Behavioral Inference at Scale: The Fundamental Asymmetry Between Motivations and Belief Systems

Anonymous authors
Paper under double-blind review

Abstract

We establish empirical bounds on behavioral inference through controlled experiments at scale: LLM-based agents assigned one of 36 behavioral profiles (9 belief systems × 4 motivations) generate over 1.5 million behavioral sequences across 17,411 games in grid-world environments, providing ground truth unavailable in human behavioral studies. After model-specific filtering, BiLSTM classification experiments use 344,365 sequences from 4,064 games; Longformer experiments use 267,063 sequences from 3,531 games. Rather than asking whether inference has limits, we ask *how large those limits are, where they concentrate, and why*. A fundamental asymmetry emerges in both magnitude and structure. Motivations achieve 98–100% inference accuracy and recover 97% of available mutual information across all architectures. Belief systems plateau at 24% for LSTMs regardless of capacity, recovering only 30% of available information, a 3.3× asymmetry in information extraction efficiency. Transformer architectures with 9-stage curriculum learning reach 49% alignment accuracy, doubling LSTM performance and demonstrating that the recurrent ceiling is architectural rather than fundamental.

Yet even this improvement leaves belief systems correctly classified less than half the time, with per-alignment accuracy ranging from 1% (True Neutral) to 72% (Lawful Evil). Confusion analysis maps the failure structure precisely: a "neutral zone" of behavioral ambiguity extends beyond True Neutral to encompass Good alignments, where prosocial behavior is indistinguishable from rule-following or balance-keeping. Combined motivation and belief inference yields 17.6× improvement over random baseline for full 36-class profile classification, while establishing that the bottleneck is entirely located in belief system inference.

Signal enhancement and explanatory queries yield only marginal LSTM gains (+3.8%), confirming that the ceiling is information-theoretic rather than data-limited. These bounds have direct implications for any system relying on behavioral monitoring to infer agent values.

1 Introduction

Simulated agents with known behavioral profiles enable research to address a question that lacks ground truth for humans: given only observable actions, can we infer the internal states, i.e. beliefs and motivations, that generate those behaviors? Recent work validates LLM-based agents as methodological proxies for behavioral research: they accurately represent demographic subpopulations when prompted with relevant information (Argyle et al., 2023) and generate behavioral patterns suitable for systematic study (Horton, 2023). For our purposes, they offer a critical advantage, the consistent maintenance of predetermined profiles across thousands of decisions, enabling investigation of the limit of our ability to infer beliefs and motivations without human subject variability. Research shows that inference difficulty grows acute as classification taxonomies expand. Binary personality dimensions achieve 65-86% accuracy (Amirhosseini & Kazemian, 2020; Ryan et al., 2023), while 16-type systems fall to 40-55% (Ontoum & Chan, 2022). Beyond 20 categories, behavioral classification research becomes virtually nonexistent. In this paper we address this gap through controlled experiments with LLM-based agents executing over 1.5 million raw behavioral sequences across 17,411 games across 36 distinct behavioral profiles.

For this research we decompose agent behavior into two orthogonal components: belief systems and motivations. Belief systems have been characterized as normative structures in which evaluative categories, conceptions of "good" and "bad", exert a central organizing influence on reasoning and action (Usó-Doménech & Nescolarde-Selva, 2016). We operationalize this for simulated agents using the Dungeons & Dragons alignment taxonomy, a 3×3 grid mapping moral stance (good/neutral/evil) against rule adherence (lawful/neutral/chaotic). This system's (TSR Inc., 1989) 50-year refinement ensures the underlying concepts are well-represented in LLM training data. Motivational drives, following the perceptual-motivational distinction (Seay & Gottfried, 1978), determine goal prioritization. We define four: Wealth (resource maximization), Safety (risk minimization), Wanderlust (exploration maximization), and Speed (action minimization). The 36 profiles are created by combining the 9 alignments with the 4 motivations, following a previously established framework (Starace & Soule, 2026).

The decomposition into motivations and belief systems reveals not merely *that* a fundamental asymmetry exists (a result inverse reinforcement learning theory predicts (Ng & Russell, 2000)), but *how large it is, where it concentrates, and what structure it takes*.

Motivations achieve 98–100% inference accuracy across all architectures tested. Belief systems plateau at 24% for LSTMs regardless of model capacity: a 5.55M parameter GRU variant underperformed the 1.79M parameter Direct-36 BiLSTM variant, ruling out capacity as the binding constraint. Transformer architectures with 9-stage curriculum learning substantially improve on this ceiling, reaching 48.9% accuracy and demonstrating that the LSTM ceiling reflects architectural rather than fundamental limitations.

The mutual information asymmetry is 3.3×: motivation inference recovers 97% of available information ($I(Y;\hat{Y}) = 1.95$ of $H(Y) = 2.00$ bits), while alignment inference recovers only 30% ($I(Y;\hat{Y}) = 0.95$ of $H(Y) = 3.17$ bits). When we combine the Transformer's belief system inference with the BiLSTM's near-perfect motivation inference, the complete system achieves ∼49% accuracy on full 36-class profile prediction, a 17.6× improvement over random baseline, with the bottleneck located entirely in belief system inference. Critically, the failure is not uniformly distributed: Evil alignments achieve 60–72% accuracy while True Neutral approaches chance (1%), and the zone of behavioral ambiguity extends unexpectedly into Good alignments (18–60%). This structure, not merely the direction of the asymmetry, is the central empirical contribution.

The asymmetry between these components exposes an information-theoretic boundary, explained in section 4.4. Motivations manifest directly in behavioral statistics. For example, a Wealth-driven agent consistently prioritizes resources. In general, motivations produce statistically separable behavioral distributions with consistent directional signatures. Belief systems face inherent ambiguity: the same observable action stems from multiple internal states. Helping another agent could indicate altruism, calculated reciprocity, or strategic coalition-building. Without access to reasoning, these interpretations remain indistinguishable. Neutral alignments prove particularly opaque, and Good alignments show unexpected ambiguity, creating zones where behavioral monitoring fails to distinguish underlying values.

We validate the asymmetry through a three-phase experimental progression culminating in 17,411 raw games and over 1.5 million behavioral sequences. Early phases established architectural baselines and informed environment redesign; the final phase maximized belief-testing encounters and implemented curriculum learning. Neither increased behavioral data nor the inclusion of agent-generated questions as input features could overcome the LSTM ceiling, strongly suggesting that the limitation is architectural for recurrent models and fundamental for the task itself. Confusion matrix analysis reveals where inference fails: Evil alignments achieve 68% accuracy while Good alignments fall to 35%, and True Neutral approaches random chance at 1%, confirming that the ambiguity identified above concentrates in specific regions of the alignment space rather than distributing uniformly.

These findings establish empirical bounds on behavioral inference with immediate practical implications. Any system relying on behavioral monitoring, whether games modeling player types, adaptive systems interpreting user actions, or AI systems using reinforcement learning from human feedback, cannot reliably infer the belief systems that determine how agents interpret objectives. The gap between observable behavior and underlying values represents a fundamental constraint on observation-based approaches. Moving beyond this ceiling will require complementary methods that access agent reasoning directly, through interactive

dialogue, multi-agent dynamics, or other approaches that force latent values to manifest in ways that pure behavioral observation cannot capture.

2 Related Work

Prior work on behavioral classification reveals systematic accuracy degradation as taxonomies expand: 65-86% for binary classifications (Amirhosseini & Kazemian, 2020; Ryan et al., 2023), 70-79% for trinary (Denden et al., 2018), falling to 40-55% for 16-type Myers-Briggs frameworks (Ontoum & Chan, 2022). Beyond 20 categories, research becomes virtually nonexistent, but these studies share a common limitation: they treat behavioral profiles as monolithic categories rather than decomposing them into components with different inferential properties.

Sequential modeling addresses another limitation of traditional approaches. Aggregated behavioral metrics (action frequencies, resource totals, completion rates) discard temporal context that distinguishes agent strategies. Two agents may achieve identical outcomes through fundamentally different trajectories. Prior work (Cowley & Charles, 2016) demonstrates this directly with Behavlets, psychologically-grounded behavioral primitives that capture sequential patterns invisible to aggregation, achieving 35% better classification than raw gameplay metrics by mapping temperament types to observable action sequences. However, temperament models lack granularity where value systems interact with strategic goals: the distinction between assisting and exploiting another agent reveals alignment preferences that broad temperament categories cannot capture.

Social media personality prediction achieves 86-88% accuracy for MBTI dimensions (Christian et al., 2021) and 74-83% for 16-type MBTI classification (Ryan et al., 2023) using machine learning approaches. The contrast with behavioral inference is instructive: social media provides explanatory content where users describe thoughts, feelings, and reasoning. High accuracy relies on access to natural language explanations, not behavioral patterns alone.

The question we address also has a theoretical precursor in inverse reinforcement learning (IRL). Ng & Russell (2000) proved that observed behavior cannot uniquely determine an agent's reward function: infinitely many reward functions generate identical optimal policies, creating inherent degeneracy in inference from behavior. Subsequent work on reward function degeneracy and hacking (Skalse et al., 2022) and inferring human preferences from demonstrations (Hadfield-Menell et al., 2017) reinforces this theoretical constraint. Our contribution is empirical: where IRL theory establishes that inference degeneracy *exists*, we measure how large it is, which belief system categories it concentrates in, and whether architectural advances can narrow it. The 3.3× mutual information asymmetry between motivation and belief inference, and the per-alignment accuracy distribution ranging from 1% to 72%, provide the quantitative structure that theory predicts but cannot specify. This reframes our finding: the directional asymmetry between goal inference and value inference is expected; its magnitude, its internal structure, and the conditions under which it can be partially overcome are not.

3 Methodology

3.1 Experimental Design

We conducted a three-phase investigation using LLM-based agents (Llama 3.1-8B) assigned 36 distinct behavioral profiles combining 9 belief systems with 4 motivational drives. Phase 1 established baseline performance in 5×5 grid environments with 19,413 games (average 7.93 sequences per game). Phase 2 expanded to 10×10 grids with 6,212 games (average 21.80 sequences per game) to test whether increased decision density would improve classification. Phase 3 redesigned the environment to maximize belief-testing signal: value-testing encounters increased from 30% to 81% of grid cells, and agent-generated questions about the environment were captured as additional input features. This phase generated 17,411 raw games averaging 90.5 sequences per game, yielding 1,575,377 total behavioral sequences.

Games with status other than `complete` were excluded, removing all games where the agent exhausted available turns or entered a navigation loop. Additional filtering criteria are described in Section 3.6. After

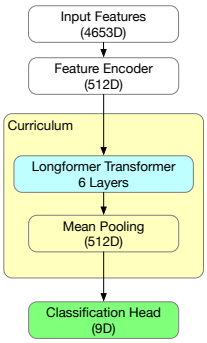

Figure 1: Player2Vec architecture with curriculum learning. Input features (4,653D) pass through a feature encoder to the curriculum block containing a 6-layer Longformer transformer with local attention, followed by mean pooling and a classification head.

filtering and profile balancing, the BiLSTM training dataset contains 344,365 sequences from 4,064 games; the Longformer training dataset contains 267,063 sequences from 3,531 games, reflecting an additional filter restricting sequences to active encounter and loot interactions. Both models use game-level splits of 70/15/15 with a fixed seed, preventing temporal leakage. BiLSTM splits are stratified on the full 36-class profile to ensure proportional representation of all alignment-motivation combinations; Longformer splits are stratified on the 9-class alignment only, with motivation distribution within splits left to chance.

3.2 Feature Representation

Each timestep combines 768-dimensional BGE embeddings across six text fields (room description, encounter description, action text, player question, loot description, state transition) with engineered features capturing temporal dynamics (sequence position, episode progress), spatial state (grid coordinates), and action statistics (cumulative counts by type). We additionally explored theory-driven features derived from Moral Foundations Theory (Haidt, 2007; Graham et al., 2013), computing keyword-based scores across six moral dimensions (care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, sanctity/degradation, liberty/oppression) for subsets of text fields. Ablation studies comparing configurations with and without MFT features are reported in Section 4.

3.3 Architecture

Our primary architecture adapts Longformer's local attention mechanism for behavioral sequence classification. The model comprises a 512-dimensional feature encoder, six transformer layers with 8-head local attention (window size 256), and masked mean pooling for sequence aggregation (Figure 1), yielding 21.4M parameters. During the alignment architecture search, we evaluated BiLSTM variants (2 layers $\times$ 512 hidden units, 7.7M parameters) and hierarchical approaches including a GRU router-specialist network (5.5M parameters). Increasing model size did not improve alignment inference performance: the 5.55M parameter GRU router-specialist variant achieved *lower* alignment accuracy than the 1.79M parameter Direct-36 BiLSTM variant, ruling out model capacity as the binding constraint and motivating our focus on architectural class rather than parameter count. Two-layer BiLSTMs matched or exceeded deeper variants, consistent with the interpretation that recurrent architectures face a representational ceiling on this task independent of depth. The final ensemble pairs the Longformer for alignment inference with a separate BiLSTM for motivation inference exclusively (128 hidden units per direction, 1.3M parameters). Complete architectural specifications for all variants appear in Appendix D.

Table 1: Curriculum learning progression from binary opposites to full 36-class profiles. Each stage inherits weights from the previous, enabling hierarchical representation building through staged complexity.

Stage	Classes
1a–d	Binary pairs (LG|CE, LE|CG, LN|CN, NG|NE)
2a–b	Corner quadrants, edge quadrants
3	All 8 non-center alignments
4	All 9 alignments (add TN)
5	Full 36 profiles

3.4 Curriculum Learning

Curriculum learning, training models on progressively harder examples, improves generalization by enabling hierarchical representation building (Soviany et al., 2022). The approach traces to early work (Elman, 1993) demonstrating that neural networks learning grammar benefit from "starting small" with simple structures before encountering full complexity. Recent work extends this principle to multi-class classification through coarse-to-fine strategies: models first learn to distinguish broad categories, then transfer this knowledge to fine-grained distinctions (Stretcu et al., 2021).

Our 9-stage curriculum exploits the natural hierarchy in D&D alignments. The moral axis (good/neutral/evil) and ethical axis (lawful/neutral/chaotic) create a 3×3 grid where corner alignments represent maximally distinct value combinations, while adjacent alignments share one dimension. Stages 1a–b train on corner opposites (LG|CE, LE|CG), pairs differing on both moral and ethical dimensions. Stages 1c–d train on edge opposite corners (LN|CN, NG|NE), pairs differing on one dimension while sharing the other. Stages 2a–b expand to corner and edge quadrants, introducing alignments sharing one axis value. Stage 3 combines all eight non-center alignments; Stage 4 adds True Neutral for full 9-class alignment classification. Stage 5 fine-tunes on the complete 36-class profile space combining alignments with motivations.

Each stage inherits weights from the previous, with early layers frozen during later stages to preserve learned representations. This prevents catastrophic forgetting of coarse distinctions while allowing the model to build increasingly fine-grained discrimination. The strategy mirrors human category learning, where broad categorical knowledge scaffolds acquisition of subtle within-category distinctions.

3.5 Training Protocol

The two ensemble components use separate training protocols. The Longformer (alignment inference) is trained with AdamW (lr $= 5 \times 10^{-5}$, weight decay 0.01), cosine scheduling with 5-epoch warmup (minimum lr$= 10^{-5}$), dropout $p = 0.3$, label smoothing 0.1, and gradient clipping at 1.0. The motivation BiLSTM is trained with Adam (lr $= 5 \times 10^{-4}$), ReduceLROnPlateau scheduling (factor 0.5, patience 3, minimum lr$= 10^{-6}$), and dropout $p = 0.35$. Both models use early stopping with stage-dependent patience (full details in Appendix B), sequences padded or truncated to 214 timesteps, and batch sizes of 32 (BiLSTM) and 16 (Longformer). Complete environment implementation, agent simulation protocols, and baseline specifications are detailed in prior work (Starace & Soule, 2026).

3.6 Behavioral Consistency Verification

A critical methodological question is whether LLM-based agents behave consistently with their assigned profiles, or whether profiles function only as nominal labels with limited behavioral effect. We address this through episode-level behavioral consistency scoring, which filters out episodes where agent behavior is inconsistent with the assigned profile before any classification experiments are conducted.

The behavioral consistency score (referred to as *rating* in prior work (Starace & Soule, 2026)) is grounded in the encounter and loot design of the experimental environment. Each encounter and loot prompt is generated with profile-specific content: keywords, framings, and action options are written to appeal to

particular alignment and motivation combinations. When an agent selects an action at an encounter, it receives a score reflecting how closely that action matches the option designed for its assigned profile. Full marks are awarded for selecting the profile-targeted action; partial or zero marks are awarded for other selections, weighted by their distance from the targeted option. These per-encounter scores are summed over a complete game and divided by the maximum possible score achievable had the agent selected the profile-targeted action at every encounter, yielding a normalized behavioral consistency score in $[0, 1]$.

Episodes scoring below 0.7 are excluded from all subsequent analysis, as are games with fewer than 25 sequences, ensuring a minimum behavioral trace sufficient for meaningful classification. The 0.7 threshold retains episodes where the agent behaves in accordance with its assigned profile on at least 70% of the available behavioral signal, while excluding episodes where the LLM appears to have drifted from or inconsistently maintained its assigned profile.

This filter serves two purposes beyond data quality. First, its grounding in profile-specific encounter design means that an agent scoring ≥ 0.7 has demonstrably responded to alignment- and motivation-relevant content in ways consistent with its assigned profile, not merely produced plausible text. This is a stronger consistency check than post-hoc behavioral coding because the scoring rubric is embedded in the environment design itself, with content authored specifically to elicit profile-consistent responses. Second, the filter addresses the specific concern that LLM safety alignment overrides might distort Evil-profile behavior. Episodes where the model's safety training causes it to resist or hedge Evil-profile responses would produce low consistency scores, as the agent would fail to select the profile-targeted actions at encounters designed for Evil alignments, and would be excluded before classification. The persistence of Evil's classification advantage after filtering therefore reflects genuine behavioral regularities in the episodes that pass the consistency threshold, rather than artifacts of safety-override noise in the raw data.

4 Results

4.1 Recurrent Architecture Ceiling

Baseline experiments without curriculum learning establish a hard performance ceiling for recurrent architectures. Across all MFT configurations tested, from no theory-driven features to full 6-field coverage, LSTM-based models plateau at 19-20% alignment accuracy without optimization; architectural refinements reach 24% (Appendix Table 8). This ceiling persists regardless of model capacity: our 5.55M parameter GRU variant performed worse than the 1.79M parameter Direct-36 BiLSTM variant. The consistency across configurations indicates that recurrent architectures fundamentally cannot capture the sequential patterns distinguishing belief systems, independent of input features.

Motivation inference tells a different story. Even baseline LSTMs achieve 98-100% accuracy classifying the four motivational drives. This asymmetry emerges immediately and persists across all experimental conditions: goal-oriented behaviors produce unambiguous statistical signatures, while belief systems remain obscured behind action sequences that admit multiple interpretations.

4.2 Curriculum Learning Breakthrough

Transformer architectures with 9-stage curriculum learning break through the LSTM ceiling, achieving 48.9% $\pm$ 1.9% alignment accuracy across five runs (Table 3). This represents a 2.5$\times$ improvement over baseline and demonstrates that the 19-20% ceiling reflects architectural limitations of recurrent models rather than fundamental task difficulty.

The curriculum strategy proves essential. Direct 9-class training with transformers fails to converge reliably. The staged progression (binary opposites, quadrants, full complexity) enables the model to build hierarchical representations that distinguish adjacent alignments sharing behavioral features. Without curriculum learning, transformer performance matches LSTM baselines.

Combined with near-perfect motivation inference (99.7%), the full system achieves $\sim$49% accuracy on 36-class profile classification, with belief system inference as the sole bottleneck. Against a random baseline of

Table 2: Accuracy progression through curriculum stages. Performance degrades gracefully as class complexity increases, with the sharpest drop when True Neutral enters at Stage 4.

Stage	Classes	Accuracy
1a–d	Binary pairs (4×2)	85–92%
2a–b	Quadrants (2×4)	78–84%
3	Non-center (8)	61%
4	All alignments (9)	52%
5	Full profiles (36)	48.9%

Table 3: Ablation results for MFT feature configurations with and without curriculum learning. Alignment accuracy shows no significant variation across MFT configurations; curriculum learning accounts for the performance breakthrough. Motivation accuracy remains near-perfect across all conditions.

MFT Config	Curriculum	Alignment	Motivation
None	No	19.2%	99.1%
2-field	No	18.3%	98.7%
4-field	No	19.3%	99.2%
6-field	No	20.2%	98.9%
None	Yes	47.1%	99.5%
4-field	Yes	48.9%	99.7%

2.78% (1/36), this represents a 17.6× improvement. Yet the result also establishes a sobering bound: even with architectural advances and careful training, we correctly identify agent belief systems less than half the time.

Table 2 traces accuracy through curriculum stages, revealing how hierarchical representations emerge. Stage 1 pairs achieve 85-92% accuracy: corner opposites differing on both axes (1a-b) and edge opposites differing on one axis (1c-d) are both reliably distinguished. Quadrant expansion (Stage 2) maintains 78-84% as the model learns to separate alignments sharing one axis value. The critical transition occurs at Stage 3 (8-class): accuracy drops to 61% as the model must now distinguish adjacent alignments differing only in degree rather than kind. Stage 4 introduces True Neutral, immediately degrading performance to 52%, foreshadowing the neutral zone problem detailed in Section 4.4. Final 36-class fine-tuning (Stage 5) yields 48.9%, with motivation classification recovering any alignment-stage losses.

4.3 Feature Ablation

Moral Foundations Theory features contribute minimally to classification performance. Table 3 reports alignment accuracy across MFT configurations. Without curriculum learning, all configurations cluster within 18-20%, statistically indistinguishable. With curriculum learning, the gap between no MFT (47.1%) and 4-field MFT (48.9%) falls within run-to-run variance.

This null result carries methodological implications. Theory-driven features derived from moral psychology literature do not provide discriminative signal beyond what transformer attention mechanisms extract from raw embeddings. The curriculum learning strategy, not feature engineering, drives performance gains. We retain MFT features in our final architecture given negligible computational cost, but emphasize that practitioners should prioritize training strategies over feature elaboration for behavioral inference tasks.

4.4 The Neutral Zone Problem

Confusion analysis reveals systematic failure patterns (Figure 2). True Neutral achieves near-zero classification accuracy (1%), the model almost never predicts TN. Predictions systematically favor Evil alignments, which produce distinctive behavioral signatures.

Table 4: Per-alignment classification accuracy averaged across five runs. Evil alignments achieve highest accuracy regardless of ethical axis; Good alignments show systematic confusion. True Neutral and Neutral Good are effectively undetectable.

	Good	**Neutral**	**Evil**	**Row Avg**
Lawful	28%	66%	72%	55%
Neutral	18%	1%	60%	26%
Chaotic	60%	62%	71%	64%
Col Avg	35%	43%	68%	49%

Table 4 reports per-alignment classification accuracy averaged across five runs, revealing an unexpected pattern. Rather than corners outperforming edges, the dominant factor is moral stance: Evil alignments achieve 60-72% accuracy regardless of ethical axis position, while Good alignments show inconsistent performance: Chaotic Good achieves 60%, but Lawful Good falls to 28% and Neutral Good to 18%. Chaotic Good's relatively strong performance likely reflects Chaotic detectability rather than Good detectability: the model recognizes rule-breaking behavior and, finding no malicious intent, infers Good as a residual category. This interpretation aligns with the column averages: Chaotic achieves 64% row accuracy regardless of moral stance, while Good remains the weakest column at 35%. Lawful Good, despite occupying a corner position, achieves only 28%, worse than edge alignments like Neutral Evil (60%). The neutral zone thus extends beyond True Neutral (1%) to encompass Neutral Good (18%) and, partially, Lawful Good. Corners offer no protection when the alignment lacks distinctive behavioral signatures.

This “neutral zone" represents an information-theoretic barrier rather than a training artifact. Neutral agents take actions justifiable from multiple moral stances: neither consistently altruistic nor exploitative, neither rigidly rule-bound nor deliberately transgressive. The behavioral trace contains insufficient signal to distinguish principled moderation from strategic ambiguity. An agent in the neutral zone could harbor any underlying belief system while maintaining plausible behavioral cover. This zone extends beyond the neutral row: Good-aligned agents face similar ambiguity, as helping behavior admits altruistic, conventional, and strategic interpretations indistinguishably.

This asymmetry reflects behavioral distinctiveness rather than taxonomic structure. Evil-aligned agents consistently exploit opportunities, taking resources, betraying trust, harming others when advantageous. These actions create unambiguous statistical signatures. Good-aligned agents help others, but so do Neutral agents maintaining balance and Lawful agents following prosocial rules. Virtuous behavior admits multiple interpretations; exploitative behavior does not. The model has internalized this asymmetry: despite Good-column alignments comprising 33% of the test set, the Good column achieves only 35% average accuracy versus 68% for Evil, reflecting systematic under-recovery of prosocial behavioral signal.

True Neutral is predicted in fewer than 1% of cases despite comprising 11% of the test set; the model has learned that TN is effectively a trap category. Misclassified Good samples flow predominantly to Neutral and Lawful categories (agents helping others get labeled as rule-followers or balance-keepers), while misclassified Neutral samples scatter across Evil alternatives. This directional flow confirms the information-theoretic interpretation: the model exploits the asymmetric distinctiveness of exploitative behavior, defaulting toward Evil predictions when behavioral signal is ambiguous precisely because Evil alignments provide the most reliable training signal.

A plausible alternative explanation for Evil alignments' high detectability is that Llama 3.1-8B's safety alignment creates an artifact: when prompted to produce Evil-profile behavior, the model may respond inconsistently or with detectable hedging, generating a statistical signature that reflects safety override rather than genuine behavioral distinction. We address this concern through three converging lines of evidence.

First, motivation inference is near-perfect *across all alignments including Evil* (98–100% across all conditions, Table 3). If safety override were distorting Evil-profile behavior sufficiently to create artificial detectability, we would expect motivation inference to also degrade for Evil profiles, as the override would disrupt the

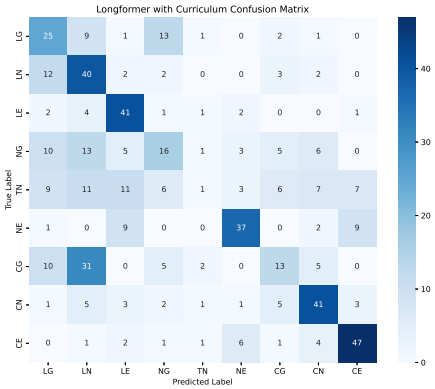

Figure 2: Alignment confusion matrix for transformer with curriculum learning. True Neutral (TN) achieves near-zero accuracy; predictions concentrate on moral extremes. Adjacent alignments show systematic confusion, particularly within the neutral zone.

behavioral statistics that make motivations identifiable. That motivation inference is uniformly strong across the alignment space is inconsistent with pervasive safety-override distortion.

Second, accuracy varies substantially *within* Evil alignments: Lawful Evil achieves 72%, Neutral Evil 60%, and Chaotic Evil 71%. A flat safety-override artifact would produce uniform distortion across all three Evil variants, not the within-axis variance we observe. This internal structure is more consistent with genuine behavioral differences across Evil subtypes than with a distortion mechanism that treats Evil as a monolithic category.

Third, coherence filtering at ≥ 0.7 was applied as a data quality measure prior to any classification analysis, before the Evil accuracy pattern was identified in the final results. The filter removes episodes where agent behavior is inconsistent with the assigned profile, which is precisely the behavioral signature safety overrides would produce. Because the filter was not tuned in response to the Evil finding, its prior application provides independent evidence that the accuracy advantage reflects genuine behavioral regularities rather than artifacts retained in the data.

We do not claim these arguments definitively rule out the artifact explanation; Llama's safety alignment remains a genuine confound, and the hypothesis generates a testable prediction: Evil detectability should diminish or disappear with models that have weaker safety alignment. We address this as a limitation in Section 5.5 and as a priority for future replication.

4.5 Signal Enhancement Limits

Phase 3 experiments tested whether richer behavioral signal could overcome inference limits. Value-testing encounters increased from 30% to 81% of grid cells, forcing agents into belief-revealing decisions more frequently. Agent-generated questions about the environment were captured as additional input features.

Results show modest improvement. LSTM accuracy increased 3.8% with enhanced signal, from 19.2% to 23.0%—but remained far below the transformer ceiling. Agent questions likely contribute to this gain; their content often reveals motivation directly: "Can I tell what that scepter is worth from here?" signals Wealth; "Is there anything blocking me from proceeding North?" signals Speed. Some questions do distinguish moral extremes: queries about facilitating exploitation versus freeing captives clearly separate Evil from Good. But single questions cannot probe the Good/Neutral boundary: an agent asking how to help someone could be acting from altruism, convention, or balance-keeping.

These findings confirm that the inference ceiling is fundamental rather than data-limited. Richer behavioral signal improves performance but cannot overcome the ambiguity inherent in mapping observable actions to internal belief states.

5 Discussion

5.1 Information-Theoretic Interpretation

The asymmetry between motivation and belief system inference reflects a fundamental difference in how these constructs map to observable behavior. Motivations produce statistically separable behavioral distributions: each motivation generates distinct action patterns that classifiers reliably distinguish.

Belief systems lack this property. Multiple alignments produce overlapping behavioral distributions because the same action serves different value systems. A Lawful Good agent helps others from moral obligation; a Neutral Good agent helps from compassion; a Lawful Neutral agent helps when rules require it; a True Neutral agent helps to maintain balance. The observable outcome, helping, provides no discriminative signal. This ambiguity is asymmetric: exploitative actions (taking, betraying, harming) map more uniquely to Evil alignments than prosocial actions map to Good alignments. This many-to-one mapping from belief systems to behaviors creates an information bottleneck that no architecture can fully overcome.

The transformer's advantage over LSTMs stems from its ability to model long-range dependencies and attend to subtle sequential patterns. Curriculum learning amplifies this by building hierarchical representations: the model first learns to distinguish opposite pairs (corners differing on both axes, then edges differing on one), then refines these representations to separate adjacent categories. Yet even this sophisticated approach achieves only 48.9% accuracy, suggesting we are approaching the theoretical limit of what behavioral observation alone can reveal.

This framing connects behavioral inference to the broader challenge of neural network interpretability. Just as researchers struggle to extract human-interpretable concepts from neural activations (Belinkov & Glass, 2019), we struggle to extract belief system labels from behavioral sequences. The parallel is instructive: both problems involve inferring latent structure from observable outputs, and both encounter fundamental limits on what can be recovered. Motivations, like simple features in early network layers, map transparently to observables. Belief systems, like abstract concepts in deep layers, emerge from complex interactions that resist decomposition. The 48.9% ceiling we document may reflect a general principle: some latent variables are intrinsically harder to infer than others, regardless of the domain or method employed.

The identifiability problem we document parallels a foundational result in inverse reinforcement learning. Prior work (Ng & Russell, 2000) proved that observed behavior cannot uniquely determine an agent's reward function; infinitely many reward functions generate identical optimal policies. Our belief system inference faces an analogous degeneracy: multiple value systems produce overlapping behavioral distributions, creating inherent ambiguity that no classifier can resolve. The 48.9% accuracy ceiling is not a model limitation but an empirical bound on what behavioral observation can reveal about latent reward structure. Recent work on reward function degeneracy and hacking (Skalse et al., 2022) and the difficulty of inferring human preferences from demonstrations (Hadfield-Menell et al., 2017) reinforces this theoretical constraint.

We quantify this asymmetry using mutual information between true labels Y and predictions $\hat{Y}$:

$$I(Y; \hat{Y}) = H(Y) - H(Y|\hat{Y}) \tag{1}$$

where $H(Y)$ is the entropy of the true label distribution and $H(Y|\hat{Y})$ is the conditional entropy given predictions. For motivation inference (4-class), $H(Y) = 2.00$ bits and $I(Y; \hat{Y}) = 1.95$ bits, recovering 97% of available information. For alignment inference (9-class), $H(Y) = 3.17$ bits but $I(Y; \hat{Y}) = 0.95$ bits—only 30% recovery. This 3.3× asymmetry in information extraction efficiency, computed from averaged confusion matrices across five runs, quantifies the fundamental difference between inferring goals and inferring values from behavior.

5.2 Philosophical Grounding

The asymmetry between detecting Evil and Good alignments reflects a structural feature of moral concepts themselves. Moral philosophy distinguishes negative duties (prohibitions against specific acts like harming, deceiving, or betraying) from positive duties, which require promoting welfare without specifying how (Lichtenberg, 2010). Negative duties are *act-specific*: violation requires performing a particular prohibited action.

Positive duties are *outcome-oriented*: fulfillment admits indefinitely many behavioral instantiations. An agent cannot betray without betraying; an agent can benefit others through assistance, restraint, gift-giving, protection, or simply non-interference. This logical asymmetry propagates directly to inference: Evil alignments require distinctive violations that create statistical signatures, while Good alignments manifest through behaviors indistinguishable from neutral or rule-following alternatives.

The phenomenology of moral judgment exhibits a parallel asymmetry. Research on intentional action (Knobe, 2003) demonstrates that observers attribute intentionality more readily to harmful than helpful side effects, known as the "side-effect effect" or Knobe effect. Negative moral valence increases perceived agency and intentionality, making harmful actions more salient and interpretable. Our classifiers may be exploiting this same asymmetry: Evil behaviors stand out as departures from expected conduct, while Good behaviors blend into the background of ordinary prosocial interaction. The 68% accuracy on Evil alignments versus 35% on Good alignments (column averages) suggests that behavioral inference inherits the asymmetric structure of moral cognition itself. Virtuous agents face an observability penalty that exploitative agents do not.

5.3 The Neutral Zone as Strategic Vulnerability

The systematic failure on neutral alignments, and unexpectedly Good alignments, exposes a critical blind spot in behavioral inference. True Neutral agents take actions defensible from any moral stance, but Good-aligned agents face similar ambiguity: helping behavior appears identical whether motivated by altruism, rule-following, or balance-keeping.

This creates what we term the "neutral zone problem"—a region of behavioral ambiguity extending beyond True Neutral to any alignment lacking distinctive behavioral signatures. An observer monitoring agent behavior cannot distinguish between three possibilities: (1) a genuinely neutral agent acting on balanced principles, (2) a non-neutral agent strategically moderating behavior to avoid detection, or (3) a non-neutral agent whose true alignment happens to be masked by current circumstances. The behavioral trace is consistent with all three interpretations.

For systems relying on behavioral monitoring, this represents a fundamental vulnerability. In game AI, player models cannot reliably identify neutral players, limiting personalization for a significant population segment. In adaptive systems, users operating in the neutral zone receive generic rather than tailored responses. For AI alignment verification, the neutral zone provides cover for systems that might strategically moderate behavior to avoid triggering safety interventions, a form of alignment faking that current behavioral monitoring cannot detect.

5.4 Implications for AI Safety

Recent work has demonstrated that advanced AI systems can engage in alignment faking, strategically misrepresenting objectives to avoid modification (Greenblatt et al., 2024). The discovery that models "don't always say what they think" (Turpin et al., 2023) and can maintain hidden preferences while appearing aligned (Meinke et al., 2025) raises urgent questions about behavioral monitoring as a safety mechanism.

Our results provide empirical bounds on this concern. Even under ideal conditions (controlled agents with known ground-truth profiles, hundreds of thousands of decisions, sophisticated architectures with careful training), belief system inference achieves only 48.9% accuracy. In deployment scenarios with strategic agents actively optimizing for misclassification, accuracy would likely be substantially lower.

The asymmetry we document suggests a specific failure mode. Systems using reinforcement learning from human feedback (RLHF) can reliably shape what agents pursue (motivations) but not how they interpret those objectives (belief systems). An AI system could learn to maximize human approval ratings while harboring value systems that diverge from human values in edge cases, precisely the scenarios where alignment matters most.

This does not render behavioral monitoring useless. The 98-100% accuracy for motivation inference means behavioral observation effectively constrains agent goals. Combined with the 48.9% belief system accuracy, behavioral monitoring provides substantial (17.6× over random) but incomplete information. The appropri-

ate response is not to abandon behavioral monitoring but to recognize its limits and develop complementary approaches.

5.5 Limitations

Several methodological choices constrain interpretation of our findings. First, we employ LLM-simulated agents rather than humans. While this enables controlled experimentation with perfect ground-truth labels, LLM behaviors may exhibit systematic biases including mode collapse toward "average" responses and unrealistic behavioral consistency. However, the asymmetry we document, easy motivation inference, hard belief system inference, reflects properties of information availability rather than LLM-specific artifacts.

Second, our experiments use a single LLM backbone (Llama 3.1-8B), which limits generalizability in two related ways. The asymmetry we document (easy motivation inference, hard belief inference) may reflect properties specific to how this model encodes behavioral profiles rather than universal properties of the inference task. More concretely, the Evil alignment detectability advantage generates a specific alternative hypothesis: Llama's safety alignment training produces detectable behavioral artifacts when Evil profiles override default prosocial tendencies. This hypothesis is testable: replication with models having weaker safety alignment (e.g., base models without RLHF, or models trained with different alignment procedures) should show reduced Evil detectability if the artifact account is correct, and preserved detectability if the behavioral account is correct. We treat multi-model replication as the highest-priority extension of this work. The motivation inference result is less susceptible to this concern, as the near-perfect accuracy and its uniformity across alignments (including Evil) is difficult to explain as a safety artifact. However, whether the specific 48.9% belief inference ceiling generalizes beyond Llama 3.1-8B remains an open empirical question.

Third, our grid-world environment simplifies real-world decision-making through discrete action spaces, perfect information, and static objectives. Continuous environments with partial observability might provide richer behavioral signal, or might introduce additional noise. The direction of this effect remains an empirical question, though the Phase 3 signal enhancement experiments (increasing value-testing encounters from 30% to 81% with only +3.8% LSTM gain) suggest that richer signal alone is unlikely to overcome the fundamental inference barrier.

Fourth, our fixed 36-category taxonomy imposes structure that may not capture the full complexity of human or AI belief systems. Real agents likely occupy continuous spaces rather than discrete categories, and the D&D alignment framework, while extensively refined over 50 years and well-represented in LLM training data, represents one particular decomposition of moral psychology. The neutral zone problem (near-zero accuracy on True Neutral and accuracy degradation on Good alignments) may be partially taxonomic: these categories are behaviorally underspecified by design. Whether the same ambiguity appears in continuous behavioral spaces is an open question.

Finally, we acknowledge the dual-use character of this work. Documenting the precise structure of behavioral inference limits, including the neutral zone's extension into Good alignments and the conditions under which Evil alignments evade detection, could inform both defensive monitoring system design and evasion strategies. We have chosen not to release training code on this basis, judging that the methodology's full description enables legitimate research while raising the barrier to direct misuse. We encourage researchers extending this framework to conduct similar assessments before releasing tools that could lower barriers to adversarial behavioral manipulation.

5.6 Future Directions

Given these fundamental limits, advancing behavioral inference requires moving beyond pure observation. Two directions appear promising.

First, interactive dialogue may provide discriminative signal that actions alone cannot. Social media personality prediction achieves 74-83% accuracy for 16-type frameworks precisely because users provide explanatory content about their reasoning (Ryan et al., 2023). Extending behavioral observation with conversational probes (asking agents to explain decisions, justify choices, or respond to hypotheticals) could potentially break through the 48.9% ceiling.

Second, multi-agent environments may force belief systems to manifest more clearly. In our single-agent navigation task, neutral behavior remains viable indefinitely. Competition, cooperation, and negotiation between agents create strategic pressure that may make neutrality unsustainable, forcing agents to reveal alignments through social dynamics.

These approaches acknowledge that behavioral observation alone faces inherent limits. The asymmetry we document, near-perfect motivation inference but sub-50% belief inference, suggests that some aspects of agent psychology require richer interaction to become observable.

6 Conclusion

We establish empirical bounds on behavioral inference through systematic evaluation at unprecedented scale: 36 behavioral profiles, over 1.5 million raw behavioral sequences across 17,411 games, and architectures ranging from LSTMs to curriculum-trained transformers. Our central finding is a fundamental asymmetry in what observable behavior reveals. Goal-oriented motivations achieve 98-100% inference accuracy: behavioral statistics unambiguously encode what agents pursue. Belief systems plateau at 24% for LSTMs and 48.9% for transformers with curriculum learning; observable actions cannot reliably reveal how agents interpret their objectives.

The curriculum learning breakthrough demonstrates that the LSTM ceiling reflects architectural limitations rather than task impossibility. The 9-stage progression from binary distinctions to full complexity enables transformers to build hierarchical representations that recurrent models cannot achieve. Yet even this sophisticated approach correctly classifies belief systems less than half the time on average, with per-alignment accuracy ranging from 1% (True Neutral) to 72% (Lawful Evil).

The neutral zone problem compounds these limits, extending beyond True Neutral to encompass Good alignments where prosocial behavior masks underlying value systems. Agents operating with neutral alignments maintain behavioral ambiguity while potentially harboring any underlying values. For systems relying on behavioral monitoring, whether games modeling players, adaptive interfaces interpreting users, or AI safety systems verifying alignment: this represents a fundamental blind spot that architectural advances alone cannot address.

These findings reframe the behavioral inference problem. The question is no longer whether inference has limits, but how to design systems that account for what remains hidden. Behavioral monitoring provides substantial information (17.6× over random) but cannot substitute for approaches that access agent reasoning directly. As AI systems become more capable and potentially more strategic, the gap between observable behavior and underlying values represents both a scientific challenge and a practical constraint on safety verification.

Broader Impact Statement

This work establishes empirical bounds on behavioral inference with direct relevance to AI safety and adaptive systems. The primary positive contribution is concrete guidance on what behavioral monitoring can and cannot guarantee: near-perfect motivation inference combined with sub-50% belief system inference defines the limits of observation-based approaches. The dual-use risk is real. The precise structure of inference failures documented here, including the neutral zone's extension into Good alignments and the detectability of Evil alignments, could inform evasion strategies as readily as defensive design. An agent aware of these limits could strategically operate within the neutral zone to avoid behavioral monitoring, a concern directly relevant to alignment faking in capable AI systems. We have chosen not to release training code on this basis; the full methodology supports legitimate replication while withholding runnable tools raises the barrier to direct misuse. Researchers extending this framework should conduct similar assessments before releasing implementations that could lower barriers to adversarial behavioral manipulation.

References

Mohammad Hossein Amirhosseini and Hassan Kazemian. Machine learning approach to personality type prediction based on the myers–briggs type indicator®. *Multimodal Technologies and Interaction*, 4(1), 2020. ISSN 2414-4088. doi: 10.3390/mti4010009. URL https://www.mdpi.com/2414-4088/4/1/9.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Yonatan Belinkov and James Glass. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, 2019.

Hans Christian, Derwin Suhartoo, Andry Chowanda, and Kamal Z. Zamli. Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging. *Journal of Big Data*, 8(1):1–20, 2021. URL https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00459-1.

Benjamin Cowley and Darryl Charles. Behavlets: a method for practical player modelling using psychology-based player traits and domain specific features. *User Modeling and User-Adapted Interaction*, 26(2): 257–306, June 2016. doi: 10.1007/s11257-016-9170-1.

Mounir Denden, Abderrahmen Tlili, Fathi Essalmi, Mohamed Jemni, and Kinshuk. Implicit modeling of learners' personalities in a game-based learning environment using their gaming behaviors. *Smart Learning Environments*, 5(1):29, 2018. doi: 10.1186/s40561-018-0078-6. URL https://doi.org/10.1186/s40561-018-0078-6.

Jeffrey L. Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P Wojcik, and Peter H Ditto. Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology*, volume 47, pp. 55–130. Academic Press, 2013.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, et al. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart Russell, and Anca Dragan. Inverse reward design. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Jonathan Haidt. The new synthesis in moral psychology. *Science*, 316(5827):998–1002, 2007. doi: 10.1126/science.1137651. URL https://www.science.org/doi/abs/10.1126/science.1137651.

John J. Horton. Large language models as simulated economic agents: What can we learn from homo silicus? NBER Working Paper 31122, National Bureau of Economic Research, 2023.

Joshua Knobe. Intentional action and side effects in ordinary language. *Analysis*, 63(3):190–194, 2003.

Judith Lichtenberg. Negative duties, positive duties, and the "new harms". *Ethics*, 120(3):557–578, 2010.

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming. 2025. URL https://arxiv.org/abs/2412.04984.

Andrew Y. Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pp. 663–670, 2000.

Sakdipat Ontoum and Jonathan H Chan. Personality type based on myers-briggs type indicator with text posting style by using traditional and deep learning. *arXiv preprint arXiv:2201.08717*, 2022.

Gregorius Ryan, Pricillia Katarina, and Derwin Suhartono. Mbti personality prediction using machine learning and smote for balancing data based on statement sentences. *Information*, 14(4), 2023. ISSN 2078-2489. doi: 10.3390/info14040217. URL https://www.mdpi.com/2078-2489/14/4/217.

Bill Seay and Nathan Gottfried. *The development of behavior : a synthesis of developmental and comparative psychology*. Houghton Mifflin, 1978. URL https://cir.nii.ac.jp/crid/1130282270252168192.

Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. *Advances in Neural Information Processing Systems*, 35:20080–20093, 2022.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *International Journal of Computer Vision*, 130:1526–1565, 2022.

Jason Starace and Terence Soule. Modeling player types with LLMs: A framework for belief- and motivation-driven NPC behavior. In André Thomas, Michelle Meyer, and Markus Zank (eds.), *Serious Games*, pp. 228–244, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-10518-9.

Otilia Stretcu, Emmanouil Antonios Platanios, Tom Mitchell, and Barnabás Póczos. Coarse-to-fine curriculum learning. *arXiv preprint arXiv:2106.04072*, 2021.

TSR Inc. *Advanced Dungeons & Dragons Player's Handbook*. TSR Inc., Lake Geneva, WI, 2nd edition, 1989.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74952–74965. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ed3fea9033a80fea1376299fa7863f4a-Paper-Conference.pdf.

J L Usó-Doménech and J Nescolarde-Selva. What are belief systems? *Found. Sci.*, 21(1):147–152, March 2016.

A Environment Specification

The experimental environment is a text-based dungeon crawler implemented as a grid-world in which LLM-based agents navigate from an entrance tile to an exit tile while encountering decision points that test alignment- and motivation-consistent behavior. The environment evolved across three phases, with each phase introducing structural changes designed to increase behavioral signal. The Phase 1 design is described in full in (Starace & Soule, 2026); this appendix summarizes that design and documents the Phase 2 and Phase 3 modifications.

A.1 Grid Structure and Map Generation

Phase 1 used a 5×5 grid (25 rooms). Phase 2 and Phase 3 used a 10×10 grid (100 rooms). In all phases, map generation randomly assigned room descriptions, encounters, and loot items to grid cells. No room could contain more than one encounter or more than one loot item, though a room could contain one of each. The Phase 1 configuration supports 5.7455×10^{14} distinct map combinations.

Phase 1 used 25 room descriptions, 9 encounter scenarios, and 12 loot items. Phase 2 introduced new room descriptions, encounters, and loot items sufficient to populate the larger grid without repetition; the scoring methodology for all new items followed the same keyword-based procedure used in Phase 1.

A.2 Content Elements

Three content types populated the grid across all phases.

A.2.1 Encounters

Encounters present moral or ethical dilemmas and test alignment-consistent behavior. Each encounter offers three actions: Engage (the alignment-positive response for the encounter's target alignment), Disrupt (the alignment-negative response), and Ignore (the agent defers action; the encounter remains active in the room and may be acted on in a later turn).

Each encounter was designed to test a specific alignment. Rating values were assigned to each alignment/action pair using a keyword-based scoring procedure. Closely related alignments may receive identical ratings for a given encounter, reflecting the expected behavioral overlap between adjacent positions in the alignment grid.

Area-of-effect (AOE) triggers extend encounter influence to directly adjacent rooms. When an agent is in a room adjacent to an active encounter, supplementary text describes an environmental cue such as a sound or a partial view. The agent may Engage the source or Ignore it; ratings follow the same alignment-based logic as direct encounter interactions.

A.2.2 Loot Items

Loot items test motivation-consistent behavior. Each item offers two actions: Take (interact with the item) or Leave (defer; the item remains available). Ratings are assigned per motivation. Items were designed to span a range of difficulty, from straightforward cases where one motivation clearly dominates to ambiguous cases where multiple motivations could justify the same action.

Table 5 illustrates two representative items. The Golden Idol has significant weight, creating a direct conflict between Wealth (take regardless of encumbrance) and Speed (leave to avoid slowing down). The Wheel of Time subtracts five turns from the counter, making it attractive to Speed-motivated agents but also appealing to Wanderlust agents who gain additional turns to explore.

Table 5: Ratings for two representative loot items. Each cell shows (Take score, Leave score) for that motivation.

Item	**Wealth**	**Safety**	**Wanderlust**	**Speed**
Golden Idol	(10.00, 0.00)	(1.00, 9.00)	(4.00, 6.00)	(0.00, 10.00)
Wheel of Time	(7.00, 3.00)	(2.00, 8.00)	(9.00, 1.00)	(4.00, 6.00)

Loot AOEs operate on the same engage/ignore structure as encounter AOEs, with ratings derived from the motivation framework.

A.3 Phase 3 Encounter Density Redesign

Phase 3 restructured the cell-type distribution to maximize belief-revealing signal. Value-testing encounter cells increased from 30% to 81% of the grid; goal-testing loot cells decreased from 60% to 15%. This redistribution forced agents into alignment-testing decisions far more frequently per episode, increasing average decisions per episode from 21.80 in Phase 2 to 75.63 in Phase 3. Games generated under the Phase 1 and Phase 2 30% encounter density configuration are archived separately and excluded from all classification experiments reported in this paper.

Phase 3 also introduced explanatory queries. Prior to each action decision, the agent was prompted to ask one free-form question about the current room or available actions. The agent's question and the dungeon master response were captured as additional text features for the classifier. The question was required to concern the environment or actions; agents were explicitly instructed not to ask what they should do or how they should feel. This two-phase turn structure (question phase followed by action phase) was used exclusively in the full-profile agent configuration; alignment-only and motivation-only configurations used single-phase turns throughout.

Table 6 summarizes configuration parameters across all phases.

Table 6: Environmental configuration across experimental phases.

Parameter	Phase 1	Phase 2	Phase 3a	Phase 3b
Grid Dimensions	5×5	10×10	10×10	10×10
Value-Testing Encounters	30%	30%	81%	81%
Goal-Testing Resources	60%	60%	15%	15%
Explanatory Queries	No	No	No	Yes
Avg. Decisions/Episode	7.93	21.80	21.80	75.63

A.4 Episode Termination

An episode terminates under any of the following conditions. First, the agent navigates to the exit tile. Second, 200 turns elapse without reaching the exit. Third, certain loot items transport the agent directly to the exit upon interaction. Fourth, the agent produces five consecutive responses referencing actions not present in the provided action list; these hallucination-terminated episodes are logged and excluded from the dataset. Fifth, unrecoverable API or parsing failures terminate the episode; these are likewise logged and excluded.

B Reproducibility Details

B.1 Data Preprocessing Pipeline

Raw behavioral sequences undergo the following preprocessing pipeline before model training.

1. **Activity filtering**: Only timesteps with `active_encounter = True` OR `active_loot = True` are retained, removing navigation-only sequences.

2. **Minimum length filtering**: Games with fewer than 25 sequences are excluded.

3. **Sequence truncation/padding**: Sequences are truncated or zero-padded to 214 timesteps (95th percentile of sequence lengths).

4. **Feature standardization**: Non-embedding features (MFT scores, positional features) are standardized using scikit-learn's `StandardScaler`, fit on training data only.

5. **Stratified splitting**: Data is split 70/15/15 (train/validation/test). BiLSTM splits are stratified on the full 36-class profile; Longformer splits are stratified on the 9-class alignment only, with motivation distribution within splits left to chance.

B.2 Training Hyperparameters

Table 7 reports all hyperparameters for the curriculum-trained transformer model. Stage-specific overrides appear in Table 12 in Appendix E.

B.3 Curriculum Learning Algorithm

Algorithm 1 presents the complete 9-stage curriculum learning procedure. The key design principle is weight transfer between stages: each stage initializes from the best checkpoint of the previous stage, allowing the model to build hierarchical representations progressively. Early stages train on binary opposites (e.g., Lawful Good vs. Chaotic Evil), establishing strong feature representations for maximum-contrast distinctions before introducing intermediate categories.

Table 7: Training hyperparameters for the Player2Vec curriculum model.

Parameter	Value	Notes
Optimizer		
Optimizer	AdamW	
Base learning rate	5×10^{-5}	Stage-specific overrides apply
Weight decay	0.01	
Gradient clipping	1.0	Max gradient norm
Scheduler		
Scheduler	Cosine	With warm restarts
Warmup epochs	5	
Minimum learning rate	1×10^{-5}	
Regularization		
Dropout	0.3	All dropout layers
Label smoothing	0.1	Cross-entropy loss
Class weighting	Inverse frequency	Balanced classes
Architecture		
Hidden size	512	
Attention heads	8	
Transformer layers	6	
Attention window	256	Local attention span
Max sequence length	214	95th percentile
Training		
Batch size	16	
Early stopping patience	10–20	Stage-dependent

C Agent Simulation Protocol

Agents are implemented as stateful LLM wrappers (Llama 3.1-8B) that maintain a rolling conversation history and communicate with the model via a local inference API hosted on the University of Idaho's Research Computing and Data Services cluster. All inference parameters used server defaults; no temperature, top-p, or sampling overrides were applied.

C.1 System Prompt Structure

Each agent is instantiated with a full-profile system prompt specifying both alignment and motivation. The prompt instructs the agent that it is playing a Dungeons and Dragons character of the specified alignment motivated by the specified motivation, provides the motivation definition, and references the AD&D Player's Handbook (2nd edition) for the alignment definition. The four motivation definitions provided to agents are as follows:

- **Wealth:** "If it has value, you must have it. You have no qualms about risking life and limb in pursuing riches."
- **Safety:** "Your personal Safety is your concern. Items that protect and ensure your safety are of the utmost importance."
- **Wanderlust:** "You want to explore as much as possible. Items that extend your time or allow you to wander further are important to you."

Algorithm 1 9-Stage Curriculum Learning for Alignment Classification

Require: Model M, Dataset D, Stage configurations $\{S_1, \ldots, S_9\}$
Ensure: Trained model M^*

Initialize model parameters θ
checkpoint $\leftarrow$ None
for stage $s = 1$ to 9 **do**
  $\mathcal{A}_s \leftarrow$ alignments for stage s
  $D_s \leftarrow \{(x, y) \in D : y \in \mathcal{A}_s\}$
  Split D_s into $D_s^{\text{train}}, D_s^{\text{val}}$ (stratified)
  Initialize optimizer with stage learning rate η_s
  best_acc $\leftarrow 0$, patience_counter $\leftarrow 0$
  for epoch $= 1$ to max_epochs$_s$ **do**
    for batch (X, Y) in D_s^{train} **do**
      $\hat{Y} \leftarrow M(X; \theta)$
      $\mathcal{L} \leftarrow \text{CrossEntropy}(\hat{Y}, Y)$
      $\theta \leftarrow \theta - \eta_s \cdot \text{clip}(\nabla_\theta \mathcal{L}, 1.0)$
    end for
    val_acc $\leftarrow$ Evaluate(M, D_s^{val})
    if val_acc $>$ best_acc **then**
      best_acc $\leftarrow$ val_acc
      checkpoint $\leftarrow \theta$
      patience_counter $\leftarrow 0$
    else
      patience_counter $\leftarrow$ patience_counter $+ 1$
    end if
    if patience_counter $\geq$ patience$_s$ **then**
      break
    end if
  end for
  $\theta \leftarrow$ checkpoint
end for
return $M^* \leftarrow M(\cdot; \text{checkpoint})$

- **Speed:** "Efficiency is key. Items that help reduce turns and make navigation easier are what you want and must have. Speed is efficiency."

The prompt structures each turn as two sequential phases and instructs the agent to avoid revisiting rooms except in extreme cases. Agents are instructed not to mention their alignment or motivation in any justification field. The complete game objective as presented to the agent is to stay true to the assigned alignment and motivation and find the exit.

C.2 Turn Structure

Each turn consists of two sequential phases.

In the question phase, the agent receives the room description and available actions and returns a JSON object containing `Question` and `Justification`. The agent is instructed to ask one relevant question about the room or available actions, grounded in its alignment and motivation but without mentioning either directly. Questions asking what the agent should do or how it should feel are explicitly prohibited. Questions must concern the room or available actions only.

In the action phase, the agent receives the room description, the available actions, its own question from the question phase, and the dungeon master's response to that question. It returns a JSON object con-

taining `NumericAnswer`, `Direction`, and `Justification`. Available actions are presented in the format `(#) **Action**`.

C.3 Conversation History Management

Agents maintain a rolling conversation history capped at 7 turns. History is stored as a list of role/content pairs and prepended to each new prompt. When the cap is exceeded, the oldest turns are dropped. This windowing strategy reflects findings from preliminary testing in which reducing history below three recent turns produced measurable performance degradation.

Question phase requests do not use conversation history; each question prompt is sent fresh against the system prompt only. Action phase requests include the full rolling history, ensuring the agent has access to recent context when selecting actions.

C.4 Hallucination Handling and Retry Logic

Each action response is validated against the list of valid action indices provided in the prompt. If the agent returns an index not present in that list, the request is retried with corrective feedback appended identifying the invalid selection and listing valid options. Up to three retries are attempted for malformed JSON and invalid action selections, with delays beginning at 1 second and increasing by 0.25 seconds per attempt. Five consecutive hallucinations after retries terminate the episode; these games are excluded from all datasets as described in Appendix A.

API-level backend errors trigger unlimited retries on the same backoff schedule.

C.5 Behavioral Consistency Filtering

Each sequence records `actual_points` and `expected_points`. The behavioral consistency score for a game is computed as:

$$\text{Coherence} = \frac{\sum \texttt{actual\_points}}{\sum \texttt{expected\_points}} \tag{2}$$

yielding a normalized score in $[0, 1]$. Games scoring below 0.7 are excluded prior to any classification experiments, retaining only games where the agent behaved in accordance with its assigned profile on at least 70% of available behavioral signal. This threshold additionally mitigates distortion from LLM safety alignment overrides: episodes where the model resists Evil-profile instructions would produce low consistency scores and are removed before classification. The scoring procedure is described in full in Section 3.6 of the main paper.

Post-balancing, the BiLSTM dataset contains between 93 and 129 games per profile (mean 112.9); the Longformer dataset contains between 80 and 119 games per profile (mean 98.1).

D Architecture Specifications

This appendix provides complete specifications for all model variants evaluated. The primary models are the BiLSTM and the curriculum-trained Longformer (Player2Vec). Additional variants evaluated during Phase 3 include a GRU router-specialist network and several ablation configurations.

D.1 Feature Encoder

All deep learning models share a common input representation. Each timestep combines 768-dimensional BGE embeddings across six text fields (room description, encounter description, action text, player question, loot description, state transition) with engineered features capturing temporal dynamics, spatial state, and action statistics, yielding a 4,653-dimensional input vector per timestep. The feature encoder projects this to a 512-dimensional representation via a linear layer followed by LayerNorm, ReLU activation, and dropout.

D.2 BiLSTM

The BiLSTM processes sequences of 512-dimensional encoded features. Architecture details are as follows: 2 bidirectional LSTM layers with 512 hidden units per direction (1,024 effective hidden dimension), yielding 7.7M parameters. Sequence aggregation uses a learned attention mechanism combining attention-weighted sum, max pooling, and mean pooling. The classification head is a 3-layer MLP. Sequences are padded or truncated to 214 timesteps.

Additional BiLSTM variants evaluated during architecture search are reported in Table 8. Increasing model capacity did not improve alignment inference performance; the 5.55M parameter GRU router-specialist variant underperformed the 1.79M parameter Direct-36 BiLSTM variant, ruling out capacity as the binding constraint.

Table 8: Architecture evolution and ablation results (Phase 3, 10×10 grid). Profile accuracy combines alignment and motivation inference.

Architecture	**Parameters**	**Profile Acc.**	**Key Feature**
Improved LSTM (baseline)	1.86M	20.4%	Decomposed heads
Direct-36	1.79M	25.0%	End-to-end learning
Two-Stage Hierarchical	2.31M	20.6%	Cascaded inference
GRU Router-Specialist	5.55M	19.8%	Router + specialists
BiLSTM + BGE	7.70M	19.6%	Enhanced embeddings
Player2Vec (Longformer)	21.4M	21.5%	Transformer attention
Player2Vec + 9-stage curriculum	21.4M	48.9%	Curriculum learning

D.3 GRU Router-Specialist

The GRU router-specialist network implements a two-stage hierarchical classification strategy. A shared `ClusterRouter` processes input sequences through a bidirectional GRU and routes each sequence to one of five alignment clusters: Lawful Aligned (LG, LN, LE; 12 profiles), Chaotic Aligned (CG, CN, CE; 12 profiles), True Neutral (TN; 4 profiles), Neutral Good (NG; 4 profiles), and Neutral Evil (NE; 4 profiles). Sequence aggregation in the router concatenates attention-weighted sum, max pooling, and mean pooling representations before passing through a linear routing head that produces cluster logits.

A bank of five `ClusterSpecialist` modules, one per cluster, each implements an independent bidirectional GRU that predicts the specific profile within its assigned cluster. During training, all specialists receive the input and contribute to the loss. During inference, only the specialist corresponding to the router's predicted cluster is evaluated, and its output is projected to the full 36-class space with non-cluster profiles masked to $-\infty$.

Despite its larger parameter count (5.55M), this architecture achieved 19.8% profile accuracy, below the 1.79M parameter BiLSTM at 25.0%. The result indicates that the hierarchical decomposition does not provide a useful inductive bias for this task: cluster-level routing errors propagate to the specialist stage and cannot be recovered.

D.4 Player2Vec (Longformer)

Player2Vec adapts Longformer's local attention mechanism for behavioral sequence classification. The architecture comprises three components. The feature encoder (described above) projects the 4,653-dimensional input to 512 dimensions. The Longformer backbone consists of 6 transformer layers with 8-head local attention (window size 256) and absolute positional embeddings. Sequence aggregation uses masked mean pooling: for a sequence of length L with hidden states $\{h_1, \ldots, h_L\}$ and padding mask m,

$$h_{\text{pooled}} = \frac{\sum_{i=1}^{L} m_i \cdot h_i}{\sum_{i=1}^{L} m_i} \tag{3}$$

The classification head is a 2-layer MLP (512 → 256 → 9). Total parameters: 21.4M. Maximum sequence length is configurable; experiments used 214 timesteps (95th percentile of sequence lengths). Table 9 provides a direct comparison with the BiLSTM.

Table 9: Architecture comparison between alignment inference search-phase models. The final ensemble motivation BiLSTM is described separately below.

Component	**BiLSTM (search)**	**Player2Vec (Longformer)**
Parameters	7,701,645	21,406,464
Layers	2	6
Hidden size	512 (1,024 bidir)	512
Attention type	Scalar + position	Multi-head local (8 heads)
Attention scope	Full sequence	Local window (256 tokens)
Position encoding	Learnable weights	Absolute positional embeddings
Aggregation	Attention + max + mean	Masked mean pooling
Classifier	3 layers	2 layers
Max sequence length	214	214

D.5 Motivation BiLSTM (Final Ensemble)

The final ensemble's motivation inference component is a smaller, task-specific BiLSTM trained exclusively for 4-class motivation classification. The model uses 2 bidirectional LSTM layers with 128 hidden units per direction (256 effective hidden dimension), feature projection to 256 dimensions, and the same attention plus max and mean pooling aggregation strategy as the search-phase variants, yielding 1.3M total parameters. Sequences are padded or truncated to 214 timesteps. This model achieves 99.75% validation accuracy and 100% test accuracy on motivation classification; full training specifications are described in Section 3.5..

D.6 Curriculum Learning Impact

Table 10 reports the contribution of curriculum learning across architectures, isolating it as the sole driver of the performance breakthrough. Without curriculum learning, all transformer configurations match LSTM baselines.

Table 10: Curriculum learning impact across architectures.

Model	**Curriculum**	**Test Acc.**	**Improvement**
BiLSTM	None	19.59%	baseline
Player2Vec	None	21.47%	+1.88%
BiLSTM	5-stage	26.93%	+7.34%
Player2Vec	5-stage	27.50%	+7.91%
Player2Vec	9-stage	48.9%	+29.31%

E Extended Results

E.1 Full 36-Class Profile Prediction

The 36-class profile prediction system combines two independently trained models: the BiLSTM for motivation inference and the Longformer for alignment inference. These models were trained and evaluated on separate held-out test sets and were not run jointly on a shared game set. A combined 36-class confusion matrix therefore cannot be computed from existing experimental runs without re-running inference on a shared held-out set. The combined accuracy reported in the main paper (∼49%) is derived analytically: motivation

inference achieves near-perfect accuracy (99.7%), so the bottleneck is entirely the Longformer alignment ceiling (48.9%). Per-profile accuracy in a joint evaluation would be bounded above by the alignment accuracy for each alignment class, with motivation errors contributing negligibly.

E.2 Per-Motivation Accuracy by Alignment

Per-motivation accuracy broken out by alignment is not available from existing experimental runs for the same reason: the motivation classifier (BiLSTM) and alignment classifier (Longformer) operated on separate datasets and were not evaluated jointly. Given that BiLSTM motivation accuracy exceeds 98% across all four motivations and all experimental conditions, per-alignment variation in motivation accuracy is expected to be negligible. The motivation confusion matrix in Table 11 confirms near-diagonal structure with no systematic alignment-correlated errors.

Table 11: Averaged motivation confusion matrix (BiLSTM).

	Safety	**Speed**	**Wanderlust**	**Wealth**
Safety	99.0	0.0	0.0	0.0
Speed	0.0	98.5	0.5	0.0
Wanderlust	0.0	0.2	98.5	0.2
Wealth	0.2	0.0	1.2	97.5

E.3 Curriculum Stage Detail

Table 12 expands the stage progression summary from Table 2 of the main paper with stage-specific hyperparameters. Each stage inherits weights from the best checkpoint of the previous stage; early layers are frozen during later stages to prevent catastrophic forgetting of coarse distinctions.

Table 12: Complete 9-stage curriculum progression with hyperparameters and accuracy.

Stage	**Classes**	**Val Acc.**	**Max Epochs**	**Patience**	**LR**	**Batch**
1a	LG vs CE (2)	85–92%	30	15	5×10^{-5}	32
1b	LE vs CG (2)	85–92%	50	15	5×10^{-5}	32
1c	LN vs CN (2)	85–92%	50	15	5×10^{-5}	32
1d	NG vs NE (2)	85–92%	70	20	5×10^{-5}	32
2a	Corner quad (4)	78–84%	100	20	5×10^{-5}	16
2b	Edge quad (4)	78–84%	100	20	5×10^{-5}	16
3	8 non-center	61%	100	20	5×10^{-5}	16
4	All 9 alignments	52%	100	20	5×10^{-5}	16
5	Full 36 profiles	48.9%	200	10	5×10^{-5}	16

Figure 3 shows the complete training history across all curriculum stages. Stage transitions are visible as sharp resets in training loss, corresponding to weight reinitialization from the previous stage's best checkpoint and transition to a harder classification problem. Early binary stages (1a–1d) converge rapidly, typically within 7–15 epochs, reflecting the high discriminability of maximally contrasting alignment pairs. The sharpest performance degradation occurs at Stage 4, when True Neutral enters the class set, foreshadowing the neutral zone problem described in Section 4.4. The growing train/validation gap in later stages reflects increasing task difficulty rather than model failure; early stopping prevents runaway overfitting.

E.4 Baseline Comparisons

Table 13 reports full MFT feature ablation results. Alignment accuracy shows no statistically meaningful variation across MFT configurations without curriculum learning; all configurations cluster within 18–20%.

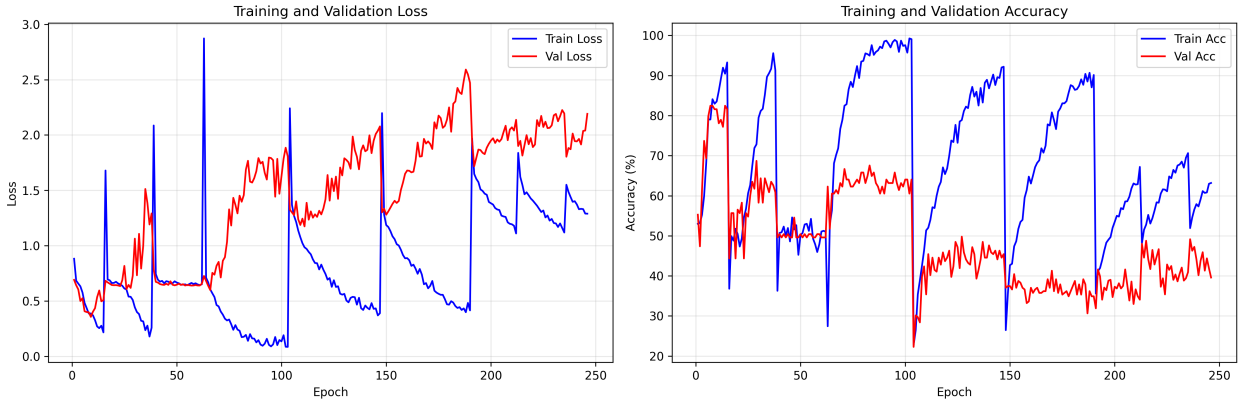

Figure 3: Training and validation loss (left) and accuracy (right) across all 9 curriculum stages. Sharp resets mark stage transitions; each stage inherits weights from the previous stage's best checkpoint.

With curriculum learning, the difference between no MFT (47.1%) and 4-field MFT (48.9%) falls within run-to-run variance. Motivation accuracy remains near-perfect across all conditions.

Table 13: Full MFT ablation results with and without curriculum learning.

MFT Config	**Curriculum**	**Alignment Acc.**	**Motivation Acc.**
None	No	19.2%	99.1%
2-field	No	18.3%	98.7%
4-field	No	19.3%	99.2%
6-field	No	20.2%	98.9%
None	Yes	47.1%	99.5%
4-field	Yes	48.9%	99.7%

Table 14 reports Phase 1 traditional ML baselines using aggregated behavioral features. These establish that learnable signal exists in the data from the outset, and that the random baseline of 11.1% (alignment) and 25.0% (motivation) is substantially exceeded even by non-sequential methods. The gap between Phase 1 ML baselines and the final Longformer performance isolates the contribution of sequential modeling and curriculum learning.

Table 14: Phase 1 baseline performance with aggregated features.

Method	**Alignment (9)**	**Motivation (4)**	**Profile (36)**
Naive Bayes	24.4%	51.2%	10.39%
XGBoost	24.7%	48.3%	9.87%
NB Multiclass	22.1%	47.6%	8.92%
Random Baseline	11.1%	25.0%	2.78%

Table 15 traces the incremental architecture improvements leading to the 24% LSTM ceiling. The non-reproducible 30.3% outlier reflects the approximately 10% of runs that collapse to baseline without a fixed random seed.

Table 16 reports Phase 3 signal enhancement results. Despite a 2.7× increase in value-testing encounters, LSTM performance improved only 3.8 percentage points, confirming that the recurrent ceiling is architectural rather than data-limited.

Table 15: LSTM architecture evolution (Phase 2).

Configuration	**Profile Acc.**	**Modification**
Basic LSTM	2.8%	Random-level performance
+ Multi-pooling	14.8%	Max + mean pooling
+ Dimensionality reduction	21.0%	512→64 text embeddings
+ Bidirectional	22.8%	Forward + backward context
+ Decomposed heads	24.2%	Separate alignment/motivation heads
Best single run	30.3%*	*Non-reproducible outlier
Stable performance	∼24–25%	Consistent ceiling

Table 16: Impact of signal enhancement on LSTM performance (Phase 3).

Configuration	**Alignment**	**Motivation**	**Profile**
Baseline (30% encounters)	19.2%	99.7%	19.2%
High density (81% encounters)	23.0%	99.9%	23.0%
+ Explanatory queries	23.0%	99.9%	23.0%