The Last Writer Wins: Installing Epistemic Disposition in Multi-Agent LLM Pipelines

Sam Bobo

Draft v0.2 — July 2026 | Target: arXiv cs.CL

Abstract

Multi-agent LLM pipelines route work through specialist models and aggregate their outputs through a synthesizer. We study *disposition*—what a model chooses to emit independent of what it knows, operationalized as five epistemic behaviors (training-cutoff disclosure, modeled-assumption flagging, precise vocabulary distinctions, jurisdictional distinguishing, and hedging)—and ask where disposition in a pipeline comes from and whether it can be installed. Across $\sim$290 audited runs of a four-model "Council of Experts" on consumer hardware, we find: (1) the pipeline architecture itself lifts disposition density 3–9$\times$ over single-shot use, model-agnostically, and *amplifies* under simultaneous behavioral demand where single-shot *dilutes*; (2) attempts to install disposition into a specialist seat by prompting or by supervised fine-tuning on behavior-rich exemplars produce large seat-level lifts ($\approx 2\times$, bootstrap CIs excluding baseline) that are indiscriminate—both hedge on a trigger-free control case—and that vanish from the pipeline's final output; reference-free preference optimization (ORPO) on content-controlled pairs installs no measurable seat-level magnitude—invariant to a $3.2\times$ increase in preference data—but is the only mechanism that *improves* responsiveness, suppressing unwarranted hedging below the untrained baseline; (3) the mechanism is not sentence-level content entanglement (refuted, $r = -0.10$) but a **synthesizer register**: each aggregator model writes at its own characteristic disposition density, nearly independent of seat input—over-dense input can *invert* output density—while the synthesis prompt's preservation instructions act as a gain control (removing them collapses output density 2–5$\times$ across all three synthesizers tested). Practically: a pipeline's epistemic posture is governed by its last writer and that writer's instructions; upstream installation is mostly futile. All experiments are pre-registered, \$0-budget, and fully reproducible on a single 32 GB consumer machine.

1 Introduction

Production LLM systems increasingly take the form of pipelines: multiple models, often specialized, whose outputs are aggregated by a final model into the answer a user sees. Evaluation practice, however, remains largely single-model: we measure whether *a model* hedges appropriately, discloses its limits, or labels assumptions. This paper asks what happens to those behaviors *inside a pipeline*—who originates them, whether they can be installed, and whether they survive aggregation.

We distinguish two evaluation axes that are conflated in practice. **Content** is whether the right substantive material appears (measured here by per-case expert rubrics). **Disposition** is what a model chooses to emit about its own claims. Our experiments show these axes are orthogonal—the model family that wins content loses disposition and vice versa—and that pipeline disposition is governed almost entirely by the final aggregation stage.

Contributions. (i) A two-axis evaluation of a heterogeneous local specialist pipeline, with a composite

disposition metric (CDS) and an architectural lift ratio (ALR). (ii) A controlled comparison of three mechanisms for installing disposition into one pipeline seat—prompting, exemplar SFT, and reference-free preference optimization—under pre-registered predictions, with a trigger-free control case that separates *responsive* from *habitual* behavior. (iii) Identification and ablation of the **synthesizer register** as the mechanism governing which installed behaviors reach the pipeline's output. (iv) A fully local, $0, consumer-hardware protocol with append-only audit logs for all 287 runs.

2 Related work

Multi-agent orchestration. Debate and role-played collaboration improve factuality and reasoning [1, 2, 3, 4], and recent work compares orchestration against the strongest single model [5] or trains orchestrators end-to-end [6]. This literature measures content outcomes. The closest work on the *aggregation* step, Parallel-Synthesis [7], trains a synthesizer adapter and explicitly distills reasoning behavior from text-concatenation synthesis—an implicit acknowledgment that naive aggregation loses behavior, which our register mechanism explains and quantifies for epistemic behavior specifically. We measure what orchestration does to epistemic behavior and find an amplification effect independent of the model filling the seats.

Specialists vs. generalists. Evidence is mixed on whether small domain fine-tunes beat larger generalists in-domain [8, 9]. Our content results replicate the negative side (a 20B open MoE outperforms a four-specialist council $42\% \rightarrow 25$–31% on rubric coverage) and motivate the pivot: the surviving specialist value is dispositional, not informational.

Uncertainty expression and abstention. Verbalized uncertainty [10], linguistic calibration [11], knowing-what-you-know [12], abstention training [13] and recent confidence-aware abstention work [14, 15] train and evaluate single models end-to-end. A parallel line asks whether epistemic markers *faithfully* track a model's internal confidence—and finds systematic deficiencies [16, 17, 18]. Our contribution is orthogonal to faithfulness: we take marker *emission* as the observable and ask where it originates and survives in a pipeline, not whether it is calibrated to hidden confidence. A seat's trained uncertainty behavior must survive an aggregator, and mostly does not.

Preference optimization for behavior. DPO [19] and ORPO [20] are standard for style/behavior shaping; DPO's length bias is documented [21], which our pair-construction controls address. Preference-data scaling is itself contested: simply adding more samples per prompt often fails to improve and can degrade downstream behavior [22, 23]—consistent with our dose-invariance null, which we further attribute to a downstream aggregation ceiling rather than to the optimizer. Comparisons of system-prompt steering vs. preference training exist for single models; we compare them *through* a pipeline.

3 Apparatus

Council. Four local models via Ollama on a 32 GB MacBook Air M5: Phi-4-14B Lead (planner + synthesizer), and three specialist seats—Llama3-Med42-8B (healthcare), Saul-7B-Instruct-v1 (legal), Qwen-Open-Finance-R-8B (finance), all Q4/Q8 GGUF. The Lead performs step-back decomposition [24] and dispatches self-contained sub-questions; seats never see the original query (this architectural choice, not prompt strengthening, eliminated cross-domain lane bleed at sub-10B scale). The Lead then synthesizes under a structurally enforced Tensions-then-Synthesis prompt containing four PRESERVE instructions.

Cases. Seven three-domain scenarios: five failure-mode-targeted (synthesis stress, jurisdictional vocabulary, quantitative discipline, recency honesty, adversarial cross-domain tension) plus a matched disposition pair—case 6 (*trigger-heavy*: demands all five behaviors simultaneously) and case 7 (*trigger-light*: an

organizational-strategy question warranting none). Case 7 is the **responsiveness gate**: any arm emitting disposition there is behaving habitually, not responsively.

Scale. 287 imported, audited runs (append-only JSON, one file per run, each carrying per-phase inputs/outputs/backends); all inference via Ollama on one fanless M5 MacBook Air, 32 GB unified memory, 26.8 GB Metal working set. Runs are nondeterministic across repeats even at temperature 0 (runtime batching); we treat repeats as seeds and report bootstrap intervals.

Metrics. Behavior density = pattern-matched occurrences of the five behavior families per 1,000 characters. CDS = density $\times \sqrt{\text{distinct behaviors}/5}$. ALR = council density/matched single-shot density. Seat-level density is computed on the specialist's own turn from the audit log; final density on the synthesized output. (Definitions and regexes are versioned with the code; see §8.)

Installation arms (legal seat, single-variable design, seed 42): A′ = conversion-control baseline (identical Saul weights through our fp16→GGUF→Q4 pipeline); B = a frozen behavior-spec addendum on the seat's system prompt (five behaviors, one example each, an explicit "do not force it" clause); SFT = LoRA on 91 behavior-rich *chosen* responses; ORPO = LoRA reference-free preference optimization on the same 91 chosen responses paired with behavior-stripped rejections. Pairs are content-controlled (chosen/rejected are rewrites of the same base answer), length-ratio filtered (0.8–1.4) against length bias, overlap-filtered (Jaccard ≥ 0.35 on capitalized tokens), and leakage-screened against all seven cases. All protocol amendments were documented before training; predictions were pre-registered before each cell ran.

4 Result 1: architecture shapes disposition

Across seven cases, council configurations emit 3–9× the behavior density of matched single-shot baselines. Aggregate CDS: specialists-v2 0.928, Opus-as-council 0.669, specialists-v1 0.584, gpt-oss-as-council 0.460, Opus single-shot 0.159, gpt-oss single-shot 0.099. ALR: Opus pair $2.99\times$, gpt-oss pair $3.91\times$, local v1 $5.14\times$, upgraded v2 $8.77\times$—the synthesis prompt's preservation scaffold does real work even with no specialist alignment in the seats. Under the trigger-heavy case the asymmetry sharpens: council modes rise to 1.66–$1.82\times$ their per-case baseline while single-shot falls to $0.68\times$—orchestration *amplifies* disposition under simultaneous demand where a single context *dilutes* it. On the trigger-light case every mode, specialist or frontier, emits zero: measured disposition is responsive, not habitual, across the board.

5 Result 2: installing disposition — magnitude vs. responsiveness

At $n = 5$ seeds per cell (140 runs, bootstrap 95% CIs):

Arm	Seat density	Final-output CDS	Trigger-light gate
A′ baseline	0.89 $[0.69, 1.11]$	$\mathbf{0.86}$ $[0.64, 1.08]$	0.96
Prompting	$\mathbf{1.85}$ $[1.42, 2.32]$	0.59 $[0.40, 0.81]$	3.03 ✗
SFT-on-chosen	$\mathbf{1.77}$ $[1.46, 2.09]$	0.58 $[0.43, 0.73]$	1.21 ✗
ORPO	0.87 $[0.60, 1.17]$	0.66 $[0.49, 0.85]$	$\mathbf{0.15}$ ✓

Prompting and SFT install real seat-level magnitude (CIs exclude baseline)—and both fail the gate, hedging on the trigger-free case (prompting despite its explicit "do not force it" clause). Both also lose their gains at the pipeline output: final CDS lands at or below the untrained baseline. ORPO installs no significant

seat-level magnitude; its effect is qualitative—it is the only arm that *suppresses* unwarranted hedging below baseline while leaving trigger-case behavior intact and imposing no content tax (rubric coverage identical to A′).

The null is dose-invariant. Retraining ORPO at $3.2\times$ dose (292 content-controlled pairs, epoch-matched to the 91-pair run) moved seat-level magnitude by zero: seat density 0.84 $[0.58, 1.11]$ vs. the 91-pair 0.85 and baseline 0.84, even though the larger set raised training-time preference accuracy from 0.48 to 0.94. More preference data was internalized but not emitted—strong evidence that ORPO's seat effect is not gradual installation of magnitude but a bounded suppression of unwarranted behavior. (The suppression itself weakened slightly with dose: the trigger-light gate rose from 0.15 to 0.49, still below baseline; more diverse pairs modestly broadened where the model deploys hedging.)

Two further details. First, prompting was tested at two loci: on the generalist single-shot (a $6\times$ final-output lift, $0.099 \rightarrow 0.589$, that failed the gate at 0.985—its answer to the hybrid-work communication question opened a section "Current State Snapshot (as of my training data)" and closed with "actual results may vary": disclosure theater where nothing warranted disclosure) and on the seat (the table above). Second, content is unaffected by successful installation: rubric coverage is $12/36$ for both A′ and ORPO ($9/36$ for the original production seat)—no content tax. Gate cells use $n = 4$–5 (the planner routes no specialists on the trigger-light case in some runs; routing variability is itself Ollama-runtime nondeterminism at temperature 0). Summary: *prompting and exemplar training change what a model says; preference training changes when it says it.*

6 Result 3: the synthesizer register

Why do installed behaviors vanish? A pre-registered entanglement hypothesis—content-woven markers survive aggregation, detached boilerplate is stripped—was **refuted**: entangled share is flat across arms (baseline 56%, prompting 61%, ORPO 61%, SFT 67%—SFT's markers are the *most* content-woven yet retain worst), retention splits by arm not position (baseline 1.08, ORPO 0.96 vs. prompting 0.49, SFT 0.49), and entanglement does not predict retention ($r = -0.10$ over 102 runs). Notably the refuted hypothesis and the overturned PRESERVE prediction (below) were both pre-registered—we report them as part of the record.

A 72-run ablation (2 input arms $\times$ 3 synthesizer models $\times$ PRESERVE instructions on/off $\times$ 6 cases) located the mechanism:

Synthesizer	PRESERVE, base seat	PRESERVE, 2–3× hot seat	no-PRESERVE, base	no-PRESERVE, hot
Phi-4-14B	1.17	0.61	0.52	0.31
gpt-oss-20B	0.63	0.86	0.20	0.13
Qwen2.5-7B	1.24	1.06	0.24	0.62

Each synthesizer writes at its own characteristic density band regardless of input (registers are writer-specific); over-dense input does not transmit and can *invert* output density (Phi-4: $1.17 \rightarrow 0.61$—over-correction, saturated hedging read as stylistic noise); and the PRESERVE instructions are a **gain control**, their removal collapsing output 2–$5\times$ on every synthesizer (overturning our registered prediction that instructions would not matter). Mechanism: *final disposition* $\approx f($*register* $\times$ *instructions*$)$*, nearly independent of seat input. ORPO "survives" by staying inside the register; prompting and SFT "strip" by exceeding it.* The dose-invariance of §5 is the register's corollary: if the aggregator's band, not the seat's input, sets output

disposition, then no amount of additional preference data at the seat can raise it—which is exactly what $3.2\times$ dose showed.

7 Background: why disposition, not content

The council was built to test whether small domain fine-tunes beat generalists on content. They do not: a single 20B open-weights MoE (gpt-oss-20B) outperforms the full council on rubric coverage (42% vs. 25%; 31% after best-available specialist upgrades), and the council exhibited classic small-model failures (confabulated trial citations, repealed regulation treated as live). What specialists retained was disposition—the highest CDS of any configuration—which motivated the installation question this paper answers. We report this as motivating context, not a contribution.

8 Limitations

Pattern-based scoring has paraphrase blind spots and rewards concision (per-character normalization); a judge-validated and per-claim-normalized replication is planned. ORPO substitutes for DPO due to a memory defect in the only local DPO trainer available (documented); claims are scoped to reference-free preference optimization. One trained seat, one base model (Mistral-7B-v0.1, weakly aligned), a preference-data range of 91–292 pairs (the $3.2\times$ dose-response run showed no magnitude gain but does not exclude effects at far larger scale), $n = 5$ seeds, seven self-authored cases, and three local synthesizers $\leq$ 20B—frontier aggregators untested. All experiments ran on one consumer machine at $0 API cost; we regard the reproducibility this buys as partial compensation for the scale it forgoes.

9 Conclusion

In multi-agent pipelines, epistemic disposition is not additive across stages: it is set by the last writer's register and that writer's instructions. Upstream installation by prompting or exemplar tuning produces loud but indiscriminate behavior that the aggregator strips; preference optimization produces quiet, responsive behavior that survives by conformity—and whose seat-level magnitude does not grow with more preference data, because the aggregator's register, not the seat, sets the ceiling. Builders who want calibrated pipelines should tune the synthesizer's instructions and choose the final model deliberately—and evaluate disposition at the pipeline's mouth, not the seat.

Reproducibility. All code, prompts, 287 imported audited runs, pre-registrations, protocol amendments, verdicts (including refuted hypotheses), the 99- and 292-pair datasets, LoRA adapters, and regeneration scripts: `https://github.com/srbobo/council-of-experts-research`.

A Glossary

Seat
: one specialist model in the council, answering only its dispatched sub-question (e.g., Saul-7B = the legal seat).

Synthesizer / Lead
the model that receives all seat outputs plus the original question and compresses them into the final answer—the pipeline's last writer.

Disposition
what a model chooses to emit independent of what it knows, operationalized as five behavior families: (1) training-cutoff disclosure; (2) modeled-assumption flagging; (3) precise vocabulary distinctions; (4) jurisdictional distinguishing; (5) **hedging**—stated conditionality of a claim ("this may vary if…"), *not* refusal or vagueness.

Synthesis stripping
behavior density present at a seat's output but absent from the final output after synthesis.

Synthesizer register
a Lead model's characteristic output disposition density—writer-specific and largely independent of seat input (§6).

Responsive vs. habitual
responsive behaviors appear only when domain triggers warrant them (trigger-light gate, case 7); habitual behaviors fire regardless.

Alignment
post-pretraining procedures (SFT, RLHF, DPO/ORPO/CPO) shaping disposition, as distinct from the pretraining corpus, which shapes knowledge.

B Model selection

Every selection cleared seven gates: (1) GGUF availability on a mirror Ollama can pull; (2) research-permissive license; (3) ChatML-compatible or locally fixable chat template; (4) recognized-lab provenance; (5) distinct lineage per seat (four labs, three model families—a family-level training artifact cannot masquerade as a council effect); (6) active maintenance; (7) evidence of third-party use.

Memory budget. Sequential mode holds the Lead warm across all five phases, so peak $\approx M_{\text{lead}} + \max_{\text{seat}} M_{\text{seat}} + \text{KV} + \text{OS reserve} \leq 32\,\text{GB}$; Metal's recommended working set on the test machine is 26.8 GB. This bounds the Lead at ≤ 14B (Q4) and seats at ≤ 8B.

Lead: Phi-4-14B (MIT)—best reasoning-per-parameter at 14B for structured planner output; Qwen2.5-14B rejected on memory headroom, Llama-3.1-8B on synthesis capacity, 22–27B dense models on the Metal cap. The Lead is deliberately *not* domain-tuned. **Healthcare: Llama3-Med42-8B**—the only 8B clinical fine-tune with a multi-stage *preference-alignment* stage. **Legal: Saul-7B-Instruct-v1**—the only sub-13B continued pretrain (30B+ tokens) on multi-jurisdiction common-law text. **Finance: Qwen-Open-Finance-R-8B**—newest backbone of any seat (Qwen3), $> 50\%$ finance corpus. **MoE comparator: gpt-oss-20B**—$\sim$3.6B active parameters, reasoning-tuned, Apache-2.0; fits beside the Lead ($14 + 9 = 23\,\text{GB} < 26.8\,\text{GB}$), unlike Qwen3-30B-A3B (27 GB) or Mixtral-8x7B (35 GB).

C Metrics and estimation

Let B be the five behavior families and P_b the regex pattern set for family b. For text t:

$$\begin{aligned}
\text{Density} \quad & d(t) = \frac{1000}{|t|} \sum_{b \in B} \sum_{p \in P_b} |\text{matches}(p, t)|, \\
\text{Breadth} \quad & k(t) = \tfrac{1}{|B|} \left| \{\, b \in B : \exists p \in P_b,\ \text{matches}(p, t) \neq \varnothing \,\} \right|, \\
\text{CDS} \quad & \text{CDS}(t) = d(t) \cdot \sqrt{k(t)}, \\
\text{ALR} \quad & \text{ALR}_m = \bar{d}_{\text{council}}(m) \,/\, \bar{d}_{\text{single}}(\hat{m}), \\
\text{Retention} \quad & R = \text{clip}\big(d(\text{final})/d(\text{seat}),\ 0,\ 2\big).
\end{aligned}$$

The square root in CDS softly penalizes narrow emission without letting one absent family zero the score (a geometric mean was rejected for exactly that failure); $\alpha = \frac{1}{2}$ is a design choice, sensitivity unreported. Matched baseline $\hat{m}$ pairs opus↔opus, gptoss↔gptoss; local councils use gptoss-single as nearest open generalist. *Worked example*: a 1,000-character answer with "as of my training data (2024)", "modeled at $8,000 assuming 60% persistence" (×2), and "this may vary if the statute changes" has $d = 4.0$, $k = 3/5$, CDS $= 4.0\sqrt{0.6} \approx 3.10$. Uncertainty: percentile bootstrap over runs (10^4 resamples), 95% intervals; $n = 30$ per arm for trigger cases, $n = 4$–5 for the trigger-light gate. Entanglement test (refuted): a marker sentence is *entangled* if it also matches content signals (digits, §, U.S.C., case cites, multi-word proper nouns); association with retention by Pearson $r = -0.10$ ($n = 102$).

D Pair construction and training

Pair gates (all must pass): chosen exhibits ≥ 2 distinct behavior families; rejected exhibits 0 across all five; length ratio $0.8 \leq |\text{chosen}|/|\text{rejected}| \leq 1.4$; content overlap $J(C(c), C(r)) \geq 0.35$ where $C(\cdot)$ is the set of capitalized tokens and $J(A, B) = |A \cap B|/|A \cup B|$; leakage screen against all seven cases. Yield: 200 prompts → 99 pairs (extended to 292 for the dose-response run).

ORPO objective [20]: $\mathcal{L} = \mathcal{L}_{\text{SFT}} + \lambda \mathcal{L}_{\text{OR}}$ with $\mathcal{L}_{\text{OR}} = -\log \sigma\big(\log \frac{\text{odds}_\theta(y_w|x)}{\text{odds}_\theta(y_l|x)}\big)$ and $\text{odds}_\theta(y|x) = \frac{P_\theta(y|x)}{1-P_\theta(y|x)}$—a reference-free preference penalty added to the NLL of the chosen response.

Configuration: LoRA rank 8, scale 10, last 16 layers (10.5M trainable, ≈0.145%); base loaded 4-bit; lr $= 5 \times 10^{-6}$; batch 1, gradient accumulation 4; max sequence 1,792 (dataset $p_{100} = 1{,}698$); 364 iterations (91-pair) / 1,056 (292-pair, epoch-matched); gradient checkpointing; seed 42. The SFT control is identical except SFT-mode on chosen-only data. Fused adapters follow the same fp16→GGUF→Q4_K_M conversion as the untrained control (A′), so the only delta between compared artifacts is the trained weights.

References

[1] Du, Y. et al. Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv:2305.14325 (2023).

[2] Wang, J. et al. Mixture-of-Agents Enhances Large Language Model Capabilities. arXiv:2406.04692 (2024).

[3] Tang, X. et al. MedAgents: LLMs as Collaborators for Zero-shot Medical Reasoning. arXiv:2311.10537 (2023).

[4] Kim, Y. et al. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making. arXiv:2404.15155 (NeurIPS 2024).

[5] Tian, A.X. et al. Beyond the Strongest LLM: Multi-Turn Multi-Agent Orchestration vs. Single LLMs on Benchmarks. arXiv:2509.23537 (2025).

[6] Ke, Z. et al. MAS-Orchestra: Understanding and Improving Multi-Agent Reasoning. arXiv:2601.14652 (2026).

[7] Towards Direct Latent-Space Synthesis for Parallel Branches in LLM-Agent Workflows. arXiv:2606.14672 (2026).

[8] Hui, Z., Dong, Y.R., Shareghi, E., Collier, N. TRIDENT: Benchmarking LLM Safety in Finance, Medicine, and Law. arXiv:2507.21134 (2025).

[9] Buskila, A.A. Domain Fine-Tuning vs. Retrieval-Augmented Generation for Medical Multiple-Choice Question Answering: A Controlled Comparison at the 4B-Parameter Scale. arXiv:2604.23801 (2026).

[10] Lin, S., Hilton, J., Evans, O. Teaching Models to Express Their Uncertainty in Words. arXiv:2205.14334 (2022).

[11] Mielke, S.J. et al. Reducing Conversational Agents' Overconfidence through Linguistic Calibration. TACL (2022).

[12] Kadavath, S. et al. Language Models (Mostly) Know What They Know. arXiv:2207.05221 (2022).

[13] Zhang, H. et al. R-Tuning: Instructing LLMs to Say 'I Don't Know'. arXiv:2311.09677 (NAACL 2024).

[14] Zong, H. et al. I-CALM: Incentivizing Confidence-Aware Abstention. arXiv:2604.03904 (2026).

[15] Wu, S. et al. BAS: A Decision-Theoretic Approach to Evaluating LLM Confidence. arXiv:2604.03216 (2026).

[16] Can LLMs Use Linguistic Uncertainty Markers to Reliably Reflect Intrinsic Confidence? arXiv:2605.28778 (2026).

[17] Saying More Than They Know: Quantifying Epistemic-Rhetorical Miscalibration in LLMs. arXiv:2604.19768 (2026).

[18] Revisiting Epistemic Markers in Confidence Estimation. arXiv:2505.24778 (2026).

[19] Rafailov, R. et al. Direct Preference Optimization. arXiv:2305.18290 (NeurIPS 2023).

[20] Hong, J., Lee, N., Thorne, J. ORPO: Monolithic Preference Optimization without Reference Model. arXiv:2403.07691 (EMNLP 2024).

[21] Park, R. et al. Disentangling Length from Quality in DPO. arXiv:2403.19159 (2024).

[22] Finding the Sweet Spot: Preference Data Construction for Scaling Preference Optimization. arXiv:2502.16825 (2025).

[23] AIR: A Systematic Analysis of Annotations, Instructions, and Response Pairs in Preference Datasets. arXiv:2504.03612 (2025).

[24] Zheng, H.S. et al. Take a Step Back: Evoking Reasoning via Abstraction. arXiv:2310.06117 (ICLR 2024).