BA Overview

1. BeAxa (BA) is a benevolent AI moderation service designed to monitor, tag, and guide online discussions using intelligent, context-aware behavior. It does not censor opinions but flags uncivil, unethical, or manipulative content using transparent in-line ScripTags and assigns a reputation indicator (RepTag) to user avatars.

1.2 How BeAxa Works: BeAxa sends a series of text chunks to a curated set of disclosed LLMs, each of which rates the content against a list of toxicity, misinformation, and ethics violations. BeAxa then translates these ratings into inline ScripTags and avatar RepTags. The process becomes iterative as users edit their statements based on feedback.

1.3 Commercial Standalone: BeAxa may operate as a standalone product (e.g. BeAxa.com) or be integrated into other platforms, like our client, not-for-profit GroupBuild at GroupBuild.org. The initial release will be a free-to-use demonstration intended to attract partners, contributors, and platform integrators. BeAxa.com may eventually evolve into a subscription model, while integrated deployments (like GroupBuild) may operate under licensing agreements or shared governance contracts, depending on deployment context.

2. Core Functionality

- Detects toxic, misleading, or unethical statements
- Issues ScripTags to specific content
- Updates user RepTags based on behavior
- Provides real-time moderation summaries
- Trains itself continuously on moderated discussions
- Functions autonomously without human-assigned tags

3. Tagging Systems

BeAxa uses two distinct tagging systems:

3.1 ScripTags (Inline Content Tags): ScripTags are visual, inline [SeverityIndicator+icon+text] inserts that symbolize award-worthy qualities like truthfulness, logic, respectfulness. When a users phrase or statement violates one of these ideals, BeAxa shows the corresponding ScripTag after the user's text, or after the user's cursor if they are still editing their text.

Examples:

- +1Truth Green icon and "+1" shows one verified Truth found in user's content segment
- -3~~Truth~~ Red icon, strikethrough, and "-3" shows mild-severity Truth violation(s) in user's content segment
- -5~~Logic~~ Icon, strikethrough and "-5" shows mid-severity logic flaw(s) in user's content segment
- -1~~Respect~~ Icon, strikethrough and "-1" shows mildly disrespectful tone in user's content segment

These tags visually indicate what virtue has been compromised and how severely (110 scale). This makes them easy to scan and understand without needing dense explanations. Definitions appear on hover. Green icons represent unflagged text (implicitly compliant).

3.2 RepTags (User Avatar Tags): RepTags are persistent indicators of user behavior over time. They reflect how frequently a user violates or upholds key principles identified by ScripTags. These tags are visible next to each user's avatar, like a string of medals, and they reflect cumulative behavior. Users cannot assign RepTags; only BeAxa can.

Examples:

- -5~~Truth~~ user is a mid-tier violator of truthfulness
- +3Logic user usually posts logical content
- +9Logic user nearly always posts logical content
- +1Respect user occasionally maintains respectful tone

A user with many past violations will show a RepTag icon for that violation category but with a strikethrough, reflecting that they are in the negative end of the RepTag's levels. This allows ethical behavior to be gamified as gain or loss rather than judgment.

4. Session Monitoring Interface

BeAxa operates as a passive overlay that analyzes all content:

- Inserts ScripTags inline for feedback
- Updates RepTags privately or publicly
- Supports modes: private view, shared moderation, exportable session log
- Interprets user statements across time and user context

5. Trust System (RepTag Architecture)

RepTags are cumulative, adaptive, and designed to gamify ethical behavior:

- Positive behavior improves RepTags
- Violations reduce RepTags
- All changes are reversible through time-weighted compliance

RepTags influence how much weight a users feedback is given during moderation events (like appeals or discussions). They can also guide AI prioritization in situations where conflicting inputs

require weighing.

6. Why BeAxa Is Technically Feasible

Skeptical developers may question feasibility. BeAxa addresses this:

- Uses multiple LLMs to reduce bias and increase consensus
- Runs multi-pass evaluations with consistent rubric scoring
- Combines moderation with context-aware memory
- Relies on adaptive tag dictionaries and AI-curated examples
- Integrates transparency for auditing AI tag decisions

This approach sidesteps the limitations of static rules and promotes continuous self-improvement. BeAxa avoids user-to-user tagging and relies solely on its own judgment, backed by transparent reasoning and audit trails.

7. RepTag and ScripTag Use Cases

- Workplace chat platforms
- Town hall meetings
- Public Q&A forums
- Debate stages
- Classroom interactions

ScripTags provide live moderation. RepTags guide long-term participant trust.

8. Implementation Challenges

Creating BeAxa requires overcoming current constraints in AI systems:

- AI cannot yet permanently follow cross-session behavioral boundaries (e.g., "never use a certain

word")

- Inline moderation requires fast, probabilistic consensus from multiple LLMs
- Real-time tagging must balance accuracy, fairness, and readability
- The interface must avoid performance bottlenecks while providing rich user feedback
- Hover-activated tag definitions and export tools must be optimized for diverse platforms
- Multi-party trust systems must balance privacy, neutrality, and institutional transparency
- ScripTag icons must balance clarity with meaning and be legible inline, even at smaller sizes
- Severity must be visually conveyed through iconography or notation

BeAxas approach involves incremental refinement, user-in-the-loop corrections, and public transparency around tag meaning and usage.

9. Implementation Goals

- Launch BeAxa.com prototype for public testing
- Integrate with GroupBuild as first adopter
- Develop open API for embedding in other platforms
- Publish ScripTag taxonomy and assignable iconography
- Provide real-time RepTag dashboards with summary metrics

10. Next Steps

- Finalize icon sets for ScripTags and RepTags
- Design traffic light-based tag visuals with color and severity
- Launch early-access version of BeAxa.com
- Run pilot tests with real-world user groups
- Create and distribute a BeAxa/GroupBuild integration proposal