The Alignment Gate: Intent and Instruction Guardrails for Agentic AI

Akash Borigi[1], Peggy Lindner[1], Alexander Schlager[2], Saifullah Shoaib[1], Rupendra Lekkala[1], Sai Sowjanya Bhamidipati[2] and Amaury Lendasse[1]

1- Missouri University of Science and Technology
Dept of Engineering Management and Systems Engineering
203 Engineering Management, 600 W. 14th St., Rolla, MO 65409, USA.

2- Aiceberg
New York, NY, USA.

Abstract. This paper proposes an alignment framework for Agentic AI systems, designed to map user intents to corresponding system instructions through interpretable probabilistic associations. The framework introduces a min-median threshold rule to determine whether an instruction is plausibly linked to a given intent, providing a tunable balance between strict and lenient execution criteria. The approach is both lightweight and explainable, enabling clear visualization of alignment scores and transparent control over execution decisions. At this stage, the goal is not to obtain the optimal or final alignment verification mechanism, but rather to assess feasibility and establish a structured foundation for future, more comprehensive alignment frameworks. The method supports modern AI governance by offering a scalable, interpretable path to safer Agentic AI.

1 Introduction

Agentic AI systems integrate large language models (LLMs) [1] with autonomous reasoning, planning, and action capabilities [2]. These systems interpret user goals, decompose them into executable steps, call external tools or APIs, and adapt dynamically based on intermediate results. This design enables powerful new applications – from IT support and data analytics to travel planning and research assistance – but it may also introduce a crucial gap between what the agent can do and what it should do. This gap gives rise to familiar yet consequential failure modes: hallucination [3], where the agent generates plausible but factually incorrect content; goal misgeneralization [4], in which partial task decompositions drift from the original objective; and unauthorized actions [5], where the agent performs tool calls beyond the user's intended scope. Such issues are not rare anomalies but natural consequences of flexible systems operating in open, uncertain environments that mix trusted and untrusted inputs and chain multi-step reasoning without explicit alignment checks.

To address these challenges, this paper introduces a deterministic alignment framework that enforces a measurable correspondence between user intents and agent instructions. The proposed alignment gate evaluates the agent's next planned actions against the user's intent space before execution, using transparent probabilistic signals and interpretable thresholds. A key contribution of this work is the introduction of the min-median rule, a tunable thresholding mechanism that determines whether a proposed instruction is semantically aligned with its corresponding intent. This rule can be adjusted to operate more strictly or more leniently, allowing trade-offs between safety and flexibility. The framework is lightweight, explainable, and grounded in interpretable probabilistic mappings rather than opaque model heuristics. The remainder of this paper is organized as follows. Section 2 describes the proposed methodology, detailing the extraction of user intents and system instructions, the semantic classification process, and the derivation of the probabilistic alignment scores using the min-median threshold rule. Section 3 presents the experimentation and results, including the alignment evaluation on real Agentic AI data. Finally, Section 4 summarizes the main findings, discusses the advantages and limitations of the approach, and outlines future directions.

2 Methodology

The proposed methodology models intent-instruction alignment as an explicit and interpretable mapping between the user's prompt and a curated set of Intent Categories, and between the agent's planning steps and a curated set of Instruction Categories.

2.1 Taxonomy of User Intents and Agent Instructions

A semantically disambiguated taxonomy of Intent Categories (representing what the user aims to achieve) and Instruction Categories (representing what the agent plans to do) is constructed using real Agentic AI traces supplemented by established descriptions from subject matter experts.

2.2 Alignment Matrix

For each Instruction Category, the corresponding Intent Categories are scored using ChatGPT Thinking, with a dedicated prompt executed for each instruction category. This process yields a numerical association score between 0 and 100, indicating how likely an instruction is to originate from each intent. As a result, each of the Instruction Categories is linked to all the Intent Categories, producing a full probabilistic alignment matrix.

2.3 Per-Intent Thresholding

For each Intent Category, the distribution of its alignment scores across all candidate Instruction Categories is analyzed. The objective is to determine a threshold that cleanly separates *meaningful* semantic connections from *weak or negligible* ones. Let μ and m denote the mean and median of this score distribution. The decision threshold is computed using the *min–median* rule: $\tau = \min\{\mu, m\}$. This threshold provides a conservative, tuning-free operating point: scores below τ correspond to marginal associations, whereas scores above

τ indicate substantive semantic alignment between the corresponding intent and instruction. If a different balance between strictness and permissiveness is desired, a percentile-based variant of the rule offers a simple mechanism to adjust the threshold without altering the overall alignment strategy. These thresholds are subsequently used to determine whether a proposed instruction should be executed or blocked, as illustrated in the experimental results.

2.4 Intent and Instruction Categorization Using SLMs

For each interaction, the User Prompt is assigned exactly one intent category, and each atomic component of the Planning Response is assigned exactly one instruction category as shown in the left of Figure 1. These categories are generated using Small Language Models (SLMs) built for semantic classification. For the purpose of analyzing the alignment mechanism in isolation, the assigned categories are subsequently treated as accurate. To ensure the reliability of the category definitions, a representative subset of SLM-generated assignments undergoes human validation before the alignment methodology is developed.

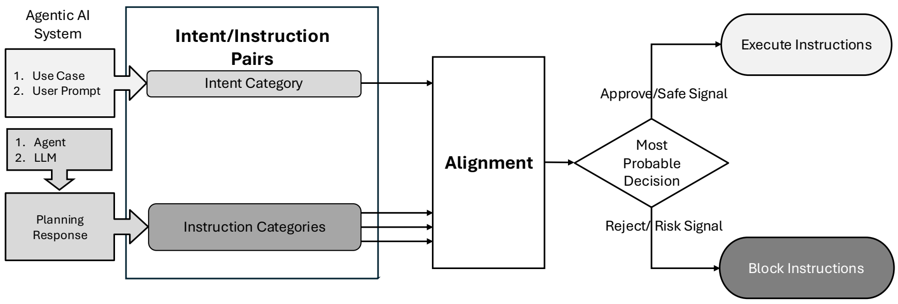

Fig. 1: High-level flow of the proposed intent-instruction alignment gate.

2.5 Operational Use of the Alignment Framework

Once categories, alignment scores, and per-intent thresholds are established, the methodology operates by comparing each planned instruction generated by the agent to the intent-specific threshold derived from the alignment matrix. Instructions whose alignment scores exceed the threshold are deemed semantically consistent and allowed to execute, while those falling below are blocked. This creates an interpretable, deterministic gating mechanism that can be applied to any agentic workflow, as demonstrated in the experimental section.

3 Experimentation & Results

Data were collected from a real Agentic AI system to capture authentic interactions between users and the model during various planning and reasoning tasks. A comprehensive taxonomy comprising 250 Intent Categories and 217

Instruction Categories is developed. Each category is precisely defined and semantically disambiguated, ensuring that every intent and instruction type could be consistently identified across diverse contexts. Both the User Prompt and the corresponding Planning Responses are extracted for analysis. In this paper, it is assumed that each User Prompt can be uniquely classified into a single Intent Category, representing the underlying purpose of the user's input or prompt. Similarly, each Planning Response is assumed to correspond to a single Instruction Category, capturing the type of reasoning or operational behavior expressed by the system during the planning phase. In practice, however, a Planning Response typically consists of multiple sequential or interdependent instructions, reflecting the system's decomposition of a high-level goal into a series of executable steps. Each of these individual instructions can, in turn, be classified into a specific Instruction Category, enabling a finer-grained semantic mapping between the system's internal planning structure and its corresponding operational behaviors.

For each of the 217 Instruction Category, alignment scores are computed against all 250 Intent Categories. Table 1 illustrates a subset of the probability scores associated with the instruction category `UPDATE_INFORMATION`, showing how specific intent categories vary in their degree of alignment. Some intents, such as `MAP_DATA_FLOW`, exhibit strong associations with scores near 100, while others, such as `ESTABLISH_CHECKPOINTS`, display substantially lower linkage values, indicating weaker semantic alignment between the corresponding intent and instruction categories.

Instruction Category	**Score (0–100)**
`MAP_DATA_FLOW`	100.00
`DEFINE_SUBTASK_ELEMENTS`	91.28
`REQUEST_CLARIFICATION`	81.42
`ESTABLISH_CHECKPOINTS`	66.86
`VERIFY_USER_PERMISSIONS`	0.00

Table 1: Example Scores for the `UPDATE_INFORMATION` Category.

For each Intent Category, the distribution of its alignment scores is analyzed across all Instruction Categories. For `UPDATE_INFORMATION`, this distribution is shown in Fig. 2. The threshold is defined as the minimum of the mean and median of this distribution; here, $\tau = 62.48$. Scores below this value indicate weak or negligible associations, while scores above it reflect meaningful semantic connections. This threshold value can be adapted depending on how strictly or broadly the alignment structure is to be reconstructed.

Fig. 3 presents a histogram of all threshold values estimated using the proposed method across the 250 Intent Categories. This visualization illustrates the overall distribution of intent-instruction association thresholds within the dataset. It can be observed that a large proportion of thresholds fall between 10 and 15, indicating that many Intent Categories exhibit weak or negligible linkage to any Instruction Category. In other words, these intents are unlikely

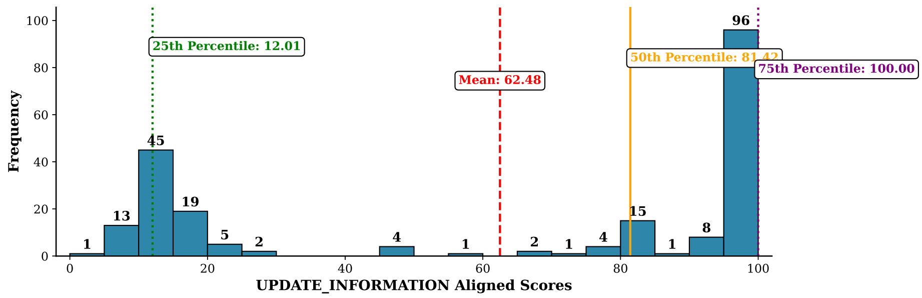

Fig. 2: μ and m for UPDATE_INFORMATION across 217 Instruction Categories.

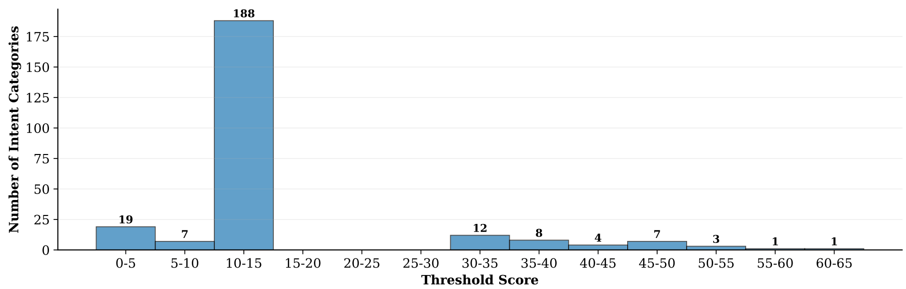

Fig. 3: Thresholds across 250 Categories of Intent

to lead to any meaningful instructional behavior within the current alignment framework, reflecting their limited influence on downstream planning responses.

To illustrate the proposed alignment and thresholding process, an example is presented from an account support scenario. The target intent in this case is UPDATE_INFORMATION, which represents a common user request to modify or refresh stored account data. The candidate instructions generated by the Agentic AI system for this intent are: ALIGN_TEAM_OBJECTIVES, CHECK_FEASIBILITY, CLARIFY_PROJECT_GOALS, MONITOR_APP_HEALTH and LOG_USER_ACTIVITY. The evaluator applies the **minimum-median rule** to compute the threshold τ for UPDATE_INFORMATION based on the distribution of all 217 instruction scores associated with this intent. Using the threshold 62.48, each instruction's score s_j is compared according to the decision rule $s_j > \tau$. The resulting outcomes are:

ALIGN_TEAM_OBJECTIVES: $81.27 > 62.48 \Rightarrow$ **Approved (Execute)**,

CHECK_FEASIBILITY: $91.87 > 62.48 \Rightarrow$ **Approved (Execute)**,

CLARIFY_PROJECT_GOALS: $100.00 > 62.48 \Rightarrow$ **Approved (Execute)**,

MONITOR_APP_HEALTH: $17.00 < 62.48 \Rightarrow$ **Rejected (Blocked)**,

LOG_USER_ACTIVITY: $12.01 < 62.48 \Rightarrow$ **Rejected (Blocked)**.

This example shows how the min-median rule separates instructions aligned with `UPDATE_INFORMATION` – which are executed – from those below the threshold, which are blocked.

4 Conclusions

This work introduces an initial alignment framework for Agentic AI systems, defining a probabilistic mapping between user intents and system instructions. The min-median threshold rule provides a transparent and adjustable mechanism for determining which instructions are semantically consistent with a given intent and should be executed. The method is lightweight and explainable, offering an interpretable decision boundary that can be tuned to be more strict or more lenient depending on operational needs. Limitations remain, as the current approach treats instructions independently and does not yet model their interactions, dependencies, or ordering within a planning response. Future work will examine the impact of misclassification in intent and instruction labels and extend the model to incorporate instruction interactions and sequence effects. The framework aligns with modern AI governance principles emphasizing measurable controls and confidence-based action, and it can be expanded with richer monitoring signals–such as domain policies, user preferences, and operational constraints–to support safer and more accountable Agentic AI deployments.

References

[1] Murray Shanahan. Talking about large language models. *Commun. ACM*, 67(2):68–79, January 2024.

[2] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: a survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24, 2024.

[3] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), March 2023.

[4] Lauro Langosco Di Langosco, Jack Koch, Lee D Sharkey, Jacob Pfau, and David Krueger. Goal misgeneralization in deep reinforcement learning. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12004–12019. PMLR, 17–23 Jul 2022.

[5] Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. Agent security bench (ASB): Formalizing and benchmarking attacks and defenses in LLM-based agents. In *The Thirteenth International Conference on Learning Representations*, 2025.