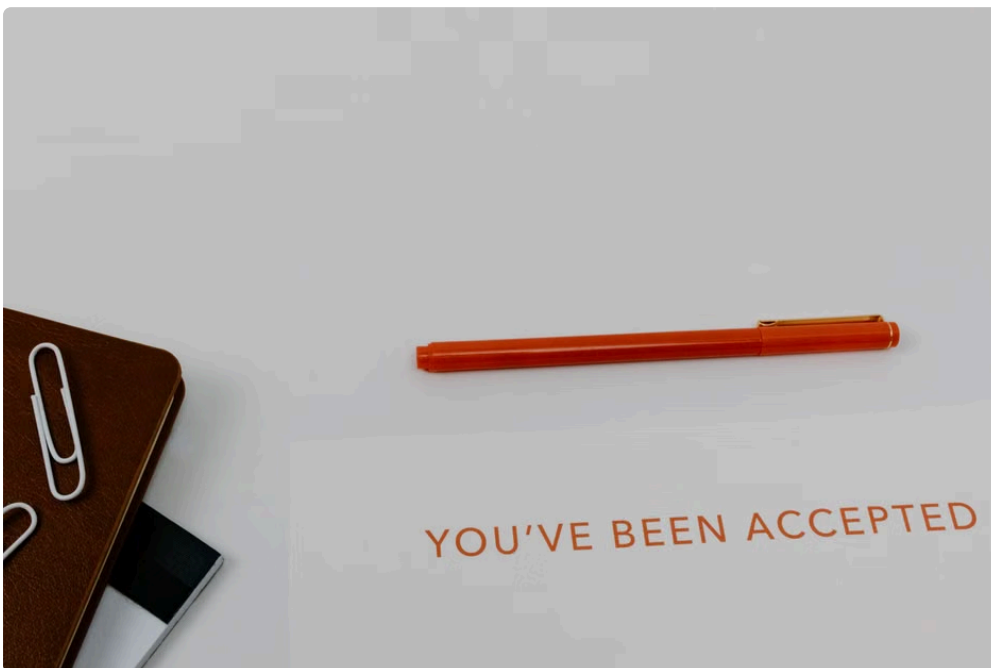
In this post, we’ll walk through practical strategies to ship responsibly when you don’t have ground truth, grounded in four critical pillars:

- Understanding why *the model is not the product*
- Establishing an **eval first** mindset with measurement plans before launch
- Recognizing *task-dependent model behavior and failure modes*
- Choosing between *automation, augmentation, and innovation* patterns

Along the way, we’ll integrate lessons from PM Toolkit’s robust calculators—like RICE prioritization, sample size estimation, and NPS tracking—and share how spreadsheets remain an invaluable ally for lightweight evaluations. Let’s dive in!

Why Ground Truth Matters—and What to Do When It’s Missing

Ground truth usually means a known, verified “correct answer” for each item you want the AI to handle. For example, in fraud detection at **PayPal**, ground truth labels might be confirmed fraudulent transactions. But many AI applications lack this luxury—think of novel document summarization, creativity assistants, or complex anomaly detection where human judgment differs widely.



Without ground truth, you can’t measure “accuracy” in the traditional sense. You risk launching models that appear to perform well on proxies, yet fail [Check out this site](#) silently in production, eroding user trust.

But all is not lost! Ground truth is a spectrum, not a binary. Even if you lack perfect labels, you can:

- Define alternative **labeling plans** involving human agreement or consensus
- Design evaluations that capture meaningful user outcomes

- Focus on model behavior aligned with real human work—avoiding pure benchmark chasing

The Model Is Not the Product: Centering Workflow and User Outcomes

This mantra is crucial. AI models don't operate in isolation; they embed into workflows, driving decisions or automating tasks. The end user cares about outcomes—time saved, errors reduced, revenue gained—not raw model scores.

In B2B SaaS, this means you must understand:

- What are users doing *today* without AI?
- How will this model alter their work?
- What failures or edge cases can disrupt the workflow?

For example, **H2O.ai** designs AutoML tools that empower data scientists by automating model search but keep human-in-the-loop checkpoints to verify predictions. This combination of *augmentation* and trusted automation is what drives adoption.

Building your AI experience without ground truth means modeling these workflows explicitly. That's where PM Toolkit's calculators come in handy: Use **impact/effort** to weigh what parts of the workflow the model affects, and estimate ROI with the **business case** calculator by plugging in conversion rates, activation rates, and CAC figures.

Define Your Eval Before Release: The “Eval First” Checklist

One of my non-negotiables is “Define an eval before release.” Without a concrete way to measure success, you're flying blind. When ground truth is fuzzy, you must pivot to innovative, task-aligned evaluations.

Here's a step-by-step approach to building your evaluation plan:

1. **Outline your user story-based evals:** Write example scenarios, not lab prompts. For instance, “When a fraud analyst reviews the flagged transactions generated by this model, do they reduce investigation time by 20%?”
2. **Establish labeling plans:** Use human agreement metrics where possible. If labels are subjective, measure inter-rater reliability to understand consensus.
3. **Leverage A/B testing:** Plan the experiment with PM Toolkit's **A/B test** and **sample size** calculators to ensure statistical power.
4. **Track cohort metrics:** Use metrics like DAU/MAU, NPS, and PMF score to assess user adoption and satisfaction over time.

Many teams underestimate how spreadsheets can be a powerful evaluation hub here. I've found that lightweight, collaborative evaluation suites in Google Sheets can track labeling consistency, calculation of human agreement scores, and align model outputs with defined KPIs without waiting for full data pipelines.

Task-Dependent Model Behavior and Failure Modes

AI models behave differently depending on the task and data context. Without ground truth, recognizing failure modes becomes critical before release to avoid catastrophic user impact.

Common failure modes include:

- *Data drift*: Model struggles when data distribution changes (e.g., new transaction types in fraud detection)
- *Ambiguity*: Variability in user labeling when tasks require nuanced judgment
- *Overconfidence*: Model outputs confident but incorrect results, misleading users

To mitigate these, you can combine human judgment with automation:

- Design confidence thresholds that trigger human review for uncertain cases
- Use RICE or ICE prioritization from PM Toolkit to pick the most valuable failure modes to address first
- Monitor user feedback loops, e.g., via NPS or churn rates, as indirect signals of product health

Automation vs Augmentation vs Innovation Patterns

Understanding your AI's role shapes how you handle ground truth ambiguity:

- **Automation**: AI replaces manual work completely. Requires the highest trust level, often meaning robust ground truth is essential.
- **Augmentation**: AI assists human decisions, providing suggestions or filtering. More forgiving of imperfect ground truth since a human inspects and corrects errors.
- **Innovation**: AI explores entirely new workflows or business models where ground truth might be non-existent by definition. Here, careful eval design and iterative user feedback are your north stars.

PayPal's fraud detection blends automation and augmentation—flagging suspicious transactions automatically but allowing human analysts final say. Their approach highlights why human agreement metrics and continuous feedback loops are vital when ground truth is evolving.

Bringing It All Together: The PM Toolkit Advantage

Numerous PM Toolkit calculators and frameworks support building responsible AI products even without perfect ground truth. Here's how you can draw from them:

PM Toolkit Tool How It Helps in No-Ground-Truth Scenarios TAM/SAM/SOM Quantify market opportunity to prioritize AI impact areas. ROI & Business Case Estimate financial impact before full model labels exist. RICE / ICE / Impact-Effort Focus scarce labeling and eval resources on highest-value work. A/B Test & Sample Size Design statistically valid experiments for task-aligned evals. NPS, PMF Score, DAU/MAU Track real-world user engagement and satisfaction over time. Spreadsheets for evals Run lightweight human agreement scoring and error analysis collaboratively.

By combining these tools with a relentless focus on workflow integration and user outcomes, you create a "ship/no-ship" checklist that keeps your AI product honest and useful.

Final Thoughts: No Ground Truth? No Excuses.

Lack of ground truth is not a showstopper—it's a signal to be more deliberate. Start by deeply understanding what users do today, then build small, testable evals with humans in the loop. Don't chase abstract accuracy metrics or model wars without workflow context.



Follow a methodical eval-first approach anchored by PM Toolkit frameworks, human agreement labeling plans, and continuous impact measurement. And always keep your kill switch handy—never ship without a clear path to disable what disrupts users.

In the end, responsible AI is about persistent iteration between model development and human feedback. Trust emerges from transparent, task-aligned measurement—not from benchmark bragging rights.

Ready to get pragmatic? Try designing your first labeling plan with human agreement scores today, use PM Toolkit's impact/effort calculator to pick the next feature, and build your eval spreadsheet with real user stories as examples. That's how you turn ground truth scarcity into an advantage.

About the author: A product manager turned AI product lead with 12 years in B2B SaaS, specializing in AI-assisted workflows and lightweight eval pipelines—from spreadsheets to automated test frameworks.