In the age of AI-assisted research and analysis, one of the biggest challenges teams face is ensuring factual accuracy while navigating multiple AI outputs. Misstatements—or what practitioners call "hallucinations"—can creep in subtly, undermining investment diligence, legal review, or strategic decision-making. The solution isn't simply to trust a single model; it's to create a **multi-model validation workflow** that has these AI systems *challenge each other* under a strict fact-checking regime.

In this blog post, I'll walk you through a practical, high-integrity approach to catch **factual inconsistencies** by thoughtfully interrogating five AI models in concert. Along the way, we'll see how tools like Flatkey AI and DeepL support persistent context, reduce drift, and feed an adjudicator that provides the final seal of trust. This is how a modern AI-powered boardroom workflow looks—comprehensive, auditable, and resilient.

The Importance of Multi-Model Validation

It's tempting to use a single best-in-class language model and hope for the best when it comes to factual accuracy. But my twelve years supporting investment and legal teams have taught me the costly perils of over-reliance on one tool. Every model has blind spots and occasionally fails to ground its outputs in facts.

Enter the multi-model approach. By running the same queries across a suite of models and then *cross-comparing* their answers, inconsistencies emerge naturally as flags utilo.io for review. This approach leverages diverse training data and architectures and reduces the likelihood of the same hallucination afflicting all models simultaneously.

Some questions you need to ask each model are foundational to **uncover factual inconsistencies:**

- "Can you cite your source for this specific fact?"
- "Are there alternative perspectives or data about this topic?"
- "What is your confidence level in this statement?"
- "Is the information up to date as of [specific date]?"
- "Can you summarize the discrepancies between your answer and the others?'"

As you'll see, these interrogations form the core of an *AI-powered adjudication* process.

Building an AI Boardroom Workflow in One Thread

What sets an effective workflow apart from ad hoc queries or demos? It's a single conversational thread that maintains persistent context, reduces drift in understanding, and layers in continuous checks.

Here's a simplified blueprint of the workflow you can implement today:

1. **Initial Query:** Pose the core question or data request simultaneously to five diverse models.
2. **Individual Responses:** Collect detailed output, including source citations if available.
3. **Inter-Model Challenge:** For each output, ask the other models to verify, refute, or contextualize the facts.
4. **Translation & Clarification:** If responses are multilingual, use DeepL for precise translation to maintain nuance and avoid introducing errors.
5. **Adjudicator Analysis:** Feed all the responses and challenges to an adjudicator model that collates, highlights factual inconsistencies, and provides a final recommendation.
6. **Audit Trail & Persistence:** Use a platform like Flatkey AI to preserve this thread's entire context, responses, challenge history, and adjudicator notes for future reference and compliance.

This linear yet iterative process enables continuous refinement without losing historical context, which is critical to reducing drift—a phenomenon where repeated conversation with AI models moves further away from the original fact base.

Why Persistent Context Matters

Without persistent context, models tend to "forget" earlier conversation points, increasing the risk of contradictory or hallucinated statements. Maintaining a single, persistent thread prevents this drift and saves precious analyst time spent reconciling mismatched outputs.

Role of the Adjudicator in Fact-Checking

A true game-changer is designating one model—or a specialized system—to act as the **Adjudicator**. The Adjudicator's role is to:

- Aggregate all model outputs in one place
- Run comparative logic to identify *exact* factual inconsistencies
- Highlight potential hallucinations or unverified claims
- Provide an evidence-based summary for human reviewers

- Assign confidence or risk scores to each claim

The Adjudicator model becomes the final arbiter before humans take action, reducing the cognitive load on analysts and sharpening confidence that flagged facts withstand scrutiny.

Mapping AI Failure Modes to Questions

From my ongoing notes on AI failure modes, here are common error patterns and how targeted questions to the five models can catch them:

Failure Mode Resulting Issue Model Question to Catch It Action Confabulated facts (invented without basis) Hallucination, misleading data “Please provide your source or reference for this fact.” Challenge other models for source verification; flag unsupported claims. Outdated information Incorrect conclusions due to obsolete data “Is this fact current as of [specific recent date]?” Request updates or corroboration from latest data sources. Semantic drift over conversations Answer deviates from original query intent “Summarize our earlier discussion to ensure consistency.” Reset or reinforce context to avoid cascading errors. Partial information or omissions Incomplete or biased representation “Are there alternative explanations or opposing data points?” Enrich data by cross-checking multiple perspectives. Incorrect data aggregation Misleading summary facts “Compare your summarized data to peers’, note contradictions.” Trigger adjudication to parse conflicting aggregates.

Leveraging Flatkey AI and DeepL to Enhance the Workflow

Flatkey AI is a vital platform here because it helps you:

- Maintain a *single-threaded, persistent context* that prevents drift
- Automatically log all query-response pairs with meta data for auditing
- Integrate multiple models seamlessly via APIs to orchestrate cross-validation
- Build customizable adjudication logic that surfaces inconsistencies clearly

Meanwhile, **DeepL** plays a critical role when your workflow encounters multilingual datasets or documents. Compared to generic translation tools, DeepL offers superior nuance preservation, which is crucial to avoid introducing error in sensitive factual statements.

Example Use Case: Cross-Model Fact-Checking on Industry Data

Let's say your team is analyzing a critical market report with embedded multilingual data points. Here's how you'd apply the workflow practically:

1. Send initial extraction tasks to five models (e.g., GPT family, specialized domain AI)
2. Receive answers, some in German or French—translate these with DeepL into English
3. Perform source validation and inter-model challenges via Flatkey AI's thread
4. Aggregate results and run adjudicator to detect and highlight any factual drift or contradictions
5. Export an auditable report capturing every step of the validation

Final Thoughts: What Is the Fallback When Models Disagree?

Even with the best multi-model setup and adjudication, some discrepancies need human judgment. The crucial question is always *"what is the fallback when AI models contradict?"*

My advice:

- Use the adjudicator's confidence scores to triage items for expert review
- Maintain traceable audit trails for all assertions and challenges (e.g., within Flatkey AI)
- Document final decisions and rationales clearly for compliance and reproducibility
- Don't hide behind claims like "hallucination reduction" without transparent mechanisms

Through disciplined multi-model validation, integration of specialist translation tools like DeepL, and the strict oversight of an adjudicator pipeline, your AI-supported workflows not only reduce factual inconsistencies but create defensible, repeatable routines that boost analyst confidence and speed.

Summary: Questions to Ask Your Five Models to Catch Factual Inconsistencies

- "What sources back this fact or figure?"
- "Is this information current and verified as of [specific date]?"
- "Are there alternative data points or perspectives?"
- "Can you identify discrepancies between your answers and your peers'?"
- "What is your confidence level in your response?"

Ask these consistently, feed answers into an adjudicator, and preserve the full conversation with Flatkey AI to create a workflow that catches hallucinations before they reach decision-makers.

Ready to implement a fact-checking AI workflow that works? Start by creating your multi-model query thread, integrate DeepL for any translations, and adopt Flatkey AI for persistent synchronization and adjudication.