Why Stage-11 Rails Matter: Path A vs Path B

This 1-pager shows how skipping Stage-11 rails (Warp → Detect → Denoise) creates phantom errors, versus how rails enforce determinism. Example prompt: 'What if I borrow 500 USDC against my ETH on Aave?'

■ *Path A — No Stage-11 rails (Naive Classifier)*

- Parser recognizes 'borrow', maps to borrow_asset.
- Phantom bump appears in repay_loan trace (noise).
- Naive classifier accepts both borrow + phantom repay.
- Sandbox executes both → inconsistent state.
- Output looks plausible but includes hallucinated steps.
- False positives ~10–15%.

■ *Path B — With Stage-11 rails*

- Parser recognizes 'borrow', maps to borrow_asset.
- Warp isolates true well; phantom repay collapses.
- Detect confirms borrow signal is strong vs nulls.
- Denoise suppresses weak phantom bumps.
- Sandbox executes only borrow.
- Verifiers run → HF < 1.0 → ABSTAIN.
- User sees clear trace: borrow detected, repay suppressed, liquidation risk flagged.
- False positives <1% (system abstains instead of hallucinating).

■ *Standard AI Baseline*

- Typical LLM agents show 8–12% hallucinated operations on DeFi-style prompts.
- Errors often look fluent and plausible, making them hard for users to spot.
- No abstain mode → users only realize after losses.

Key Message: Stage-11 rails transform a noisy classifier into a deterministic reasoning engine. Only true primitives survive; phantoms are suppressed; verifiers guarantee safety. That’s the leap from 'AI that talks finance' to 'AI you can trust with money.'