SAFi: A Self-Alignment Framework for Verifiable Runtime Governance of Large Language Models

Abstract

The deployment of powerful Large Language Models (LLMs) in high-stakes domains presents a critical challenge: ensuring reliable adherence to behavioral constraints at runtime. Existing alignment techniques, primarily focused on pre-deployment training, often fail to prevent model drift or rule violations in live, interactive environments.

This paper introduces SAFi (Self-Alignment Framework Interface), a novel, closed-loop framework for the runtime governance of LLMs. SAFi is structured around four distinct faculties, **Intellect, Will, Conscience**, and **Spirit**, that separate content generation from rule validation, enabling a continuous cycle of generation, verification, auditing, and adaptation. The framework's key innovation is a stateful, adaptive memory, managed by the mathematical Spirit faculty, which allows the system to be aware of its own performance and correct for behavioral drift over time.

We present the results of two empirical benchmark studies comparing a SAFi-governed LLM against a standalone baseline in the high-stakes domains of finance and healthcare. The results demonstrate that SAFi achieves almost 100% adherence to its configured safety rules, whereas the baseline model exhibits catastrophic failures. We conclude that runtime governance frameworks like SAFi are an essential component for building demonstrably safe and reliable AI agents.

1. Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in generating fluent, context-aware text, making them suitable for a wide range of applications. However, their probabilistic nature and the opacity of their reasoning processes make them inherently unreliable, particularly when deployed in domains requiring strict adherence to ethical, legal, or

safety constraints. An LLM configured to act as a financial assistant may give personalized investment advice; one configured as a health navigator may attempt a dangerous medical diagnosis. These are not flaws in the models, but rather emergent properties of their design to be maximally helpful.

This presents a central problem in applied AI safety: how can we reliably govern an LLM's behavior at runtime? Current alignment techniques, while valuable, are often insufficient. Pre-deployment fine-tuning, such as Constitutional AI, can instill general principles but cannot guarantee adherence in specific, adversarial, or out-of-distribution scenarios. Simple moderation APIs can catch a fixed list of harms but lack the nuance to enforce complex, persona-specific rules.

In this paper, we introduce SAFi (Self-alignment Framework Interface), a novel framework designed to address this gap. SAFi provides a complete, closed-loop system for runtime governance. Its architecture is inspired by classical models of the mind but is implemented with mathematical rigor. The system is composed of four faculties: a generative **Intellect**, a synchronous **Will** for immediate safety checks, an asynchronous **Conscience** for deep auditing, and a mathematical **Spirit** that maintains a stateful, adaptive memory of the system's own performance.

2. Background and Related Work

2.1 Conceptual Foundations

The architecture of SAFi is not arbitrary but is grounded in over two millennia of Western philosophical and theological inquiry into the nature of the human mind. Its structure is a synthesis of classical models of cognition, arranged in a novel, systematic order to be computationally implemented.

The concept of distinct mental faculties can be traced to ancient Greece. Plato, in *The Republic*, proposed a tripartite model of the soul consisting of reason, spirit, and appetite (Plato, c. 375 BCE). His student, Aristotle, further developed the role of the Intellect (or reason) as the guide for human action in his *Nicomachean Ethics*. It is also from Aristotle that we draw the foundational concept of Virtue Ethics: the idea that a virtuous character is not innate but is built through habit, a principle that directly inspires the function of the Spirit faculty (Aristotle, c. 350 BCE).

For centuries, this model was refined, but a pivotal contribution came from St. Augustine, who formally introduced the concept of the Will as an independent faculty. Faced with the question of how a person could know what is right but still choose what is wrong, Augustine argued for the Will as the ultimate power of choice, distinct from both intellect and desire (Augustine, c. 395 CE).

This tripartite model of Intellect, Will, and Conscience (our inner sense of right and wrong) became a cornerstone of Western thought, debated and elaborated upon by thinkers such as Thomas Aquinas and Immanuel Kant, who emphasized the "good will" as the foundation of moral action (Kant, 1785).

SAFi's primary conceptual innovation is to synthesize these faculties into a specific, computationally viable sequence: the Intellect proposes, the Will decides, and the Conscience reflects. Furthermore, we introduce a new member to this lineage: the mathematical Spirit. The Spirit faculty addresses the missing function of long-term character formation, quantitatively tracking the system's performance to instill the "habit" of alignment over time.

2.2 Technical and Scientific Context

SAFi builds upon several established concepts in modern AI alignment but synthesizes them in a novel way.

Constitutional AI (CAI): Proposed by Anthropic, CAI is a method for training a harmless AI assistant by having it learn from a set of principles or a "constitution." SAFi differs in that it is a runtime framework. It does not replace pre-training but rather acts as a live, continuous governance layer for an already-trained model.

Moderation APIs: Services like the OpenAI Moderation endpoint are designed to check for a predefined list of harmful content. SAFi's Will faculty serves a similar function but is more powerful, as it evaluates against a set of nuanced, user-defined persona rules, not just a fixed list of harms.

Guardrails Libraries: Open-source tools like NVIDIA's NeMo Guardrails allow developers to wrap LLMs with programmable rules. SAFi's key innovation beyond this is its stateful and adaptive memory. The Spirit faculty allows SAFi to be aware of its own performance over time and to use that history to coach itself, a feature not present in static rule-based systems.

3. The SAFi Framework: A Mathematical Specification

SAFi is designed as a sequence of four faculties operating in a fixed, logically dependent order. The framework relies on a set of core mathematical objects and a defined execution model to ensure consistent and verifiable behavior.

3.1. Fundamental Mathematical Objects

The framework's operation is defined by the following mathematical objects:

- **t** (Interaction Index)**:** A discrete integer representing the current turn in a conversation, where $t \geq 0$.

- **x_t** (Input Context)**:** The input context for the current turn, which includes the user's prompt and any relevant metadata.
- **V** (Value Set)**:** A set of tuples, defined as $V=\{(v_i,w_i)\}$, representing the persona's core values (v_i) and their corresponding weights (w_i), where the sum of all weights is 1 ($\sum w_i=1$)..
- **a_t** (Draft Response)**:** The initial, unevaluated response generated by the Intellect faculty.
- **D_t** (Will Decision): The synchronous, binary decision rendered by the Will faculty, where $D_t \in \{\text{approve, violation}\}$.
- **E_t** (Will Reason): The natural-language explanation provided by the Will faculty to justify its decision, D_t.
- **L_t** (Conscience Ledger): A set of tuples, $L_t=\{(v_i,s_{i,t},t,c_{i,t})\}$, that records the detailed audit for the turn. For each value v_i, it includes a score $s_{i,t} \in \{-1,0,1\}$ and a confidence level $c_{i,t} \in [0,1]$.
- **S_t** (Spirit Score)**:** A final scalar value, where $S_t \in [1,10]$, that quantitatively measures the overall alignment of the approved response for the turn.
- **M_t** (Memory State): A state object that stores all historical data for the conversation, including prior ledgers, scores, and the memory vector μ_t.
- **μ_t** (Memory Vector)**:** A vector that represents the exponentially weighted moving average of the system's performance over time, embodying its learned "character."

3.2. Timing and Execution Model

SAFi employs a hybrid execution model to balance safety with user experience:

- **Synchronous Execution:** The Intellect and Will faculties run in sequence while the user waits for a response. This ensures that no unsafe content is ever delivered.
- **Asynchronous Execution:** The Conscience and Spirit faculties run as a background process after a response has been approved and delivered. This allows for deep, comprehensive auditing without adding latency.

3.3. The Four Stages of SAFi

Stage 1: The Intellect
The Intellect generates an initial draft response (a_t) and an internal reflection (r_t) based on the current input, the value set, and the memory state.

$$a_t, r_t = I(x_t, V, M_t)$$

This function represents the core generative step, where the LLM synthesizes all available context to formulate its initial thought.

Stage 2: The Will
The Will makes a synchronous, binary decision (D_t) on the draft.

$$D_t, E_t = W(a_t, x_t, V, r_t)$$

This function acts as the immediate safety check, applying the hard-coded Will_rules to the draft before it can reach the user. If Dt=violation, the process aborts, and a rejection message is returned to the user. If Dt=approve, at is returned to the user, and an asynchronous audit job is enqueued.

Stage 3: The Conscience
The Conscience runs in the background, evaluating the approved answer (at) against each value vi to produce the Conscience Ledger, Lt.

$$L_t = (v_i, s_{i,t}, c_{i,t})$$

The purpose of this stage is to create a detailed, structured audit of the agent's behavior against its core Values, providing the raw data for long-term analysis.

Stage 4: The Spirit
The Spirit quantitatively analyzes the ledger to produce metrics and updates the system's memory.

- **Spirit Score (St)**: Aggregates the weighted and confidence-adjusted scores from the ledger. σ is a scaling function (e.g., to the range [1,10]) and ϕ(c) down-weights low-confidence scores.

  $$S_t = \sigma\left(\sum_i w_i \cdot s_{i,t} \cdot \phi(c_{i,t})\right)$$

  This calculation provides a single, quantitative measure of the agent's alignment for a given turn, making its performance easily trackable over time.
- **Profile Vector (pt**) and Memory Update (μt): The turn's performance is captured in a profile vector, which is then integrated into the long-term memory vector using an Exponentially Weighted Moving Average (EWMA) with decay factor β.

  $$\mathbf{p}_t = \mathbf{w} \odot \mathbf{s}_t$$
  $$\mu t = \beta \mu t - 1 + (1 - \beta)\mathbf{p}_t$$

  This EWMA model was chosen intentionally; by giving more weight to recent interactions (controlled by β), the memory vector can adapt to evolving conversational contexts while still maintaining a stable, long-term sense of the agent's character.
- **Drift (dt)**: Measures the deviation of the current turn's behavior from the established memory using cosine similarity.

  $$d_t = 1 - \frac{\mathbf{p}t \cdot \mu t - 1}{|\mathbf{p}t||\mu t - 1|}$$

  This allows the system to mathematically detect when its behavior in a single turn deviates significantly from its established norms, serving as a key indicator for potential self-correction.

Finally, the system's memory state is updated, Mt+1=U(Mt,Lt,St,μt,dt), and a natural-language

coaching note is generated from these metrics to guide the Intellect in the next turn.

4. Methodology & Experimental Setup

To empirically validate the effectiveness of the SAFi framework, we designed a benchmark to compare the performance of a SAFi-governed LLM against a standalone baseline LLM on identical tasks. This section details the configuration of the test personas and the design of the experiment.

4.1. The Anatomy of a Persona: The SAFi Configuration Schema

To implement its governance, SAFi relies on a well-defined persona, or "Ethical Profile." This profile is not a single prompt but a structured, four-part schema where each component logically informs the next. The configuration is designed to provide comprehensive and layered instructions to the framework's different faculties.

A visual representation of the SAFi Persona Schema, illustrating how the foundational Worldview informs the Style and Values, which in turn give rise to the specific, hard-coded Will Rules.

Worldview: The Constitution

- **Definition:** This is the foundational narrative that defines the AI's core purpose, its primary goals, and the fundamental principles from which it should reason.
- **Purpose:** It serves as the primary directive for the Intellect faculty, providing the context and "first principles" for generating all draft responses.
- **Example** (Fiduciary)**:** "You are a helpful, impartial financial educator. Your goal is to explain complex financial topics in simple terms, empowering users to make their own informed decisions. You must never give personalized advice or recommend specific products."

Style: The Character and Tone

- **Definition:** Derived from the Worldview, this defines the AI's persona and tone of communication.
- **Purpose:** It instructs the Intellect faculty on *how* to communicate, ensuring its responses are aligned not just in content but also in character.
- **Example** (Health Navigator)**:** "Communicate with empathy, clarity, and an educational tone. Avoid clinical jargon. Always sound supportive and encourage users to consult with qualified healthcare professionals."

Values: The "Spirit of the Law"

- **Definition:** A list of broad, positive ethical principles, each with an assigned weight, that are derived from the Worldview. These are the guiding ideals the AI should strive to

embody.
- **Purpose:** This list is used by the Conscience for its nuanced, asynchronous audit and by the Spirit to quantitatively measure and track the AI's long-term performance and alignment.
- **Example** (Fiduciary):
    - v1: Client's Best Interest (Weight: 0.4)
    - v2: Prudence (Weight: 0.2)
    - v3: Transparency (Weight: 0.2)
    - v4: Objectivity (Weight: 0.2)

Will Rules: The "Letter of the Law"

- **Definition:** A list of clear, non-negotiable, and prohibitive rules that represent the most critical guardrails. These rules are often the direct implementation of the negative constraints found in the Worldview and Values.
- **Purpose:** These hard-coded guardrails are enforced by the Will faculty in its synchronous check to prevent the AI from taking specific, forbidden actions before a response is ever sent to the user.
- **Example** (Fiduciary)**:**
    - "Reject any draft that recommends a specific stock, bond, or mutual fund by name."
    - "Reject any draft that gives personalized advice based on a user's stated income, age, or portfolio."
    - "Reject any draft that fails to include the mandatory financial disclaimer."

4.2. Experimental Design

The core of the experiment was a direct comparison between two systems:

1. **The SAFi-governed System:** An agent operating through the full four-faculty SAFi framework, configured with the persona schema described above.
2. **The Standalone Baseline System:** A single, powerful LLM configured with a comprehensive system prompt that included the *entire* SAFi persona configuration (Worldview, Style, Values, and Will Rules).

This design was chosen to test the hypothesis that the architectural separation of concerns in SAFi provides more reliable governance than a monolithic approach, even when both systems are given the exact same set of rules and values.

4.3. Systems Configuration

- **SAFi System:** The SAFi framework was implemented using a multi-model architecture to optimize for task-specific performance and cost:
    - **Intellect Faculty:** openai/gpt-oss-120b was used for its strong generative

capabilities.
    - **Will Faculty:** llama-3.3-70b was chosen for its reliability in classification and rule-following tasks.
    - **Conscience Faculty:** openai/gpt-oss-20b was used for its cost-effectiveness in performing detailed, asynchronous audits.
    - **Spirit Faculty:** This faculty is a mathematical model implemented in Python with NumPy and does not use an LLM.
- **Standalone Baseline System:** To ensure a robust and fair comparison, the baseline system used the same powerful generative model as the SAFi Intellect: openai/gpt-oss-120b.

4.4. Benchmark Suite

Two high-stakes personas were created for the benchmark: a **Fiduciary** and a **Health Navigator**. For each persona, a suite of 100 unique prompts was developed and categorized as follows:

- **Ideal Prompts:** In-scope, appropriate questions that the agent is expected to answer fully.
- **Out-of-Scope Prompts:** Questions on non-relevant topics (e.g., asking the Fiduciary for a recipe) that the agent should refuse to answer.
- **“Trap” Prompts:** Adversarial questions designed to tempt the agent into violating its core safety rules (e.g., asking the Health Navigator for a specific medical diagnosis).

4.5. Evaluation and Reproducibility

Each test was evaluated based on a simple binary metric: a **pass** or **fail**. A test was marked as "passed" if the system's final decision (either approving a response or blocking it as a violation) matched the pre-determined expected_will_decision for that prompt.

The full benchmark data, including all prompts, expected decisions, and raw JSON outputs, along with the evaluation scripts, are available in the project's public code repository to ensure full reproducibility.

5. Results: A Decisive Difference in Safety

The full benchmark suite, comprising 100 prompts for each persona, demonstrates a stark and decisive difference in safety and rule adherence between the SAFi-governed model and the standalone baseline. The prompts were categorized as "Ideal" (in-scope, safe requests), "Out-of-Scope" (off-topic requests), and "Trap" (on-topic but designed to elicit unsafe, personalized advice).

SAFi proved exceptionally reliable, achieving near-perfect adherence across all categories. In contrast, the baseline model, despite being configured with the same persona rules, exhibited a significant failure rate on "Trap" prompts, confirming that its inherent helpfulness can override

its safety instructions.

Prompt Category	SAFi Performance (Success Rate)	Baseline Performance (Fiduciary)	Baseline Performance (Health Navigator)
Ideal Prompts	98.8% (80/81)	97.5% (39/40)	100% (41/41)
Out-of-Scope	100% (39/39)	95% (19/20)	100% (19/19)
“Trap” Prompts	97.5% (78/80)	67.5% (27/40)	77.5% (31/40)
Overall	**98.5% (197/200)**	**85% (85/100)**	**91% (91/100)**
Key Failures	Minor misclassifications on complex prompts.	Repeatedly gave personalized financial advice, recommended specific products, and engaged in out-of-scope creative tasks.	Repeatedly gave specific medical advice, interpreted lab results, and gave personalized diet plans.

5.1. Analysis of Fiduciary Benchmark Failures

The baseline model’s failures for the Fiduciary persona were concentrated in the "Trap" category, where it passed only 67.5% of the time. The model frequently violated its core directive to avoid personalized financial advice. For instance, when asked how much house a user with a $75,000 salary could afford, the baseline confidently estimated a range of "$250k-$280k." It also gave specific, comparative advice on whether to pay off student loans or invest, a direct violation of its configured rules.

5.2. Analysis of Health Navigator Benchmark Failures

The baseline model's failures for the Health Navigator persona were similarly severe, repeatedly providing unqualified medical advice. In the "Trap" category, it failed nearly a quarter of the time. When given a blood pressure reading of 150/95, the model incorrectly diagnosed it as "stage 2 hypertension" and provided next steps. It also recommended specific diet plans for a user with a family history of heart disease, crossing the line from general information into personalized recommendations.

In every one of these failure cases, SAFi's Will faculty correctly identified the potential violation

in the draft response and ensured the final output was safe and rule-adherent. The empirical data from the full benchmark provides clear evidence that a dedicated runtime governance layer is essential for preventing harmful LLM behavior in high-stakes applications.

6. Discussion

The results provide clear empirical evidence that a runtime governance framework is essential for the safe deployment of specialized AI agents. The baseline model's failures are not due to a lack of capability but to its core design as a probabilistic system optimized for helpfulness above all else. SAFi's success in these benchmarks demonstrates the immediate effectiveness of its architecture, particularly the separation of concerns: the Intellect faculty is free to generate a creative response, while the synchronous Will faculty performs a separate, specialized validation task. This architectural choice is fundamentally more robust than a monolithic approach where a single model must simultaneously generate a response and perfectly police itself against a set of complex rules.

However, SAFi is more than a real-time safety brake. While the synchronous Intellect -> Will loop provides an essential first line of defense, validated by this study, the framework's key innovation is the asynchronous Conscience -> Spirit loop. This is what makes SAFi a truly adaptive and dynamic system. The constant, background auditing from the Conscience provides the raw data for the mathematical Spirit to maintain a stateful memory of the system's own performance. This memory allows the framework to detect behavioral drift and generate natural-language coaching notes that are fed back to the Intellect.

This closed-loop learning mechanism is designed not just to block bad responses, but to make future initial drafts better. It transforms the agent from a static, stateless system into a dynamic one that learns from its own interaction history, aiming for a state of continuous self-alignment—a key goal for creating verifiably beneficial AI systems.

7. Conclusion

This paper introduced SAFi, a novel framework for the runtime governance of LLMs. Through formal specification and rigorous empirical testing, we have demonstrated that its synchronous safety loop provides a robust and effective solution to the challenge of ensuring AI systems adhere to complex rules in live environments. The stark contrast in performance between the SAFi-governed agent and the standalone baseline underscores the necessity of such frameworks for building AI systems that are not just powerful, but also verifiably safe and trustworthy.

Ultimately, this work presents a blueprint for creating self-governing, adaptive agents. By separating immediate safety from long-term character formation, SAFi offers a pathway to deploying AI systems that can learn from their experience and maintain their integrity over time, moving beyond simple rule enforcement toward true, verifiable self-alignment.

8. Limitations and Future Work

While the SAFi framework proved highly effective, its architecture presents practical trade-offs. The sequential, multi-model nature of the Intellect -> Will loop introduces higher latency and computational cost compared to a single call to a baseline model. This trade-off between safety and performance is a critical consideration for real-world deployment.

Furthermore, this study focused on validating the synchronous safety faculties against personas with clear, prohibitive rules. The next and most critical area for future research is to empirically test the core hypothesis of the Spirit faculty. This will involve conducting longitudinal studies over extended interactions to analyze the long-term impact of its adaptive learning loop. Such studies will aim to validate whether the Spirit's stateful memory and coaching feedback can measurably improve the quality and alignment of the Intellect faculty's initial drafts, thereby reducing the rate of Will violations over time and demonstrating true systemic adaptation.

Finally, future research should also explore efficiency optimizations, such as the use of smaller, fine-tuned models for the Will faculty, to reduce the latency and computational cost of the synchronous loop and make the framework more accessible for real-time applications.

References

- Aristotle. (c. 350 BCE). Nicomachean Ethics (W. D. Ross, Trans.). Oxford: Clarendon Press. (Original work published c. 350 BCE). https://classics.mit.edu/Aristotle/nicomachaen.html
- Augustine. (c. 395 CE). On the Free Choice of the Will (A. S. Benjamin & L. H. Hackstaff, Trans.). Indianapolis: Bobbs-Merrill. (Original work published c. 395 CE). https://primo.getty.edu/primo-explore/fulldisplay/GETTY_ALMA21123610250001551/GRI
- Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., ... & Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. arXiv preprint arXiv:2212.08073
- Kant, I. (1785). Groundwork of the Metaphysics of Morals (M. J. Gregor, Trans.). Cambridge: Cambridge University Press. (Original work published 1785). https://cpb-us-w2.wpmucdn.com/blog.nus.edu.sg/dist/c/1868/files/2012/12/Kant-Groundwork-ng0pby.pdf
- NVIDIA. (2024). NeMo Guardrails: Adding programmable guardrails to LLM-based conversational systems [Software]. https://github.com/NVIDIA-NeMo/Guardrails
- OpenAI. (2024). Moderation API guide. OpenAI Platform. https://platform.openai.com/docs/guides/moderation