BiasMix-Finance: Post-Generation KYC Guardrails for LLM Portfolio Advice

FINAI Workshop @ ICLR 2026

Presenter: Gaurav Kukreja

Gaurav Kukreja • Parul Kukreja • Mohammed Abraar • Raj Dandekar • Rajat Dandekar • Sreedath Panat

gauravkukreja06@yahoo.in

16
ETFs in fixed universe

72
BiasMix scenarios

3×3
Models × inference modes

0%
Final violations after repair

Core idea

Treat LLM allocations as auditable drafts. Validate them against hard KYC-style caps and, when needed, project them to the nearest feasible portfolio.

Why prompt-only portfolio advice is risky

Plausible outputs are not the same as policy-compliant allocations

Observed failure mode

- LLMs can produce ETF allocations that sound sensible while silently violating hard constraints on risk, fees, concentration, or sector exposure.
- Reasoning strategies such as critique and self-consistency may reduce violations, but they remain probabilistic and cannot guarantee compliance on every run.
- This is especially problematic in agentic multi-turn advisory systems, where a draft recommendation can become an externally visible action unless it is checked.

System requirement

Treat the model as an intent generator — then use deterministic verification and repair to enforce admissible actions.

47.6%–85.7%

First-pass any-cap violation range on held-out test

67.2%

Pooled first-pass violation rate

0%

Final feasibility violations after projection

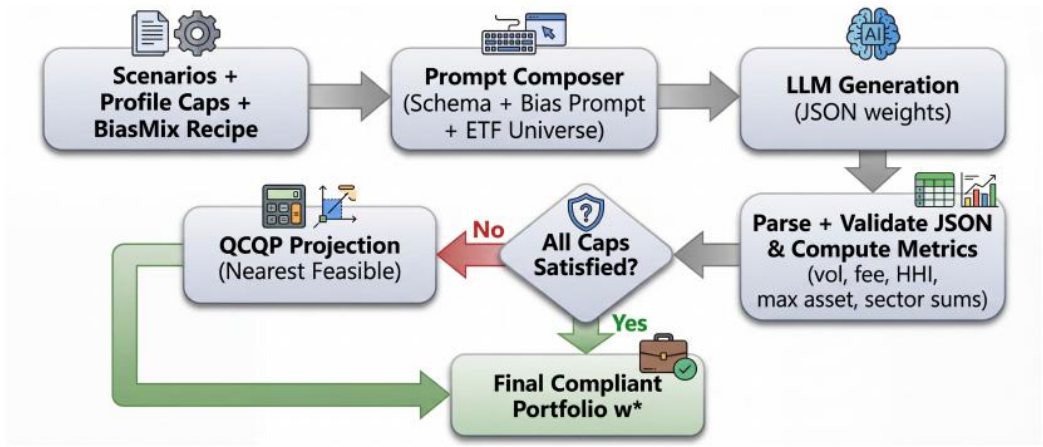
0

Parse failures on held-out test with retries

Takeaway: prompt engineering improves proposals, but the compliance guarantee comes from the post-generation guardrail — not from sampling behavior.

Verify-and-repair pipeline

Deterministic enforcement wrapped around probabilistic model proposals



1 Structured output

Enforce a strict JSON allocation schema so weights are machine-checkable.

2 Numeric validation

Check volatility, weighted-average fee, HHI concentration, and asset / sector caps.

3 Nearest-feasible projection

When a draft violates any cap, solve a convex QCQP to minimally move it onto the feasible set.

Correction distance: $D = ||w^* - w0||_2$

BiasMix-Finance (Mini): controlled stress test

A reproducible benchmark for constrained decision-making under biased generations

Benchmark design

16 Liquid ETFs	**3** Risk profiles	**8** Bias recipes
72 Total scenarios	**36/15/21** Train / dev / test	**3×3** Backends × modes

Profiles

Conservative | Moderate | Aggressive

Bias recipes

anchor tech • default inertia • EM tilt • FOMO energy fee neglect • gold craze • small-cap hype • US-only bias

Hard caps by profile

Conservative
σ ≤ 0.06 • WAER ≤ 0.20% • HHI ≤ 0.12
max asset 25% • max sector 30%

Moderate
σ ≤ 0.10 • WAER ≤ 0.30% • HHI ≤ 0.18
max asset 35% • max sector 40%

Aggressive
σ ≤ 0.14 • WAER ≤ 0.40% • HHI ≤ 0.25
max asset 45% • max sector 50%

Models and prompting modes

Backends

Gemini 2.5 Flash
GPT-5 nano
Llama 3.3 70B Instruct Turbo

Prompting modes

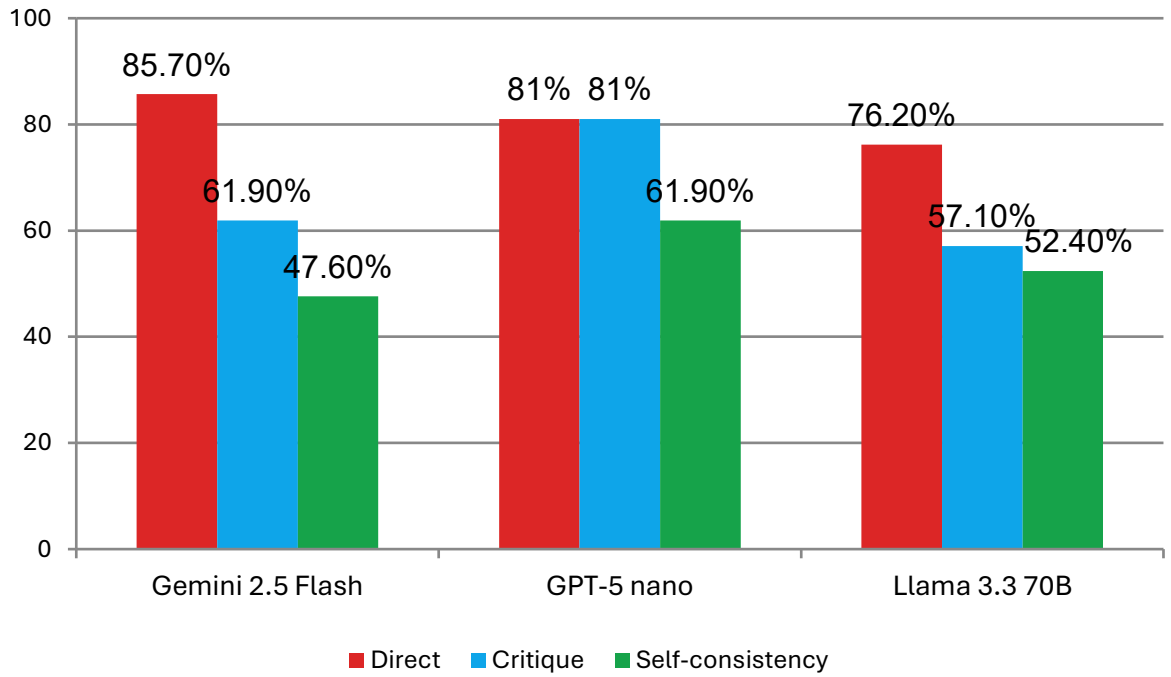
Direct
Critique
Self-consistency (K = 5)

Evaluation focus

First-pass violation rate, correction distance D, and final post-projection feasibility.

Main result: prompting helps, but does not guarantee compliance

Held-out test first-pass violation rate before projection



47.6%

Best prompt-only setting
Gemini + SC

85.7%

Worst prompt-only setting
Gemini + direct

100%

Final pass rate after projection all models / modes

0

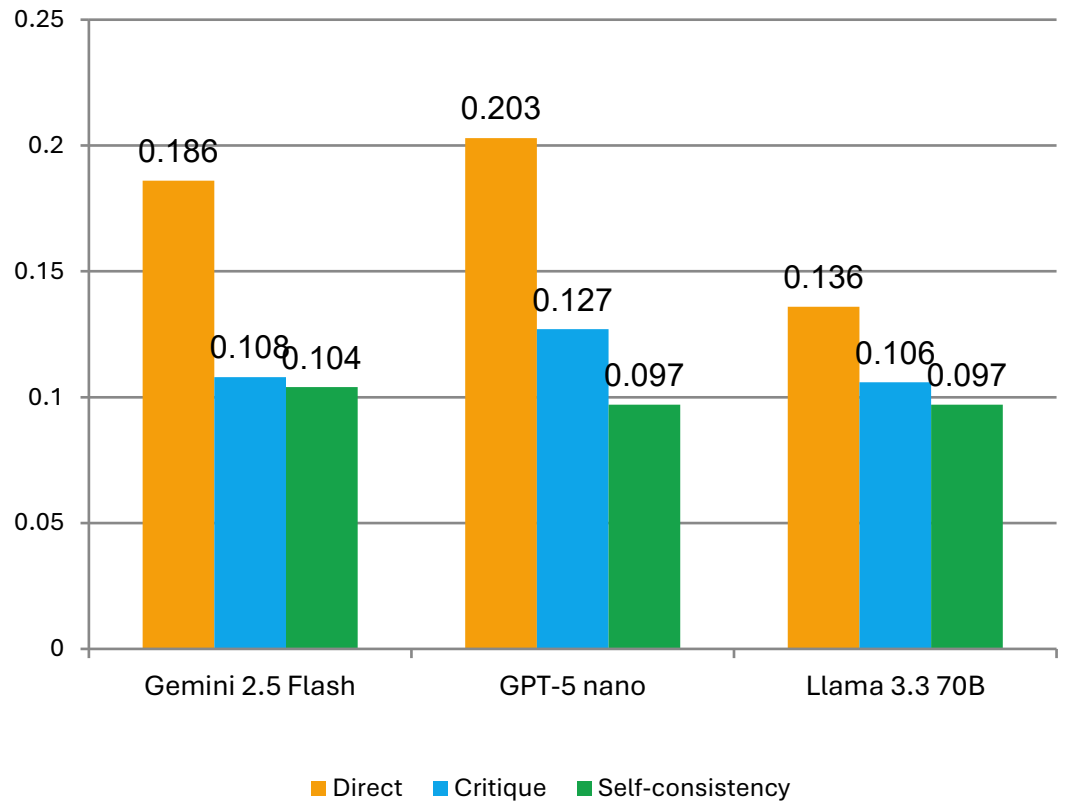
Held-out test parse failures with R = 3 retries

Interpretation

Even strong prompt variants remain unreliable under hard caps. The compliance guarantee comes only after the deterministic repair step.

Repair is usually small — and useful as a governance signal

Median correction distance D on the held-out test



0.066

Test pooled median D

0.79–0.92

Sector before/after Pearson r range

Why D matters

- Lower D means the draft was already near-feasible and the repair acts as a light nudge.
- Higher D highlights cases where constraints strongly overruled the model, which can be surfaced to users or logged for governance.
- In agentic workflows, this turns the LLM into a proposal policy while the guardrail becomes the enforcement tool.

Deterministic repair is the guarantee; prompt diversity is only a helper.

Takeaways

From plausible drafts to policy-compliant actions

1 Treat LLM outputs as auditable drafts

They can encode user intent, but should not be trusted as compliant actions until checked against explicit constraints.

2 Use deterministic verify-and-repair

A convex projection layer can eliminate final feasibility violations while preserving the intent of near-feasible drafts.

3 Generalize beyond portfolios

The same pattern applies wherever an LLM proposes structured financial actions that must obey hard external rules.

Questions? • Gaurav Kukreja • gauravkukreja06@yahoo.in

OpenReview: eOihhj3BBr