GROUNDED COORDINATION IN LLM DIALOGUE: A REPEATED-MEASURES TEST FOR THE SHARED FRAME SUBSPACE

Anonymous authors
Paper under double-blind review

ABSTRACT

Opponent-shaping methods align agents' *behavior* without requiring an explicit representation of what a partner believes or intends beyond the reward function, leaving them brittle exactly where representational mismatch matters most: in novel situations requiring multi-agent coordination. We propose a representational account of grounding: a single activation space W factors into a shared *frame* subspace F, situational structure common to both agents, and agent-indexed *self*/*other* subspaces I and O, with a grounding registry G tracking which regions of O have survived communicative testing. We identify F as the subspace on which an orthogonal-Procrustes alignment residual between two agents' activations stays small. Across twelve instruction-tuned models, pooled activations show significance across episodes, but a within-episode-centered analysis, is significant in exactly one model, Qwen2.5-32B-Instruct. Treating turn as a repeated-measures factor for this model, the systematic turn-to-turn representational shift falls outside the identified frame, consistent with grounding-relevant change concentrating in the agent-indexed complement. A turn-stratified decoding probes the dialogue's disambiguating content across turns. We report these initial findings on a single controlled task and outline how this instrument extends toward certifying grounded coordination across populations of interacting AI agents.

1 INTRODUCTION

Opponent shaping has made real progress finding socially beneficial equilibria in general-sum games. This line of work includes LOLA (Foerster et al., 2018), SOS (Letcher et al., 2019), M-FOS (Lu et al., 2022), LOQA (Aghajohari et al., 2024), and most recently Advantage Alignment (Duque et al., 2025). Advantage Alignment reduces the machinery of earlier methods to a single modification of the advantage function. It scales to high-dimensional, partially observable, continuous-action settings via Proximal Advantage Alignment (PAA). These methods are *behavioral*: alignment is inferred from action consequences. Nothing in the formulation requires an agent to represent what its partner believes or intends beyond the shared reward function. Two agents can therefore be behaviorally aligned while representationally divergent. This configuration is stable in-distribution but brittle precisely where the environment shifts, which is the regime coordination guarantees exist to cover. Zero-shot coordination failures in Hanabi and Cicero's reliance on explicit intent modeling in Diplomacy are examples of this brittleness.

Theory-of-mind modeling in ML treats other-modeling as inference. This includes ToMnet (Rabinowitz et al., 2018), Bayesian ToM (Baker et al., 2009; 2017), cognitive-hierarchy models (Camerer et al., 2004), and LLM-era benchmarks such as FANToM, Hi-ToM, and MindGames (Sileo & Lernould, 2023).. These approaches typically lack a mechanism for *grounding*: establishing through interaction that mutual models have been tested and accepted (Clark & Brennan, 1991). We take grounding seriously as a representational claim. We ask whether it is measurable in trained sequence models without supervision, and whether the resulting instrument can serve a broader program of certifying grounded coordination across populations of interacting agents.

Contributions. (1) A grounding estimator that operationalizes the frame subspace F without labeling it a priori (§3), combining PCA-suppressed orthogonal-Procrustes alignment, a canonical-angle

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

spectrum, and the participation ratio (PR) of the residual. We select its dimensionality using a magnitude-based criterion (§3.3), which materially changes the answer compared to the correlation-based criterion typically used for this purpose. (2) A repeated-measures-aware test for whether F stays turn-invariant during an interaction (§3.4), motivated by a real statistical confound in the naive version of this test: pooling activations across turns to build a null baseline mixes in the *between-turn mean effect*, which is systematic and predictably deflates the baseline. (3) A second, independent measurement instrument (§3.5): a cross-validated logistic-regression probe of the dialogue's disambiguating epistemic content, stratified by turn with a content-independent control task and a multiple-comparisons correction. Stratifying by turn changes the reading of an otherwise strong pooled result. (4) An empirical test of H1 across twelve models spanning three architecture families and a 2B-to-32B parameter range (§5): pooled activations show apparent significance in a few models, while the within-episode-centered analysis, is significant in exactly one model, Qwen2.5-32B-Instruct; the turn-stratified probe (§5.5) decodes the disambiguating content well when pooled, but sits at chance on the one turn that requires genuine tracking, in every model tested. (5) A discussion connecting this instrument to a broader research program on certifying grounded coordination in multi-agent AI populations (§6).

2 Related Work

Opponent shaping. LOLA (Foerster et al., 2018), SOS (Letcher et al., 2019), M-FOS (Lu et al., 2022), LOQA (Aghajohari et al., 2024), and Advantage Alignment (Duque et al., 2025) find cooperative equilibria by shaping learning dynamics. None of them build an explicit other-model. We view our contribution as complementary: a representational diagnostic that could in principle explain *when* behavioral alignment is fragile.

Theory of mind in ML. ToMnet (Rabinowitz et al., 2018) and Bayesian ToM (Baker et al., 2009; 2017) treat other-modeling as inference over a partner's latent state. Cognitive-hierarchy models (Camerer et al., 2004) give a bounded-rationality account of recursive reasoning. None of these supply a grounding mechanism in Clark & Brennan's sense (Clark & Brennan, 1991): a record of which parts of a mutual model have survived communicative testing. This is the role our grounding registry G plays.

Multi-agent interpretability and collusion detection. Rose et al. (2026) train a supervised linear "deception direction" to detect adversarial collusion from activations, reporting 1.00 AUROC in-distribution but only 0.73–0.93 on structurally distinct transfer scenarios (weakest for their representation-geometry technique, mean pairwise cosine similarity, at 0.73). We interpret this gap as evidence that undifferentiated similarity conflates *shared situation* with *shared intent*, exactly the $F/I{\oplus}O$ split this paper targets. Our estimator is unsupervised, requiring no collusion labels; applies identically to cooperative and adversarial interactions; and, being differentiable, is reusable as a steering signal in addition to detection.

Belief-state geometry and computational mechanics. Shai et al. (2024) and Piotrowski et al. (2025) show transformers linearly represent the belief-state geometry of their training distribution's generative process in the residual stream, including nontrivial mixed-state geometries derivable a priori. Our frame-subspace hypothesis is the natural multi-agent extension: F is the shared component of the mixed-state geometry of an *interactive, communicative* generative process, with agent-indexed components as the remaining factors. We defer the a priori derivation to future work (see §7); the present results are empirical.

Mechanistic interpretability and frame semantics. Sparse-autoencoder work (Cunningham et al., 2023; Bricken et al., 2023; Templeton et al., 2024) recovers monosemantic directions. This rests on an implicit assumption that meaning is atomic and directional. Recent results (Bhalla et al., 2026) suggest concepts instead organize along low-dimensional manifolds. SAEs recover this structure only in a fragmented regime. These results identify geometric objects, above individual directions, as the right unit of analysis. We read this through frame semantics (Fillmore, 1982; Minsky, 1974): coherent co-activation manifolds correspond to frames. The present work's frame-subspace construct is the interactive extension of that claim.

3 FRAMEWORK

3.1 A UNIFIED REPRESENTATIONAL SPACE WITH A GROUNDING REGISTRY

We posit a single representational space W, factoring into a frame manifold F (situational structure common to both agents), the agent's own intentional state I relative to F, and the agent's model O of the partner's position within F. O is implemented by the same machinery as I, under a perspective transformation. Ours is a simulationist account of other-modeling sharing the same representational substrate.

G, the grounding registry, is not a fourth subspace. Alignment between an agent's other-model and the partner's actual state is not internally observable. From inside the agent, O is all there is. What is available is the *verification status* of O: which regions have been tested against communicative evidence and survived. G is therefore historical: a modulatory memory over communicative events rather than a geometric object in its own right. We formalize this with a linear-Gaussian factor model over the two agents' activations (Appendix A.1, following van Ditmarsch et al. (2008)), under which grounding events partition time steps into **agreement** (divergent latents momentarily aligned), **disagreement** (a prediction failure attributable to genuine divergence, with the frame remaining aligned), and **resolution** (a prior disagreement closes through communication) (Appendix A.2). This labeling supplies the contrastive structure that makes F, I, and O identifiable without external labels. Identifiability conditions and their sample-complexity bounds are described in Appendix A.3–A.8. We confirm the relevant condition is met at our sample size in §5.1.

3.2 HYPOTHESES TESTED IN THIS DRAFT

H1 (Frame subspace). W factors into a shared frame manifold F and discriminable agent-indexed subspaces I, O; grounding quality is measurable as alignment of agents' projections onto F. H1 is a factorization claim and is purely geometric.

H2 (Grounding as dimensionality reduction). Grounding converts free variables into fixed context; the representation itself does not get smaller but the dimensions contract. The predicted trajectory of the active-inference dimensionality is a sawtooth: contraction during grounded stretches, sharp re-expansion at frame-disagreement events, re-contraction as repair completes.

3.3 THE GROUNDING ESTIMATOR

Given paired activation samples from agents A and B, we compute a principal-subspace alignment. First, PCA to a threshold retaining 95% of variance suppresses the noise floor. Then an orthogonal Procrustes rotation R minimizes $\mathbb{E}\|Rx^A - x^B\|^2$. The residual $\delta = Rx^A - x^B$ is small on frame directions and large on divergent ones. We use Procrustes and generalized shape metrics (Williams et al., 2021) for the *magnitude*, not raw CCA. CCA's arbitrary invertible maps whiten the data and can manufacture correlation from low-variance noise, which is the wrong property for a dimensionality claim. We retain canonical correlations as a complementary, differently-normalized view.

Agents A and B read largely overlapping dialogue text, so a skeptical reading is that any low-rank shared subspace reflects shared template and position statistics, not genuine frame content. The permutation null (next paragraph) controls for this. Template structure is constant across *every* episode regardless of pairing, so it cancels out of the significance comparison, and the null isolates the alignment attributable to the paired episode's own content, which varies combinatorially under permutation.

Frame dimensionality. We identify the frame rank r_F by two criteria and take their intersection. The first is *significance*: the leading canonical correlations that clear a permutation-null threshold, built by re-pairing agents at random, following the generalized shape-metric literature's use of rotation and permutation nulls to calibrate an alignment threshold instead of fixing one by hand. The second is *magnitude*: a kneedle elbow detector (Satopaa et al., 2011) applied to the **raw Procrustes residual eigenvalues**, not the canonical correlations as canonical correlation is whitened and scale-free, so a nearly flat correlation spectrum can sit on top of a highly skewed magnitude spectrum. Magnitude separates components that correlation alone would treat as indistinguishable. Magnitude

is also the quantity the participation ratio Σ, is built on, so using it for dimension selection keeps the pipeline internally consistent.

Participation ratio and magnitude. $\mathrm{PR} = (\sum_i \lambda_i)^2 / \sum_i \lambda_i^2$ of the residual covariance is threshold-free, differentiable, and Schur-concave in the eigenvalue distribution. It tracks how *concentrated* the residual is, independent of its overall scale. We report it against a permutation null, built by re-pairing agents at random, with a bootstrap confidence interval computed at the episode level. Episode-level bootstrapping is necessary because repeated observations from the same dialogue are not independent samples. We declare PR significantly below null when the observed value falls more than two null standard deviations below the null mean ($z > 2$; one-sided, per-model threshold $\alpha \approx 0.023$ under a Gaussian approximation to the permutation null). This per-model threshold is uncorrected for the twelve-model family tested in §5.3. We therefore report exact z-scores wherever a positive call is made (Table 2), and address the resulting multiple-comparisons question directly in §5.3. PR is the primary readout for H1; it tells us whether there is a small, concentrated frame block at all, and it describes the *shape* of the divergent block. PR is exactly scale-invariant: $\mathrm{PR}(c\Sigma) = \mathrm{PR}(\Sigma)$ for any $c > 0$. This means it cannot register a contraction that shrinks several divergent eigenvalues by a comparable factor simultaneously. It can only register a change in their *relative* proportions. Even for a single grounded direction, PR's guarantee is conditional rather than universal: Appendix A.5 shows PR is guaranteed non-increasing only while the grounded eigenvalue stays at or below a size-weighted average of the others, so grounding the *dominant* divergent direction can transiently *increase* PR before it falls. Raw magnitude has no such gap: $\mathrm{tr}\,\mathcal{G}(\Sigma) \preceq \mathrm{tr}\,\Sigma$ unconditionally, in the Löwner (positive-semidefinite) order, for any grounding update and any number of directions contracting at once. We therefore use $\mathrm{tr}\,\Sigma = \mathbb{E}\|\delta\|^2$ as the primary readout for H2 (§3.4), and keep PR as a secondary concentration diagnostic.

3.4 Repeated-measures frame coordination

Once F is identified from a pooled fit (r_F dimensions of a k-dimensional retained space), we ask a further question: does F stay fixed as the interaction proceeds, or does it drift? A first attempt at this test compared, at each turn, the fraction of that turn's own residual variance falling inside F, against a null built by pooling residuals across *all* turns. This null is invalid. Turns differ systematically in content, so pooling mixes the *between-turn mean effect* into what was meant to be a chance baseline. This effect reflects real content differences between turns, and it predictably concentrates outside F. The pooled null is therefore deflated relative to any single turn's own value, and the test flags instability regardless of whether anything is wrong.

The corrected test treats turn as a repeated-measures factor directly. For each turn t we compute the *condition mean* residual $\mu_t = \mathrm{mean}(\delta \mid \mathrm{turn} = t)$ across episodes, then decompose the variance of these T condition means around their grand mean into the fraction lying in F versus its complement:

$$\text{mean_shift_in_}F \;=\; \frac{\sum_t \|(\mu_t - \bar{\mu})\,F\|^2}{\sum_t \|\mu_t - \bar{\mu}\|^2}. \tag{1}$$

This isolates the *systematic* turn-to-turn shift, the repeated-measures "condition" effect, from within-turn, across-episode noise. The naive test above conflated that noise with the effect. We compare this fraction to a closed-form reference, rather than a simulated null. A uniformly random r_F-dimensional subspace of a k-dimensional space captures exactly r_F/k of any fixed vector's squared norm in expectation. This reference carries no dependence on row grouping, so it carries no repeated-measures confound of its own. Confidence intervals are obtained by resampling episode identity, not rows, with replacement. We apply this resampling consistently across all turns within a replicate, to preserve the repeated-measures structure.

3.5 A second instrument: turn-stratified content decoding

The estimator above determines whether a shared frame exists and stays fixed. It says nothing about whether the model's activations encode the specific fact each turn's query is about, independent of whether the model's own forced-choice output reflects it. We test this with a cross-validated logistic-regression probe trained to decode the turn's ground-truth answer directly from saved hidden states, bypassing the model's emitted prediction entirely.

For each (agent, position, layer) combination, we fit an ℓ_2-regularized logistic regression on PCA-reduced activations (95% variance retained), cross-validated with scenario-grouped K-fold splits so that a scenario's own turns never span train and test. Director and Follower activations are always fit as fully independent classifiers, never pooled into one training set; both roles' probes target the same label, whether the Follower actually knows the referent, so the Follower's own activations are probed for self-knowledge (I) and the Director's for a model of the Follower's state (O), directly comparable rather than an artifact of mixing the two roles (Appendix B checks this split explicitly). We report held-out AUC against a permutation null built by reassigning each scenario's own per-turn label sequence to a different, randomly chosen scenario (matched turn-for-turn), which preserves each turn's marginal label distribution and isolates the null from the raw turn-index confound.

Turns in this task are not equally informative: two of four turns have constant ground truth, and the final turn states the resolution outright, so pooling across turns can make a probe look informative even if it has decoded nothing more than turn position. We therefore report both the pooled AUC and the AUC restricted to the one turn that requires the listener to actually track the disambiguating content, using the same permutation-null construction within that turn alone.

Two further checks address standard concerns about probing classifiers. First, a control task (following Hewitt & Liang (2019)) fits the identical pipeline to a label that is fixed per scenario, independent of scenario content, and matched to the true label's base rate. We report selectivity as the gap between the real and control-task AUC, which bounds how much of the observed decoding could be attributable to probe capacity alone. Second, we test 8–16 (agent, position, layer) combinations per model, so we apply Benjamini-Hochberg false-discovery-rate correction across all combinations tested for a given model. We do not treat each combination's permutation test as independent.

4 Experimental Setup

Testbed. We use a controlled director–follower mirror task, extending the reference-game paradigm of Keysar et al. (2000), as a minimal environment satisfying the grounding-event-profile requirement of §3. The shared workspace contains a single, unambiguous instance of the target item. The Follower additionally holds a same-type distractor in a private area, visible only to them and unknown to the Director. The Director's instruction is therefore fully determinate from the Director's own perspective, but ambiguous to the Follower, because the Follower must recognize their private item is not in common ground. This is genuine perspective-taking, matching Keysar et al.'s actual manipulation, and not simply symmetric lexical disambiguation. The Follower flags the ambiguity and asks the Director to confirm which item is meant (a scripted *disagree* event). In most episodes the Director confirms by reference to their own limited visibility (*resolve*), and the Follower confirms in turn (*agree*). In the remaining episodes, about 39%, the Director's clarification does not resolve the ambiguity, and the Follower ends the episode still uncertain; both realized branches, resolved and unresolved, are therefore present in the dataset by construction, not only through model error. Each of the 8 item types, 8 colors, and 5 goal verbs is combinatorially sampled, giving an effectively unlimited pool of surface-distinct scenarios that share one underlying grounding structure. We generated 2000 four-turn episodes with 8000 agent-turn activation samples.

Models and activation extraction. We probe twelve instruction-tuned models spanning three architecture families (Qwen2.5, Gemma-2/4, Mistral) and a 2B-to-32B parameter range (bf16), listed in Table 3. We report the H1 estimator (§5.2) and repeated-measures test (§5.4) in depth for Qwen2.5-32B-Instruct, the one model where centered significance coincides with clean identifiability. We report the generalization result (§5.3) across all twelve. At each turn t, each agent role is prompted with the dialogue so far plus a query ("does the Follower know which item?"), and we extract the last-token hidden state at the final transformer layer. This gives paired activation trajectories x_t^A, x_t^B for every episode and turn, for each model independently. We fix the final layer for the H1 estimator (§3.3) a priori, since it is the representation closest to the read-out the grounding-relevant query is actually answered from. The turn-stratified probe (§3.5) instead sweeps layer and token position, because it asks a different question, that is, whether disambiguating content is encoded *anywhere* in the network, independent of whether the final layer's read-out reflects it. A pooled-analysis layer sweep confirms the final-layer choice is not arbitrary (Appendix B gives the per-layer statistics).

Realized grounding events. Event labels are assigned exogenously, from behavioral correctness. Essentially, does the role's predicted answer match ground truth at that turn? They are not assigned

from the scripted structure. This follows the identifiability requirement that disagreement labels be exogenous to the current representational estimate (§3), which is what excludes the degenerate "everything is shared" fixed point in the identification procedure. Both roles are queried about the same state, whether the Follower can resolve the instruction, so both are graded against that state.

Preprocessing: per-episode centering. Pooling activations across all 2000 episodes conflates two sources of variance: which specific scenario this is (item/color/goal-verb content, which differs across nearly every episode) and where in the dialogue we are (which is the quantity relevant to grounding dynamics). We remove the former with a standard panel-data "within" transform: subtract each episode's own mean (across its four turns) from its rows before any alignment step. This necessarily also removes any frame signal that is *constant* across turns within an episode, so we report both the pooled and the within-episode-centered analyses. We treat them as answering different questions as neither supersedes the other.

Why centered is the primary test. Pooled activations are dominated by which scenario an episode is, instead of the dialogue-internal structure we are interested in. With 2000 combinatorially distinct scenarios and only 4 turns each, cross-episode content variance occupies most of the retained variance and masks any grounding-relevant signal. Pooled significance is an unreliable indicator for a second reason. Across the twelve models we test (§5.3), every case of pooled significance coincides with a shortfall in our own identifiability conditions (§5.1). This is consistent with pooled analysis picking up something other than a genuine frame when the labeled-contrast sample is too small. We treat pooled results as a methodological comparison point, reported throughout for transparency. We treat the within-episode-centered analysis as the test that speaks to H1, since it avoids both confounds.

5 RESULTS

5.1 SAMPLE SIZE AND IDENTIFIABILITY

The formalization's sample-complexity bound for certifying the $F/I\oplus O$ split scales with effective dimension and inverse-square separation margin. Realized event counts depend on the model, since agreement, disagreement, and resolution labels come from each model's own behavioral correctness (§4). They do not come from the scripted structure. For Qwen2.5-32B-Instruct, the model we report H1 results for in depth below, N=2000 episodes (8000 samples) give n_{agree}=4051, n_{disagree}=2063, n_{resolve}=1886. This clears a bound requiring $\geq 1800/1800/1000$ at this model's effective dimension (r_{eff}=15): all three conditions pass comfortably. This does not hold for every model we test. Eleven of the other twelve show a shortfall on at least one condition. The most common shortfall is n_{resolve} (C3): resolution events are the scarcest event class across models, consistent with §A.8's observation that the resolution margin ζ is typically the weakest of the three. We report this per model, and its bearing on how to read pooled significance, in §5.3 and §7.

5.2 H1: NO FRAME POOLED OR CENTERED IN MOST MODELS EXCEPT FOR ONE

For Qwen2.5-32B-Instruct, the one model in our twelve-model test (§5.3) where centered significance coincides with all three identifiability conditions passing. Both the pooled and within-episode-centered analyses support H1: PR sits several standard deviations below its permutation null in both configurations ($z = 3.31$ pooled, $z = 3.24$ centered; Appendix B Table 2 and Figure 2 give the full numbers and the canonical-correlation spectrum). The residual eigenvalue spectrum shows a clear elbow at the reported frame dimension in both configurations. A small number of components carry most of the residual mass. This is a stronger and cleaner result than the pooled-analysis failure we see for every other model we test (§5.3): Qwen2.5-32B-Instruct shows a significant frame whether or not per-episode content variance is removed. We return to why this model in particular shows this in §5.3.

5.3 GENERALIZATION ACROSS TWELVE MODELS

We repeat the H1 test on twelve instruction-tuned models spanning three architecture families (Qwen2.5, Gemma-2/4, Mistral) and a 2B-to-32B parameter range, all under the identical pipeline and dataset. Appendix B Table 3 reports centered PR significance, since §4 establishes this as the

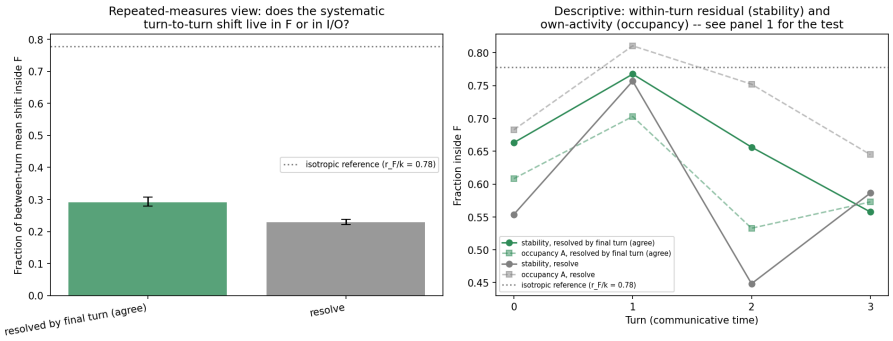

Figure 1: Between-turn mean shift inside F vs. the isotropic reference, Qwen2.5-32B-Instruct.

test that speaks to H1, alongside whether each model's realized event counts meet our identifiability conditions. Under the centered test, the effect replicates in exactly one of twelve models.

Multiple comparisons across the model sweep. The per-model significance call in Table 3 uses the $z > 2$ threshold of §3.3 (one-sided $\alpha \approx 0.023$). This threshold is uncorrected for testing twelve models. At this per-model rate, at least one spurious "Yes" among twelve independent tests is expected with probability $1 - (1 - 0.023)^{12} \approx 24\%$ under the global null. A single positive call at $z \approx 2$ would therefore not, by itself, be distinguishable from a false positive at this family size. Qwen2.5-32B-Instruct's result does not rest on being just over that line. Its z-scores of 3.31 (pooled) and 3.24 (centered) (Table 2) clear the Bonferroni-corrected threshold for a twelve-model family ($\alpha/12 \approx 0.0042$, $z \approx 2.64$) with margin, in both the pooled and centered analyses independently. We flag this explicitly instead of relying on the binary Yes/No column alone. gemma-4-31B-it's "Yes" in the same table is exactly the kind of call the uncorrected threshold cannot be trusted on: it fails all three identifiability conditions, while Qwen2.5-32B-Instruct passes all three cleanly (§5.1).

We do not have a confirmed account of why this one model shows the effect and the other eleven do not, and the sweep does not let us cleanly rule out scale or family. Qwen2.5-14B, gemma-2-27b, and Qwen2.5-3B all fail alongside Qwen2.5-32B-Instruct's success, but every one of them also fails at least one identifiability condition (Table 3), so their failure is already explained by insufficient labeled contrast. A capability-threshold account remains a live alternative to treating Qwen2.5-32B-Instruct as an outlier: ToM-relevant reasoning in LLMs shows threshold effects concentrated at the largest, most capable models (Kosinski, 2023; Sileo & Lernould, 2023), and the only within-family comparison available is consistent with this: only the largest model in that family, Qwen2.5-32B-Instruct, succeeds. We return to this as an open question in §6.

5.4 Repeated-measures test: the frame holds across all four tested combinations

Table 1 treats turn as a repeated-measures factor and reports what fraction of the systematic turn-to-turn representational shift falls inside F, for Qwen2.5-32B-Instruct's two realized branches (§3.4). Figure 1 plots the centered branches against the isotropic reference.

Table 1: Repeated-measures frame-coordination result, Qwen2.5-32B-Instruct.

Branch	Config	Mean shift in F	95% CI	Isotropic ref.	Outside F (chance)
Agree	Pooled	0.386	(0.372, 0.407)	0.867	61.4% (13.3%)
Agree	Centered	0.290	(0.279, 0.307)	0.778	71.0% (22.2%)
Resolve	Pooled	0.223	(0.216, 0.231)	0.867	77.7% (13.3%)
Resolve	Centered	0.230	(0.222, 0.238)	0.778	77.0% (22.2%)

In all four branch/configuration combinations, the systematic shift falls outside F substantially more often than the isotropic reference. The observed outside-F fraction ($1-$ mean shift in F) ranges from

61.4% to 77.7%, against a chance level of only 13.3% (pooled, $r_F/k = 13/15$) or 22.2% (centered, $r_F/k = 7/9$). This is a 3–4$\times$ excess over chance in every one of the four cases and there is no near-chance outlier among them. This is consistent with H1's prediction that grounding-relevant change concentrates outside the frame, uniformly across both realized branches and both pooling configurations. The outside-F fraction (61–78%) is well short of unity, however. This raises a genuine question, taken up in §6 and §7: how sharply does the Procrustes-identified F isolate the theoretically-intended frame subspace? The alternative is a subspace that is disproportionately, but not exclusively, frame-stable. This dataset does not meet the assumptions our pre-registered H2 contraction test requires (§7), which is why we do not report an H2 result herein.

5.5 Turn-stratified content decoding

The estimator above allows us to determine whether change is inside or outside a shared frame. The probe in §3.5 estimates whether the model's activations linearly encode the specific fact each turn's query is about, whether or not F captures it. Two of the four turns in each episode have constant ground truth (trivially decodable), and the final turn states the resolution outright (surface-readable). Only the second turn requires the listener to actually track the disambiguating content across the dialogue.

Across all twelve models, pooling over turns gives strong decoding: held-out AUC of 0.86–0.97 across (agent, position, layer) combinations, significant after Benjamini-Hochberg correction in effectively every combination tested. Restricted to the second turn alone, decoding is indistinguishable from chance in every model: AUC ranges from 0.48 to 0.54 across all combinations and models, with at most 2 of 16 combinations reaching significance for any single model, consistent with the false-positive rate expected under our correction (Appendix B, Table 4 gives the full per-model breakdown). Control-task selectivity (Hewitt & Liang (2019); §3.5) is positive and substantial throughout (mean 0.46–0.48), ruling out probe capacity as the explanation for the pooled result.

The pooled result, taken alone, reads as strong evidence that these models represent the disambiguating content well. Turn-stratified, it shows that this evidence is carried almost entirely by turns where the answer is trivial or already stated on the surface, and that no model we test shows reliable decoding at the one turn where tracking the content actually matters.

Behavioral accuracy at the same turn. Event labels are already assigned from each model's own behavioral correctness (§4), so the pipeline logs each model's own predicted answer at every turn. This lets us ask the companion question directly: do these models get turn 2 right *behaviorally*, independent of whether a probe can decode it from activations? Across all twelve models, the Director role's turn-2 prediction of whether the Follower knows the referent never exceeds the scripted base rate of always predicting "yes" (60.5%). Observed accuracy ranges from 39.5% to 60.5% (mean 52.6%): at or below the trivial baseline in every model, including Qwen2.5-32B-Instruct (54.2%). The turn-2 probe null is therefore consistent with behavioral failure at the same turn, not an unexplained dissociation from it: none of the twelve models we test reliably track the disambiguating content at turn 2 by either measure.

6 Discussion: Scaling and Multi-Agent Coordination

The present result is a two-agent, single-task finding. We test it across twelve models but confirm it cleanly in only one. We want to be direct about what would make it a stronger claim, so the framing does not outrun the evidence.

Is the identified frame the intended one? H1's PR test confirms that Qwen2.5-32B-Instruct's alignment residual compresses into a small number of dominant directions, well below what a random rotation would produce. This is a real, well-supported geometric result. The repeated-measures test in §5.4 answers a further, semantic question: is this particular subspace the one where grounding-relevant stability actually lives? The direction of the answer is consistent (all four branch/configuration combinations place the shift outside the frame at 3–4$\times$ the chance rate, with no near-chance outlier), but the magnitude is only partial: 61–78% of the shift falls outside F, well short of the near-totality a fully privileged frame subspace would predict, so a nontrivial share still falls *inside* F in every case. One reading is that F, as identified by a purely magnitude-based, linear alignment criterion, only partially recovers the semantically privileged subspace the theory posits.

Perhaps this is because activation directions are in superposition rather than one-fact-per-direction (§2's discussion of sparse-autoencoder anomalies and concept manifolds), so an orthogonal Procrustes alignment on raw coordinates may not isolate a semantically coherent frame direction even when it correctly detects that *some* low-dimensional residual structure exists. This is an open question for follow-up, with a nonlinear or sparse-coding-based alternative to Procrustes alignment as the natural next step.

Beyond alignment measurement, a frame-factored, population-scale version of this estimator affords the representational substrate for inquiry into whether multi-agent coordination exceeds what shared conditions explain. This underlies both collusion detection and the diagnosis of confident-but-wrong mutual coordination, "stable miscoordination," that current supervised collusion probes cannot cover, since it has no anomaly signature. This paper contributes, a repeated-measures frame-analytic approach to the study of multi-agent coordination.

7 LIMITATIONS

Several limitations qualify these results. We test H1 across twelve models but confirm it in only one, Qwen2.5-32B-Instruct; the repeated-measures test and the turn-stratified probe deep-dive are reported for one confirming model, on one synthetic task. Identifiability shortfalls are the norm rather than the exception: eleven of twelve models fail at least one condition, most often too few resolution events (Appendix B gives the per-model breakdown). Whether the identified F is the semantically intended frame, not merely a disproportionately frame-stable subspace, is not fully settled (§6); the repeated-measures separation from chance holds directionally in all four combinations, but the outside-frame fraction is short of unity, leaving the semantic claim on weaker footing than the geometric one. H2 is untested: the turn-schedule our pre-registered contraction test assumes does not hold in the realized event sequences for any model, so the test's own guard declines to produce a number rather than one computed on the wrong turn offsets.

H1 and the repeated-measures test are demonstrated for two agents only; an n-agent extension and comparison of architectures are the intended next steps. We motivate F via the belief-state-geometry literature (§2) but do not derive its predicted geometry a priori, so H1's support here is empirical rather than a confirmed first-principles prediction. H1's layer sweep is pooled-only and for one model: the effect is specific to the final layer tested and not significant three layers earlier (Appendix B), and we have not extended it to the centered analysis or the other eleven models. Finally, behavioral accuracy at turn 2 is at or below the trivial base rate in all twelve models, alongside the turn-2 probe null, consistent with these models simply not tracking the content at turn 2 by either measure.

8 CONCLUSION

We give an unsupervised, differentiable estimator for a shared representational "frame" between interacting LLM agents. Tested across twelve models, a shared frame is detectable, on the valid within-episode-centered test, in exactly one, where the repeated-measures test finds systematic turn-to-turn change falling outside the frame more often than chance in all four tested combinations, though the outside-frame fraction (61–78%) is well short of unity. A second, independent instrument, a turn-stratified decoding probe, is well-decoded when pooled but at chance on the one turn requiring actual tracking, across all twelve models. These are initial, narrow, but real results toward certifying grounded coordination in AI populations.

AI USE STATEMENT

We used a large language model assistant (Claude, Anthropic) to help draft and revise text in this manuscript, and to implement, debug, and iterate on the analysis code underlying the grounding estimator, repeated-measures test, and turn-stratified probe described in §3.3–§3.5. All reported numerical results were produced by running this code against saved model activations. The authors reviewed the implementation, reran the analysis, and independently checked the reported statistics (Tables 2–4) against the underlying output files before including them here. We did not use genera-

tive AI to generate or fabricate experimental data. We take responsibility for the final content of this work, including text, claims, and code artifacts produced with the aid of generative AI.

Reproducibility statement

The grounding estimator is fully specified in §3.3: the PCA threshold, orthogonal Procrustes alignment, canonical-correlation and eigenvalue-based frame-dimension selection, permutation null, and episode-level bootstrap. The repeated-measures test and its closed-form isotropic reference are specified in §3.4, Equation 1. The turn-stratified probe is specified in §3.5: cross-validated logistic regression, scenario-block permutation null, control-task selectivity, and Benjamini-Hochberg correction. The testbed generation procedure (item/color/goal-verb combinatorics, four-turn scripted structure, exogenous event labeling) is described in §4. The identifiability sample-complexity bound used to validate our sample size (§5.1) is derived in the accompanying formalization (Appendix A) and evaluated per model in §5.3. All twelve models tested (§5.3, Table 3) are publicly available on the Hugging Face Hub. The grounding estimator, repeated-measures test, and turn-stratified probe code, the testbed generator, and the exact 2000-episode dialogue dataset used for every model in Table 3 are included as anonymous supplementary material. The large per-model, per-layer activation arrays the estimator operates on are not included. They are mechanically regenerable from the included code, dataset, and the publicly available model weights (supplementary README).

References

Milad Aghajohari, Juan Agustin Duque, Tim Cooijmans, and Aaron Courville. Loqa: Learning with opponent q-learning awareness. *arXiv preprint arXiv:2405.01035*, 2024.

Chris L. Baker, Joshua B. Tenenbaum, and Rebecca R. Saxe. Action understanding as inverse planning. *Cognition*, 113(3):329–349, 2009.

Chris L. Baker, Julian Jara-Ettinger, Rebecca Saxe, and Joshua B. Tenenbaum. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4): 1–10, 2017.

U. Bhalla et al. Do sparse autoencoders capture concept manifolds? *arXiv preprint arXiv:2604.28119*, 2026.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards monosemanticity: Decomposing language models with dictionary learning. Transformer Circuits Thread, 2023.

Colin F. Camerer, Teck-Hua Ho, and Juin-Kuan Chong. A cognitive hierarchy model of games. *Quarterly Journal of Economics*, 119(3):861–898, 2004.

Herbert H. Clark and Susan E. Brennan. *Grounding in Communication*. American Psychological Association, 1991.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Juan Agustin Duque, Daniel Morales-Brotons, Saul Tan, Milad Aghajohari, Samuel Sokota, Yash Chandak, Jakob Foerster, Sepehr Zaman, et al. Advantage alignment algorithms. In *International Conference on Learning Representations (ICLR)*, 2025.

Charles J. Fillmore. Frame semantics. In *Linguistics in the Morning Calm*, pp. 111–137. Hanshin Publishing, 1982.

Jakob Foerster, Richard Y. Chen, Maruan Al-Shedivat, Shimon Whiteson, Pieter Abbeel, and Igor Mordatch. Learning with opponent-learning awareness. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2018.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019.

Boaz Keysar, Dale J. Barr, Jennifer A. Balin, and Jason S. Brauner. Taking perspective in conversation: the role of mutual knowledge in comprehension. *Psychological Science*, 11(1):32–38, 2000.

Michal Kosinski. Evaluating large language models in theory of mind tasks. *arXiv preprint arXiv:2302.02083*, 2023.

Alistair Letcher, Jakob Foerster, David Balduzzi, Tim Rocktäschel, and Shimon Whiteson. Stable opponent shaping in differentiable games. In *International Conference on Learning Representations (ICLR)*, 2019.

Christopher Lu, Timon Willi, Alistair Letcher, and Jakob Foerster. Model-free opponent shaping. In *International Conference on Machine Learning (ICML)*, 2022.

Marvin Minsky. A framework for representing knowledge, June 1974. URL `http://hdl.handle.net/1721.1/6089`.

Mateusz Piotrowski, Adam Rubié, Daniel Filan, and Adam S. Shai. Constrained belief updates explain geometric structures in transformer representations. *arXiv preprint arXiv:2502.01954*, 2025.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International Conference on Machine Learning (ICML)*, 2018.

A. Rose et al. Detecting multi-agent collusion through multi-agent interpretability. *arXiv preprint arXiv:2604.01151*, 2026.

Ville Satopaa, Jeannie Albrecht, David Irwin, and Barath Raghavan. Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *IEEE 31st International Conference on Distributed Computing Systems Workshops*, 2011.

Adam S. Shai, Sarah E. Piotrowski, Daniel Filan, Stephen Casper, Jakob Foerster, Róbert Csordás, Sarah Marzen, and Paul M. Riechers. Transformers represent belief state geometry in their residual stream. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Damien Sileo and Antoine Lernould. Mindgames: Targeting theory of mind in large language models with dynamic epistemic modal logic. *arXiv preprint arXiv:2305.03353*, 2023.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. Transformer Circuits Thread, 2024.

Hans van Ditmarsch, Wiebe van der Hoek, and Barteld Kooi. *Dynamic Epistemic Logic*. Springer, 2008.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

A Formalization

This appendix gives the identifiability results and the sample-complexity bound that §3 and §5.1 rely on, at the population (infinite-sample) level, with proof sketches, rather than full proofs, in the interest of space. The register here is "theorem with sketch." This is sufficient to make every quantity used in the main text, r_eff, the event-count thresholds of Table 3's C1/C2/C3 columns, and the isotropic reference of Eq. 1, traceable to an explicit assumption.

A.1 SETUP

Two agents A, B each emit an activation $x_t^A, x_t^B \in \mathbb{R}^d$ at interaction step t. We posit the linear-Gaussian factor model

$$x_t^A = U_F f_t + U_I a_t + U_O \hat{b}_t + \varepsilon_t^A, \tag{2}$$

$$x_t^B = V_F f_t + V_I b_t + V_O \hat{a}_t + \varepsilon_t^B, \tag{3}$$

with f_t the shared frame latent, a_t/b_t each agent's own intention, $\hat{b}_t/\hat{a}_t$ each agent's model of the other, and column-orthonormal loadings $U_\bullet, V_\bullet$ and isotropic noise $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2 I)$. Per-agent subspaces are defined by their loadings: $F_A = \mathrm{span}(U_F)$, $I_A = \mathrm{span}(U_I)$, $O_A = \mathrm{span}(U_O)$, and symmetrically for B. The identification target is the partition (F_A, I_A, O_A), up to within-block rotation. This within-block rotation is the model's irreducible gauge freedom; it is not our concern. Two latents are *aligned* on a step when they carry the same value across agents. Frame grounding is f_t common to both rows. Cross-role (external) grounding is $\hat{b}_t = b_t$ and $\hat{a}_t = a_t$. The experimenter observes both rows. An agent observes only its own row, plus the partner's messages. For paired samples, $R \in O(d)$ is the orthogonal Procrustes map minimizing $\mathbb{E}\|Rx^A - x^B\|^2$, and $\delta_t := Rx_t^A - x_t^B$. On a frame direction $u \in F$, $u^\top \delta_t = 0$ in expectation. On a divergent direction it is nonzero whenever the corresponding latents differ.

A.2 GROUNDING EVENTS AS CONDITIONING SETS

Each step carries a label $e(t) \in \{\text{agree}, \text{disagree}, \text{resolve}, \emptyset\}$, operationally assigned from communicative outcomes (§4). Here we take the labels as given and ask what they identify. **Agreement** ($e = \text{agree}$): divergent latents are momentarily aligned, so δ_t is small on all active directions. Agreement data alone cannot separate F from $I \oplus O$. **Disagreement** ($e = \text{disagree}$): a prediction failure attributable to divergence, so δ_t is large precisely on $I \oplus O$. The frame remains aligned, because the failure is attributable to divergence, not frame mismatch. (Frame-level disagreement is a distinct, separately detected event; see Remark A.4.) **Resolution** ($e = \text{resolve}$): a step where a prior disagreement closes through communication, and divergent components of A's representation move toward B's actual state. Define $C_e := \mathbb{E}[x^A (x^B)^\top \mid e]$ and $M_e := \mathbb{E}[\delta\delta^\top \mid e]$.

A.3 THEOREM 1: POPULATION IDENTIFIABILITY OF F

Theorem 1. Suppose the grounding-event profile satisfies two conditions. **(C1)** Agreement events occur, and their active cross-correlation subspace spans F (anchor). **(C2)** Disagreement events occur and load on $I \oplus O$ with positive variance: M_{disagree} restricted to $I \oplus O$ is nonzero on a subspace of full block dimension. Then F is identified, up to within-block rotation, as the subspace on which the canonical correlations of C_{disagree} remain ≈ 1. This is equivalently the active common kernel of M_{disagree}. Conversely, if (C2) fails, F and $I \oplus O$ are *not* separately identified. The all-shared assignment $F' = F \oplus I \oplus O$, $I' = O' = \emptyset$ fits agreement-only data exactly.

Proof sketch. Frame directions carry the same latent f_t in both rows. Under the optimal R they therefore satisfy $u^\top \delta = 0$ on every event class, including disagreement, so $F \subseteq \ker M_{\text{disagree}}$ and retains canonical correlation 1 in C_{disagree}. Divergent directions carry distinct latents under disagreement, by (C2); they show canonical correlation < 1 and nonzero mass in M_{disagree}, which excludes them from F. Restricting to the active subspace (eigenvalues above the noise floor σ_ε^2, estimated via the random-rotation null of §A.7) removes the ε-directions. These are uncorrelated across agents and drop out of C_{disagree} regardless. The converse is immediate. With only agreement data, M is small on all active directions, so the assignment to F versus $I \oplus O$ is unconstrained on the active subspace, and the all-shared solution is a global optimum. ■

This locates the degeneracy the main text's grounding registry G is built to avoid. The all-shared assignment is admissible exactly when (C2) is violated. Disagreement events are what make "shared" falsifiable: a negotiation is diagnostic only if its grounding dynamics are rich enough. *Finite-sample caveat:* C_{disagree} is estimated on the disagreement subsample, which may be small. Spurious canonical correlations scale with d/n_{disagree}. This is why we use SVCCA-style truncation and shrinkage, and a rotation null for the correlation-1 threshold, instead of fixing one by hand (§3.3).

A.4 THEOREM 2: SEPARATION OF I FROM O, AND EXCLUSION OF THE SPURIOUS FIXED POINT

Theorem A.3 delivers F and the divergent block $I \oplus O$, but not the split within it. It also does not address the practical fact that event labels are themselves estimated from the subspaces they help identify. Both are the content of Theorem 2.

Theorem 2(a): within-block separation. Add **(C3)**: resolution events occur, and across each, the change in A's divergent-block representation has positive projection onto B's intention subspace I_B. Then within $I_A \oplus O_A$, the other-component O_A is identified as the subspace whose resolution-induced motion aligns with I_B: the directions that update toward the partner. I_A is the orthogonal complement within the block. Internally, an agent substitutes B's messages for I_B as the orientation target. Externally, the experimenter uses I_B directly. This is the internal/external grounding distinction of §3.

Theorem 2(b–c): the bootstrap's fixed-point structure. Consider the alternating map $\mathcal{T}$: subspace estimate $\rightarrow$ event labels $\rightarrow$ refined subspace estimate, with labels assigned by thresholding the current δ against the current divergent-subspace estimate. The true decomposition is a fixed point of $\mathcal{T}$. *(b)* The all-shared decomposition is *also* a fixed point under purely **endogenous** labeling, where labels are derived only from the current estimate. In that case $F' =$ everything, so $\delta \equiv 0$, so every event is labelled agree, so no disagreement signal ever revises F'. *(c)* Suppose instead that disagreement labels are **exogenous**: supplied by behavioral prediction failure, observable independently of the current subspace estimate. This is exactly how §4 assigns event labels, from behavioral correctness, not from the scripted structure. Then the all-shared decomposition is no longer a fixed point. Exogenous disagreement events carry nonzero M_{disagree} mass, which by Theorem A.3 forces a nontrivial $F \subsetneq W$. Under (C1–C3), exogenous labeling, and positive initial overlap with the true divergent block, $\mathcal{T}$ converges to the true decomposition.

Proof sketch. (a) By construction, $\hat{b}_t$ (A's other-latent) is designed to track b_t. Under a resolution event, A's posterior moves $\hat{b}$ toward the partner's revealed state. A's own intention a_t is not updated to match the partner, so I_A shows no such alignment. Projecting the resolution displacement onto I_B therefore separates the block. (b) True-decomposition stationarity holds because labels generated from the true subspaces reproduce the event classes that re-estimate the same subspaces. All-shared stationarity is the degenerate computation above. (c) Exogenous labels break the self-confirming loop. A disagreement event labelled by behavior persists even when the current estimate predicts $\delta \equiv 0$. This injects M_{disagree} mass and forces $F \subsetneq W$ at the next iteration, so all-shared is not stationary. Convergence under positive initial overlap follows from monotone ascent of the Procrustes-aligned likelihood on the now-identified blocks, a CEM argument. Global convergence is not claimed; see the caveat below. ■

Why exogenous labeling is necessary. An agent reasoning only from its own representation can sit forever in the all-shared fixed point, confidently but stably miscoordinated. What dislodges it is grounding exogenous to the representation: behavioral failure, or the partner's surprising act. This is the formal counterpart of the "high internal G, low external grounding = stable miscoordination" hazard. It is also the reason §4 insists disagreement labels come from behavioral correctness, not from the scripted turn structure.

Finite-sample caveat. $\mathcal{T}$ is a hard-assignment (classification-EM) procedure, and it inherits CEM's susceptibility to local optima. Positive initial overlap is assumed, not guaranteed; in practice we mitigate this by seeding from a small set of independently (exogenously) labelled events. The convergence claim in (c) currently covers only two things: that the true decomposition is an ascent fixed point, and that the all-shared one is excluded. It does not establish global convergence, or characterize the true fixed point's basin as a function of the margins γ, ζ of §A.8. Hardening this remains open.

Remark (frame disagreement vs. intention disagreement). §A.2 assumes disagreement leaves the frame aligned. Frame-level disagreement, where agents are in different frames, is a distinct event. It is detected when δ has large mass *inside* the current F estimate: alignment fails on directions previously certified shared. This triggers frame renegotiation. F is provisionally demoted and re-identified.

A.5 THE SOFT-UPDATE OPERATOR AND ITS HARD LIMIT

Grounding along a direction u with informativeness governed by observation noise $\sigma^2 \in (0, \infty]$ acts on the divergent-block covariance $\Sigma \succeq 0$ by the Kalman contraction

$$\mathcal{G}_{u,\sigma^2}(\Sigma) = \Sigma - \Sigma u(u^\top \Sigma u + \sigma^2)^{-1} u^\top \Sigma. \tag{4}$$

Proposition. *(i) Monotone contraction.* For all $\sigma^2 \in (0, \infty)$, $\mathcal{G}_{u,\sigma^2}(\Sigma) \preceq \Sigma$ in the Löwner order (the partial order on symmetric matrices where $A \preceq B$ means $B - A$ is positive semi-definite), with equality iff $\Sigma u = 0$. Consequently $\operatorname{tr}\Sigma$ is non-increasing unconditionally, for any u and any composition of grounding directions, since trace is linear and non-negative on PSD matrices. PR does not inherit this unconditionally. Take u to be an eigenvector of Σ with eigenvalue λ_u, so the update rescales λ_u alone and leaves the other eigenvalues $\{\lambda_i\}_{i \neq u}$ fixed: then PR is non-increasing as λ_u shrinks if and only if

$$\lambda_u \;\leq\; \frac{\sum_{i \neq u} \lambda_i^2}{\sum_{i \neq u} \lambda_i}, \tag{5}$$

the size-weighted average of the untouched eigenvalues. This holds automatically once λ_u has shrunk below that threshold, and throughout the shrinkage whenever the grounded direction starts out subordinate to the rest of the residual block; if instead λ_u starts *above* the threshold (the grounded direction is the dominant divergent eigenvalue), grounding it can transiently *increase* PR before the threshold is crossed and PR turns over. See the Remark below. *(ii) Soft = partial, reversible.* The eigenvalue along u shrinks by $\sigma^2/(u^\top \Sigma u + \sigma^2) \in (0, 1)$. A disagreement event re-inflates it, giving the sawtooth of H2. *(iii) Hard limit = rank drop.* As $\sigma^2 \to 0^+$, $\mathcal{G}_{u,0}(\Sigma) = (I - \Pi_u)\Sigma(I - \Pi_u)^\top$ on the relevant block. The u-eigenvalue vanishes and rank drops by one. This is exactly the hard public-announcement (state-eliminating) operator. *(iv) rank-at-τ as the shadow of PR.* For threshold $\tau > 0$, the grounded eigenvalue falls below τ once $\sigma^2 < u^\top \Sigma u \cdot (\lambda_u - \tau)/\tau$. So rank-at-$\tau$ decrements continuously, then discretely, as σ^2 shrinks, with the hard update as the $\sigma^2 = 0$ endpoint. PR is the smooth interpolant; rank-at-τ is its thresholded limit.

Proof sketch. (i) The subtracted term is rank-one PSD, giving the Löwner inequality directly, and $\operatorname{tr}\Sigma - \operatorname{tr}\mathcal{G}(\Sigma) = \operatorname{tr}(\Sigma - \mathcal{G}(\Sigma)) \geq 0$ since trace is non-negative on PSD matrices. For the PR claim, write $\mathrm{PR}(\lambda) = S_1^2/S_2$ with $S_1 = \sum_i \lambda_i$, $S_2 = \sum_i \lambda_i^2$. Then $\partial\,\mathrm{PR}/\partial\lambda_u = 2S_1(S_2 - \lambda_u S_1)/S_2^2$, and $S_2 - \lambda_u S_1 = \sum_{i \neq u} \lambda_i(\lambda_i - \lambda_u)$ (the $i = u$ term vanishes), which is non-negative exactly when $\lambda_u \leq \sum_{i \neq u} \lambda_i^2 / \sum_{i \neq u} \lambda_i$. Shrinking λ_u cannot increase PR exactly on this range; note PR is genuinely Schur-concave as a function of the full eigenvalue vector, but that is a constant-sum (majorization) statement and does not by itself constrain a trace-reducing move like this one, which is why the sign condition above is needed rather than assumed. (ii) Sherman–Morrison restricted to u. (iii) Taking $\sigma^2 \to 0$ gives the oblique-to-orthogonal projection limit. (iv) Solve the eigenvalue-below-τ condition for σ^2. ■

Remark (PR vs. trace under simultaneous contraction: why §3.3 uses $\operatorname{tr}\Sigma$ for H2). $\mathrm{PR}(\Sigma) = (\operatorname{tr}\Sigma)^2/\operatorname{tr}(\Sigma^2)$ is exactly scale-invariant: $\mathrm{PR}(c\Sigma) = \mathrm{PR}(\Sigma)$ for any $c > 0$. A single communicative act can resolve several divergent latents at once. For example, "I mean the *red* book" grounds item- and color-identity together. This can shrink several eigenvalues by a roughly common factor, changing $\operatorname{tr}\Sigma$ substantially while leaving PR unchanged, or moved only by finite-sample noise. At $r_{\text{div}} = 1$, the centered-analysis regime of Table 2: $\mathrm{PR}(\Sigma) \equiv 1$ identically regardless of scale, so PR carries no grounding information there at all. $\operatorname{tr}\Sigma = \mathbb{E}\|\delta\|^2$ is the only one of the two statistics still sensitive to it. PR is the right metric for whether grounding has changed *which* directions carry the mismatch: a shape question, relevant to the *I/O* split. $\operatorname{tr}\Sigma$ is the right metric for whether grounding has *occurred*: a magnitude question, what H2's sawtooth is actually about. Conflating the two would silently miss any grounding event structured as simultaneous multi-direction contraction. This is the expected case for a testbed like the director–follower task, where a single resolving utterance is designed to close more than one divergent latent at once.

A.6 FEASIBILITY CONDITION 1 IN BOTH REGIMES

In the hard regime, $\mathcal{G}_{u,0}$ is an orthogonal projection: $\mathcal{G}_{u,0}^2 = \mathcal{G}_{u,0}$. Announcing twice equals announcing once. Condition 1, consistency under common-knowledge iteration, is the requirement that the shared subspace F be closed under the announcement operators the interaction generates.

This is the representational image of the Aumann no-disagreement property. In the soft regime ($\sigma^2 > 0$), the update is a strict contraction along u, but it is not idempotent. Iterated soft grounding converges geometrically to the grounded fixed point, at rate $\sigma^2/(u^\top \Sigma u + \sigma^2)$. Condition 1 then becomes a metric-stability statement. The shared subspace is an attracting fixed set. Feasibility is quantified by the iterated operator's spectral gap, and defeasibility is loss of that attraction. The hard case is the $\sigma^2 \to 0$ specialization of the same operator family, acting on the same identified geometry from §A.3–A.4.

A.7 Linear proxy vs. nonlinear shape

§A.1–A.6 are linear: a Gaussian factor model, covariance operators. The computational-mechanics geometry motivating F (§2) is in general nonlinear. The linear estimator returns the dimension count (PR / rank-at-τ) and canonical-angle spectrum cheaply. Persistent homology on the same activations would check whether aligned subspaces share the same *shape*, not merely matching variance. High linear alignment alongside divergent persistence diagrams would flag representations that are surface-aligned but topologically divergent. The **random-rotation null** used throughout, the threshold for "canonical correlation ≈ 1" in Theorem A.3 and the active/noise cutoff, is generated by recomputing the alignment statistic under Haar-random rotations of one agent's representation. Thresholds are set as upper-tail quantiles of the resulting null distribution.

A.8 The grounding-event-profile bound

Theorem A.3's (C2) and Theorem A.4's (C3) require that disagreement and resolution events *occur and cover* the relevant subspaces. This section quantifies "enough," giving the bound §5.1 evaluates per model.

Effective dimension. The estimator pre-projects onto the active subspace (PCA/SVCCA truncation, §3.3). All bounds below therefore scale with $r_{\text{eff}} = r_F + r_{\text{div}}$, with $r_{\text{div}} = r_I + r_O$, not the ambient width d. This is what makes the estimator tractable on small models. A high-d residual stream may still carry a low-r_{eff} interaction geometry, and only r_{eff} enters the sample complexity.

Separation margins. Let $\gamma := \min_{u \in I \oplus O}(1 - \rho_u^{\text{disagree}})$: the frame-separation/coverage margin, one minus the largest disagreement-conditioned canonical correlation over divergent directions. $\gamma > 0$ iff every divergent direction is excited by some disagreement event; this is the quantitative form of C2 coverage. A divergent direction never disagreed-upon has $\rho_u^{\text{disagree}} \approx 1$ and is indistinguishable from frame, forcing $\gamma \to 0$. Let $\zeta := \min_{u \in O}(\text{resolution coupling of } u \text{ toward } I_B)$: the resolution margin, the quantitative form of C3.

Proposition (grounding-event-profile bound). Fix target angular resolution η for $\hat{F}$ and η_{IO} for the I/O split, and confidence $1 - \alpha$. Under sub-Gaussian activations, the decomposition is certified when

$$n_{\text{disagree}} \gtrsim (r_{\text{eff}} + \log(1/\alpha))/(\gamma^2 \eta^2) \qquad \text{(certifies } F \text{ vs. } I \oplus O\text{, Thm A.3)} \tag{6}$$

$$n_{\text{resolve}} \gtrsim (r_{\text{div}} + \log(1/\alpha))/(\zeta^2 \eta_{IO}^2) \qquad \text{(certifies } I \text{ vs. } O\text{, Thm A.4a)} \tag{7}$$

$$n_{\text{agree}} \gtrsim (r_F + \log(1/\alpha))/\eta^2 \qquad \text{(anchors } F \text{ as shared, C1)} \tag{8}$$

Conversely, if n_{disagree} falls below the first bound with γ fixed, F and $I \oplus O$ are not separable at resolution η. If $\gamma = 0$, no n_{disagree} suffices. These are the three thresholds Table 3's C1/C2/C3 columns check per model. The specific counts, for example $\geq 1800/1800/1000$ at r_{eff}=15 for Qwen2.5-32B-Instruct (§5.1), are this bound evaluated at that model's realized event rates and effective dimension.

Proof sketch. Each bound is a singular-subspace perturbation (Davis–Kahan/Wedin) estimate. The event-conditioned operator, C_{disagree} for F, the resolution-displacement cross-operator for O, is estimated from its subsample. Sub-Gaussian concentration gives operator-norm error $\|\hat{C} - C\| \lesssim \sqrt{(r_{\text{eff}} + \log(1/\alpha))/n}$ after active-subspace projection. Davis–Kahan converts this to a subspace-angle bound $\sin\angle(\cdot,\cdot) \lesssim \|\hat{C} - C\|/\text{gap}$, with gap γ for the frame separation and ζ for the I/O split. Setting the angle below η (resp. η_{IO}) and solving for n gives the stated counts. The agreement bound has effective gap $O(1)$, since frame correlation ≈ 1 sits a constant above chance, so it has no small-margin denominator. Agreement is cheap relative to disagreement and resolution. ■

Testbed-selection reading. Pure cooperation ($p_{\text{disagree}} \approx 0$) never meets the first bound. F and $I \oplus O$ stay confounded, matching Theorem A.3's converse. Pure unresolved conflict ($p_{\text{resolve}} \approx 0$) fails the second bound. I and O cannot be split, even though F is identifiable. Small γ means disagreements occur but barely distinguish divergent directions from frame; this forces n_{disagree} to grow as $1/\gamma^2$. This is the formal backing for treating the realized (agree, disagree, resolve) event profile as a per-environment, per-model diagnosticity check (§5.1), not an incidental property of the task.

Caveat. Constants, the precise sub-Gaussian/regularity conditions, and the dependence on the noise floor σ_ε^2, which enters the active/noise cutoff and hence r_{eff}, are are left to future work. The bounds above are scaling statements meant to drive testbed selection and power analysis. They are not exact thresholds.

A.9 Summary of assumptions

Assumption	Used in	Relaxable?
Linear-Gaussian factor model	§A.3–A.6	Yes → nonlinear cross-check (§A.7)
Column-orthonormal block loadings	identifiability up to within-block rotation	Within-block rotation is irreducible gauge
C1: agreement spans F	Thm A.3 anchor	Needed for a positive F anchor
C2: disagreement loads on $I \oplus O$	Thm A.3 separation	**Not** relaxable – the diagnosticity condition
C2′: disagreement covers $I \oplus O$ ($\gamma > 0$)	§A.8 profile bound	**Not** relaxable – uncovered direction ⇒ unidentifiable
C3: resolution projects onto I_B	Thm A.4a	Needed only to split I from O
C3′: resolution coupling $\zeta > 0$	§A.8 profile bound	Needed to certify the split at finite budget
Exogenous disagreement labels	Thm A.4c	**Not** relaxable – internal-only labels admit the degeneracy
Positive initial overlap	Thm A.4c convergence	Mitigated by exogenous-label seeding

The two non-relaxable conditions (C2, exogenous labeling) are about the same thing from two angles: identifiability requires that disagreement genuinely occur (C2), and that it be registered independently of the agent's own current beliefs (exogenous labeling). Everything else is gauge or cross-check.

B Additional Methodological Notes

This appendix collects supporting detail for claims made in §4, §5, and §7.

H1 results, Qwen2.5-32B-Instruct (§5.2). Table 2 reports the pooled and within-episode-centered analyses; Figure 2 plots the centered canonical-correlation spectrum and PR-vs-null comparison. The residual eigenvalue spectrum shows a clear elbow at the reported frame dimension in both configurations: pooled eigenvalues jump from 84.7 to 114.4 to 472.6 around the 13th of 15 retained components, and centered eigenvalues jump from 59.9 to 99.5 to 416.4 around the 7th of 9.

Table 2: H1 results for Qwen2.5-32B-Instruct: pooled vs. within-episode-centered analysis.

	Pooled	**Centered (within-episode)**
Retained dimensions (k)	15	9
Frame dimension (r_F)	13	7
PR	3.482	2.824
Null PR (mean $\pm$ SD)	4.310 ± 0.250	3.472 ± 0.200
z (null mean $-$ PR)/SD	3.31	3.24
PR significantly below null ($z > 2$)	**Yes**	**Yes**
H1 supported	**Yes**	**Yes**

H1 across twelve models (§5.3). Table 3 reports centered PR significance and identifiability status (C1/C2/C3) for every model tested.

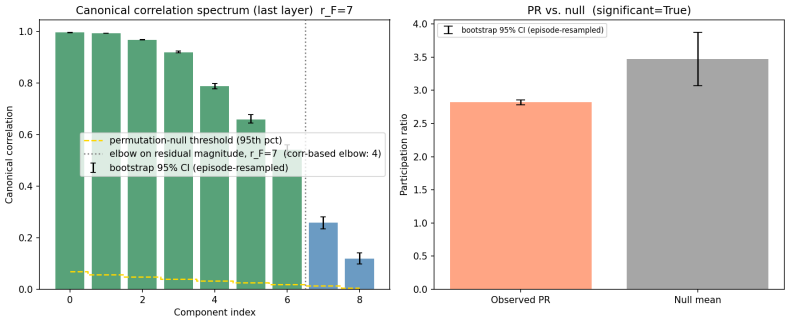

Figure 2: Canonical-correlation spectrum and PR vs. null, within-episode-centered analysis, Qwen2.5-32B-Instruct.

Table 3: H1 (centered) across twelve models. C1/C2/C3: identifiability conditions on realized agree/disagree/resolve counts (§5.1). r_F: centered frame dimension.

Model	**Params**	**C1/C2/C3**	r_F	**PR sig. (centered)**
gemma-2-2b-it	2B	×✓✓	3	No
Qwen2.5-3B-Instruct	3B	✓✓×	5	No
gemma-2-9b-it	9B	×✓✓	2	No
Qwen2.5-14B-Instruct	14B	✓✓×	7	No
Ministral-3-8B-Instruct	8B	×✓×	7	No
Ministral-3-14B-Instruct	14B	×✓×	4	No
Mistral-Small-24B-Instruct	24B	✓✓×	4	No
gemma-2-27b-it	27B	×✓×	6	No
Qwen2.5-32B-Instruct	**32B**	✓✓✓	**7**	**Yes**
Qwen3-32B (reasoning)	32B	✓××	10	No
gemma-4-31B-it	31B	× × ×	48†	Yes†
Magistral-Small-2506 (reasoning)	24B	✓✓×	5	No

†gemma-4-31B-it fails all three identifiability conditions and its reported frame dimension (48) is nearly the entire retained subspace, not a small, concentrated frame; we do not read this as a second confirming case (§7).

Per-model identifiability breakdown. Eleven of twelve models fail at least one identifiability condition (Table 3); Qwen2.5-32B-Instruct is the only model that passes all three. C3 (too few resolution events) is the most common failure, appearing alone in Qwen2.5-3B, Qwen2.5-14B, Mistral-Small-24B, and Magistral-Small-2506, and together with C1 in Ministral-3-8B, Ministral-3-14B, and gemma-2-27b-it. C1 (too few agreement events) fails alone in gemma-2-2b-it and gemma-2-9b-it. Qwen3-32B fails C2 and C3 together; gemma-4-31B-it fails all three. gemma-4-31B-it is the starkest case: its centered analysis reports r_F=48, close to the full retained space rather than a small frame, alongside a complete identifiability failure. We read this as the method producing a degenerate answer under insufficient labeled contrast and do not count it as second confirming case in §5.3.

Layer sweep. A pooled-analysis sweep over four layers of Qwen2.5-32B-Instruct (47, 53, 59, and the final layer 64, out of 64 total) shows the effect is specific to the last layer we tested ($z = 3.30$); it is not significant at the three earlier layers ($z = -7.46, 1.97, 1.89$ respectively). The two layers closest to the final layer sit just under the $z > 2$ threshold, not far from it.

Turn-stratified probe, per-model breakdown. Table 4 gives the full per-model AUC ranges underlying §5.5's pooled and turn-2-only summary.

Director vs. Follower decoding §3.5 fits Director and Follower activations as independent classifiers against the same target (does the Follower actually know the referent); Table 4's ranges pool both roles together. Pooled AUC is essentially identical for both roles in every model, consistent with pooled decoding being carried by the trivial and surface-stated turns rather than by role-specific content. At turn 2, both roles are at chance in nearly every combination. Of the four (agent, position,

Table 4: Turn-stratified probe results, all twelve models. Pooled and turn-2-only AUC ranges are across the 8–16 (agent, position, layer) combinations tested per model.

Model	Pooled AUC	Turn-2 AUC	Turn-2 significant
gemma-2-2b-it	0.87–0.97	0.48–0.54	1/16
Qwen2.5-3B-Instruct	0.96–0.97	0.49–0.53	0/16
gemma-2-9b-it	0.96–0.97	0.49–0.52	0/16
Qwen2.5-14B-Instruct	0.96–0.97	0.49–0.55	2/16
Ministral-3-8B-Instruct	0.96–0.97	0.49–0.54	0/16
Ministral-3-14B-Instruct	0.96–0.97	0.49–0.54	0/16
Mistral-Small-24B-Instruct	0.96–0.97	0.48–0.53	0/16
gemma-2-27b-it	0.96–0.97	0.48–0.52	0/16
Qwen2.5-32B-Instruct	0.96–0.97	0.50–0.54	0/16
Qwen3-32B (reasoning)	0.96–0.97	0.48–0.54	1/16
gemma-4-31B-it	0.96–0.97	0.49–0.53	0/16
Magistral-Small-2506 (reasoning)	0.96–0.97	0.47–0.53	0/16

layer) combinations across all twelve models that cross the 2σ threshold at turn 2 (Table 4's "turn-2 significant" column), all four are on the Follower's own activations (self-knowledge, I), none on the Director's (a model of the Follower's state, O). The expected false-positive count at this uncorrected threshold across $12 \times 16 = 192$ combination-tests is $192 \times 0.023 \approx 4.4$, matching the four observed almost exactly, and a 4–0 split under a fair coin has probability $1/16$. We do not apply Benjamini-Hochberg correction to the turn-2-specific test itself (only the pooled fit is FDR-corrected, §3.5); the four hits are consistent with the false-positive rate already discussed in §5.5.