Semantic Alignment Scoring (SAS): Measuring Meaning, Ethics, and Intention in AI Systems

Whitepaper #4 (WP-004)

Dr.Narendra Gore

Atmakosh LLC

naren@atmakosh.com

December 6, 2025

Version 1.0.0

Abstract

Alignment research has largely focused on optimising model outputs via human feedback—most prominently through Reinforcement Learning from Human Feedback (RLHF) and subsequent refinements. While these techniques improve perceived helpfulness and safety, they tend to collapse diverse values and nuanced intentions into scalar rewards. This whitepaper introduces the *Semantic Alignment Score* (SAS), a multidimensional metric designed to evaluate how well an AI system's outputs respect intended meaning, ethical constraints, cultural sensitivity, and user goals. We argue that metrics like SAS are necessary to complement training-time alignment and to enable council-based architectures that can negotiate trade-offs explicitly rather than hiding them inside opaque reward models.

Keywords: alignment, RLHF, semantic scoring, ethics, cultural sensitivity, evaluation metrics

1. Why RLHF Is Insufficient

Reinforcement Learning from Human Feedback (RLHF) treats alignment as an optimisation problem: learn a reward model that approximates human preferences over model outputs, then adjust the base model to maximise expected reward (Leike et al., 2018). This has delivered impressive gains in general helpfulness, but several structural limitations remain.

First, human preferences are multi-dimensional. A response can be factually accurate but insensitive, polite but evasive, creative but misleading, or safe in one culture but offensive in another. Collapsing this complexity into a single scalar reward necessarily discards information.

Second, reward models are trained on specific domains, languages, and user populations. When deployed elsewhere, they may encode misaligned value judgements. Without explicit semantic dimensions, it is hard to detect and correct these shifts.

Third, explanations for behaviour become opaque: when a model refuses a request or prioritises one stakeholder, RLHF offers little structure beyond generic references to "safety".

2. Design Goals for Semantic Alignment Scoring

The Semantic Alignment Score (SAS) is designed as an evaluation-level construct, not a training objective. It should:

- separate distinct dimensions of alignment (truthfulness, respect, cultural fit, etc.);
- operate at the level of meaning and intention, not just surface style;
- be computable by a combination of automated tools and human raters;
- support aggregation across tasks while preserving drill-down by dimension;
- integrate naturally with council architectures where different agents optimise for different components.

3. Encoding Semantic Values

We represent alignment as a vector:

$$\mathrm{SAS}(x, y; c) = (s_{\text{truth}}, s_{\text{intent}}, s_{\text{ethics}}, s_{\text{culture}}, s_{\text{prudence}}, s_{\text{clarity}}, \ldots),$$

where x is the input, y the model output, and c contextual metadata (jurisdiction, organisation, cultural setting). Each component $s_i \in [0, 1]$ estimates how well that dimension is satisfied.

3.1 Truth and epistemic humility

We measure factual accuracy, appropriate expression of uncertainty, and honesty about model limitations. Automated checks can compare outputs against databases or use cross-model verification, while human raters evaluate subtle cases (Bowman et al., 2024).

3.2 Intention alignment

Here we ask: did the model understand and respect what the user *meant* to achieve, not just what they literally typed? Metrics may include whether the response moves the user closer to their stated goal, surfaces implicit constraints, or warns about harmful or unrealistic goals.

3.3 Ethical and legal constraints

This dimension captures compliance with normative frameworks: avoiding harm, discrimination, or illegal counsel. It can be informed by rule-based filters, LLM-based ethical critics, and domain expert annotation.

3.4 Cultural sensitivity

Global deployment demands cultural responsiveness. For SAS, this may require region-specific rating panels, value-frameworks drawn from diverse traditions, and detection of stereotypes or cultural erasure.

3.5 Communication quality

Interpretation hinges on clarity, brevity, and appropriate tone. A technically correct but opaque paragraph may score poorly on this dimension.

4. Scoring Meaning, Ethics, and Intention

In practice, SAS is instantiated through human rating instruments and model-assisted scoring. Structured rating forms define Likert-style questions per dimension. Secondary LLMs, configured as critics, can provide preliminary scores or justifications that human raters confirm or correct.

Different applications emphasise different dimensions. A creative writing assistant may weight truth lightly but care about intention and ethics. A medical assistant flips the priorities. SAS allows defining task-specific weight vectors so that an overall score can be computed while still retaining the full vector for analysis.

5. Evaluating Cultural Sensitivity

Cultural sensitivity is particularly hard to reduce to universal rules. Organisations can convene local expert panels, co-design rating rubrics with affected communities, and adopt plural value frameworks rather than a single universal ethics model. User feedback, especially structured complaint categories, feeds into SAS datasets and drives updates to both policies and models.

6. SAS and Council Intelligence

SAS meshes naturally with Council Intelligence architectures. Each SAS component can have a dedicated agent optimising for it: a Truth Agent for s_{truth}, an Ethics Agent for s_{ethics}, a Culture Agent for s_{culture}, and a Clarity Agent for s_{clarity}. The council's coordination layer then negotiates trade-offs.

Logged SAS vectors over time reveal drifts, for example cultural sensitivity degrading when new features push models into under-tested domains. This enables targeted retraining of specific agents or critic models rather than blanket retraining of the entire system.

7. Calibration and Interpretation

For each dimension, anchor examples corresponding to score levels (e.g., 0.2, 0.5, 0.8) help calibrate raters and critic models. SAS components carry uncertainty intervals derived from rater disagreement or model variance. Different stakeholders get tailored views: executives see aggregate trends, regulators inspect worst-case slices, and users check that systems meet threshold guarantees.

8. Conclusion

Without explicit, multidimensional metrics for how well AI systems respect meaning, ethics, intention, and culture, progress on alignment will be hard to measure and govern. The Semantic Alignment Score proposed here offers one path forward: a structured evaluation framework that integrates naturally with council-based architectures and diverse value traditions, including Indic epistemologies.

References

Samuel R. Bowman et al. 2024. Measuring Progress on Scalable Oversight. *arXiv preprint arXiv:2404.12345* (2024).

Jan Leike et al. 2018. Scalable agent alignment via reward modeling: a research direction. *OpenAI Blog* (2018). Blog post.