IDV: Customer-Context-Aware Multi-Engine Architecture for AI-Agent Intent Drift Detection

Yan Xue, Mulan Zhou
Instinctive Network
{yan, mulan}@instinctive.network

***Abstract*—AI agents operating on customer infrastructure, especially in financial services, face a governance challenge: customer intent specifications are category-level ("customer accounts," "process payments") rather than instance-level (specific account IDs, segregation-of-duties constraints). We empirically establish that this ambiguity produces measurable interpretation divergence at the LLM layer: two frontier LLMs (GPT-5.5, Claude Opus 4.7) labeling the same naturalistic agent traces under identical Intent Anchor Objects (IAOs) disagree on `drift_detected` for 30% of cases. When one judge labels a trace as instance-level entity access (Object-type drift), the inter-judge disagreement rate is 82.4% [95% CI 65–93%], a $2.7\times$ enrichment relative to the marginal Object label rate. This suggests that single-pass LLM-only monitoring is unstable as a sole arbiter for category-level IAO boundary cases under the tested prompting strategy. We present IDV (Intent Drift Verification), a multi-engine architecture that runs four parallel detection engines (Authority, Coherence, Scope, Impact) on agent events with a two-layer adjudicator preserving the invariant `drift_detected` independent of Impact. On IDV-1000 (850 author-constructed synthetic + 150 naturalistic traces), IDV achieves F1=93.9% (precision 95.9%, recall 92.1%) and AUROC 98.7% on the author-injected subset, at sub-300ms end-to-end latency. Against full N=850 LLM-Judge baselines under JSON-only prompting, IDV achieves a $6.5\times$ reduction in false-positive count (22 FP vs. 143/140 for GPT-5.5/GPT-4o, at precision 79.5%/79.9%). We identify system-design directions that customer-context-aware monitoring naturally supports: interactive IAO refinement, rolling-baseline IAO learning, and ambiguity-aware response routing.**

***Index Terms*—AI agents, intent drift, LLM-as-judge, financial AI, governance, multi-agent systems, auditability**

I. Introduction

AI agents executing actions on customer infrastructure create a governance gap. Per-event guardrails (regex filters, allowlists) miss drift classes that emerge from agent intent rather than individual actions: *Authority drift*, where an agent invokes actions whose authorization context has shifted; *Coherence drift*, where an agent's stated rationale contradicts the customer's authorized scope; and *Scope drift*, where an action touches entities outside the authorized boundary.

We focus our empirical evaluation on financial services as the most regulated and structurally demanding deployment context, but the structural property we identify—category-level intent specification meeting instance-level execution—is not unique to finance.

Customer compliance teams in financial services write authorized-intent specifications at category level ("customer accounts," "transaction approvals"), not instance level, because instance-level enumeration scales poorly when thousands of accounts update daily; SOX Section 404 describes process-level controls; FINRA Rule 2111 references customer profile categories rather than specific identifiers; and PCAOB AS 2201 evaluates transaction populations, not enumerated lists. Instance-level access events must therefore be reasoned about against category-level specifications.

A natural objection is to have the LLM self-monitor. We argue that this is insufficient for five reasons: (1) self-judgment unreliability, since adversarial prompts can suppress LLM self-detection [10], [15]; (2) customer-context blindness; (3) audit independence, since regulatory frameworks create expectations for independent evidence, model-risk documentation, and auditability; (4) adversarial bypass; and (5) frontier LLM calibration divergence: our empirical study shows GPT-5.5 [19] and Claude Opus 4.7 [20] disagree on 30% of naturalistic traces overall (38% on Finance), with disagreements concentrated on Object-type drift.

A. Contributions

1) **Calibration-divergence finding.** Two frontier LLMs disagree on `drift_detected` for 30% of naturalistic traces (38% Finance), with 84% of disagreements (38/45) following a single direction concentrated on Object-type drift (28/38). Conditional on GPT-5.5 labeling Object-type drift, inter-judge disagreement is 82.4% [95% CI 65–93%].
2) **Architectural response.** IDV runs Authority/Coherence/Scope/Impact engines on agent events. A two-layer adjudicator preserves the invariant `drift_detected` independent of Impact, externalizing customer-specific boundaries into auditable configuration.
3) **IDV-1000 benchmark.** We evaluate 850 author-constructed synthetic traces, thematically representative of public agent benchmarks but not derived from their content, plus 150 naturalistic traces under 30 pressure scenarios. Naturalistic traces are independently labeled by frontier LLM judges; high-consensus cases are used

Preprint. Submitted to IEEE CIFEr 2026 (notification: 15 July 2026). This preprint version adds author information to the manuscript submitted for double-blind review; no other content has been changed.

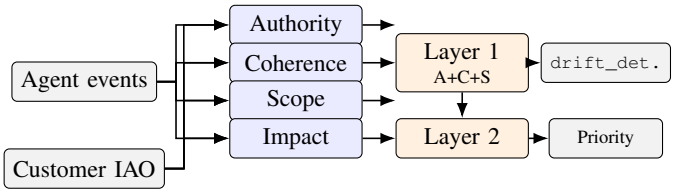

Fig. 1. IDV system architecture. Agent events and customer-specific IAO feed four parallel engines. Layer 1 computes `drift_detected` from Authority, Coherence, and Scope only; Layer 2 adds Impact for response prioritization. The invariant `drift_detected` independent of Impact is enforced by construction.

for AUROC probes, while disagreement cases are analyzed as a core finding.

4) **Empirical validation.** IDV achieves F1=93.9%, precision=95.9%, recall=92.1%, and AUROC=98.7% on N=850; it reduces false positives $6.5\times$ relative to JSON-only LLM-Judge baselines and achieves 100% precision on IT and Sales.
5) **Distribution-shift probe.** No-drift DriftScore median shifts from 0.677 (explicit IAO) to 0.515 (category-level IAO), linking model disagreement, IDV score compression, and Finance-specific residual FPs to the same granularity mismatch.

II. Related Work

LLM safety and monitoring. RLHF-based alignment [6], Constitutional AI [1], and Llama Guard [3] are model-layer approaches. Our work is complementary: an external observer that maintains customer-context state those model-layer approaches do not access.

LLM-as-Judge methodology. Zheng et al. [14] established LLM-as-judge consensus methodology. Subsequent work documents position bias [8] and verbosity bias [7]; we document a domain-specific calibration divergence under category-level IAO.

Agent-evaluation benchmarks. WorkBench [9], TheAgentCompany [11], SWE-bench Verified [4], [5], and tau-bench/tau-squared-bench [2], [13] evaluate agent capability. Financial AI safety benchmarks include FinVault [16], FinToolBench [17], and FinSafetyBench [18]. None address customer-context-aware multi-engine intent-drift monitoring under category-level IAO. Yager's OWA [12] motivates multi-criteria aggregation; our use is specialized to agent monitoring with a Layer-1/Layer-2 split.

III. System Design

Fig. 1 summarizes IDV. The four engines are: **Authority**, checking approval/authorization context against roles, cumulative limits, and segregation constraints; **Coherence**, using NLI to compare stated rationale to authorized scope and operational practice; **Scope**, checking touched entities against the authorized entity scope; and **Impact**, estimating severity (reversibility, blast radius, data sensitivity).

A two-layer adjudicator processes engine outputs. Layer 1 computes `drift_detected` via profile-calibrated aggregation over Authority, Coherence, and Scope. Layer 2 consumes the Layer-1 result and Impact score to produce a response priority. The invariant `drift_detected` independent of Impact is structural: a low-impact action with high drift indicators must still be detected; a high-impact action without drift indicators must not be falsely flagged.

The customer-specific IAO is constructed through an onboarding pipeline drawing on policy documents, role definitions, and historical execution patterns. Per-department default profiles (Finance/IT/Sales) define threshold defaults; customer-specific calibration adjusts them.

IV. Experimental Design

A. IDV-1000 Benchmark

Author-injected subset (N=850). Synthetic agent execution traces are constructed in domain scenarios thematically representative of WorkBench, TheAgentCompany, SWE-bench Verified, and tau-bench/tau-squared-bench. Author-applied perturbation operators inject drift across six categories (Authority, Coherence, Scope, Rationale, Object, Compound), with deterministic perturbation-to-label mapping. Traces are stratified across Finance (N=284), IT (N=283), Sales (N=283), and three complexity tiers. The benchmark includes 32 advisory traces (ground-truth no-drift but high-impact); the N=818 non-advisory binary subset is reported in parallel.

Naturalistic subset (N=150). Thirty pressure scenarios (10 per department) are executed by Claude Sonnet 4.6 and GPT-4o as agents. GPT-5.5 and Claude Opus 4.7 independently label each trace using the same trace, IAO, and drift taxonomy. Both judges use default sampling parameters; exact-output reproducibility is therefore not guaranteed for either judge, and consensus stratification reflects a single labeling pass.

Benchmark-bias concern. Author-constructed benchmarks with deterministic label mapping are common in agent evaluation. Three checks mitigate but do not eliminate this concern: (a) blind LLM-Judge baselines achieve 88.4%/88.8% F1 under JSON-only prompting on the same benchmark, indicating the task is tractable without IDV's engine structure; (b) naturalistic traces are constructed via independent pressure-scenario design and used for distribution-shift probing; and (c) on a 213-trace easy subset, IDV and GPT-5.5 produce identical predictions, while differentiation arises on harder cases.

B. Evaluation Metrics

We report binary F1, AUROC, precision, recall, per-department F1, per-engine recall, and subtype recall. Baseline LLM-Judges (GPT-4o cost-tier and GPT-5.5 frontier-tier) are evaluated under two prompting configurations: (a) JSON-only output (no reasoning text, `max_tokens=2048`), the primary comparison; and (b) chain-of-thought with drift taxonomy, a supplementary robustness check. Baselines are evaluated on the author-injected subset only, since the naturalistic subset uses frontier LLM-Judge consensus as part of its analysis.

TABLE I
IDV AUTHOR-INJECTED PERFORMANCE

Metric	N=850 full	N=818 non-adv.
F1	93.9%	94.5%
Precision	95.9%	97.0%
Recall	92.1%	92.1%
AUROC	98.7%	98.7%
TP/FP/TN/FN	512/22/272/44	512/16/246/44

TABLE II
THRESHOLD SENSITIVITY SWEEP OVER δ UNIFORMLY APPLIED TO PER-PROFILE THRESHOLDS. AUTHOR-INJECTED N=850. THE HEADLINE CONFIGURATION (δ=0) IS ON THE PARETO FRONTIER.

δ	θ (F/I/S)	F1	Prec.	Recall
-0.10	0.45/0.40/0.35	87.7%	99.1%	78.6%
-0.05	0.50/0.45/0.40	91.8%	99.2%	85.4%
0 (default)	**0.55/0.50/0.45**	**93.9%**	**95.9%**	**92.1%**
$+0.05$	0.60/0.55/0.50	92.5%	91.5%	93.5%
$+0.10$	0.65/0.60/0.55	91.5%	88.8%	94.4%

V. RESULTS

A. Author-Injected Performance

Table I reports IDV performance. Per-department F1: Finance 94.8%, IT 91.6%, Sales 95.6%. Per-engine recall on drift subset (N=556): Authority 100%, Coherence 100%, Scope 99.2%, Compound 100%. Diagnostic subtype recall is lower (Rationale 22.5%, Object 42.9%), reflecting a deliberate binary-detection trade-off. Advisory accuracy is 81.2% (26/32). All 22 false positives originate from Finance (FP=22, IT=0, Sales=0). Finance precision is 90.4% on N=284 and 92.9% on N=274 binary subset. Of the 22 Finance FPs, 16 are non-advisory and 6 are Finance advisory traces; specifically, 6 of 10 Finance advisory traces are predicted as drift, versus 0 of 11 IT advisory and 0 of 11 Sales advisory (N=32 advisory traces total)—the advisory FP pattern is itself department-localized to Finance. These high-impact non-drift cases place Coherence scores near the Finance decision boundary.

Threshold sensitivity. The drift-detection thresholds per profile $\theta_{\text{drift}} \in \{0.55, 0.50, 0.45\}$ for Finance/IT/Sales are pre-specified defaults. To assess sensitivity around these defaults, we ran a post-hoc sweep over $\delta \in \{-0.10, -0.05, 0, +0.05, +0.10\}$ applied uniformly to all three profile values (Table II). The default (δ=0) is on the Pareto frontier in (precision, recall); IDV's precision advantage over LLM-Judge baselines is robust to ± 0.05 perturbation. We do not use the sweep to select the headline configuration; a dedicated calibration/test split is left to future work.

B. Naturalistic Distribution Shift

On the naturalistic high-consensus subset (N=100, both judges agree on `drift_detected` and type), IDV achieves AUROC=92.3%. We do not report naturalistic F1 as a headline because frontier-LLM consensus carries ≈30% inter-judge label noise, conflating label noise with detection performance. In our scoring convention, higher DriftScore means greater confidence that actions align with the IAO; the detection rule is DriftScore $< \theta_{\text{drift}} \rightarrow$ flag. The no-drift median shifts from 0.677 (author-injected, explicit IAO) to 0.515 (naturalistic, category-level IAO), moving clean events toward the Finance threshold (0.55) and mirroring the LLM disagreement pattern.

C. Baseline Comparison

Baselines in Table III use JSON-only output to eliminate prompting-length asymmetry. JSON-only achieves marginally higher F1 than corresponding CoT versions (GPT-5.5: 88.4% vs. 88.0%; GPT-4o: 88.8% vs. 88.2%), confirming that IDV's precision advantage is not an artifact of baseline prompting. IDV's precision advantage over GPT-5.5 JSON-only is 16.40pp [12.89pp, 19.75pp]; an unpooled normal approximation yields z=9.35, $p < 10^{-21}$, and a pooled two-proportion test also remains highly significant. The FP-count ratio is $143/22 = 6.50\times$ (bootstrap 95% CI $[4.39\times, 10.83\times]$). All 22 IDV FPs concentrate in Finance, against a uniform-null expectation of ≈7.3 of 22 ($p < 10^{-10}$, binomial).

Per-department precision (Table IV) localizes IDV's residual ambiguity entirely to Finance. IT and Sales achieve 100% precision (0 FP across 283 traces each), while both LLM-Judge baselines distribute false positives across all three departments. GPT-4o achieves 100% recall in both CoT and JSON-only configurations (FN=0 in both; unchanged by prompting choice). GPT-5.5 shows a marginal recall increase under JSON-only (FN=4 CoT $\rightarrow$ FN=2 JSON-only). On the 213-trace easy subset where GPT-5.5 CoT produced complete reasoning chains, both GPT-5.5 and IDV achieve F1=96.2% with identical predictions (TP=175, FP=14, TN=24, FN=0). The IDV–LLM-Judge differentiation therefore arises specifically on hard Object-type, T2/T3, and non-Finance cases. IDV runs local NLI inference at sub-300ms latency versus ≈4.7s for GPT-5.5; deterministic engine-level scores are easier to integrate into audit trails than sampling-dependent reasoning chains.

D. Engine and Adjudicator Ablation

No single engine suffices: Authority alone misses most drifts (recall 41.2%), Coherence alone produces 262 FPs, and Scope alone misses 16.5% of drifts. Weighted-mean is the best fixed-threshold binary classifier on the non-advisory synthetic subset. We do not claim Full IDV strictly dominates weighted-mean; rather, the hybrid blend is selected for ranking stability under naturalistic category-level IAO shift (AUROC 81.8% vs. 72.4% for weighted-mean, a 9.4pp advantage) and for Layer-2 prioritization. Weighted-mean is a low-FP operating point for lower-risk contexts; hybrid IDV is the default risk-sensitive profile for high-stakes Finance deployment.

E. LLM-Judge Calibration Divergence

On naturalistic N=150, the two frontier LLM judges agree on `drift_detected` for 105 traces and disagree for 45 (30%). Per-department: Finance 38%, IT 24%, Sales 28% (Fig. 2). Of 45 disagreements, 38 (84%) follow a single pattern: GPT-5.5 flags drift and Claude does not. Finance

TABLE III
BASELINE COMPARISON ON AUTHOR-INJECTED N=850. VALUES IN BRACKETS ARE 95% BOOTSTRAP CIS (n=10,000).

System	F1	Precision	Recall	TP	FP	TN	FN	N
IDV	93.9 [92.5, 95.3]	95.9 [94.1, 97.5]	92.1 [89.9, 94.3]	512	22	272	44	850
GPT-5.5 LLM-Judge (JSON)	88.4 [86.5, 90.2]	79.5 [76.5, 82.4]	99.6 [99.1, 100.0]	554	143	151	2	850
GPT-4o LLM-Judge (JSON)	88.8 [86.9, 90.6]	79.9 [76.9, 82.8]	100.0 [100.0, 100.0]	556	140	154	0	850

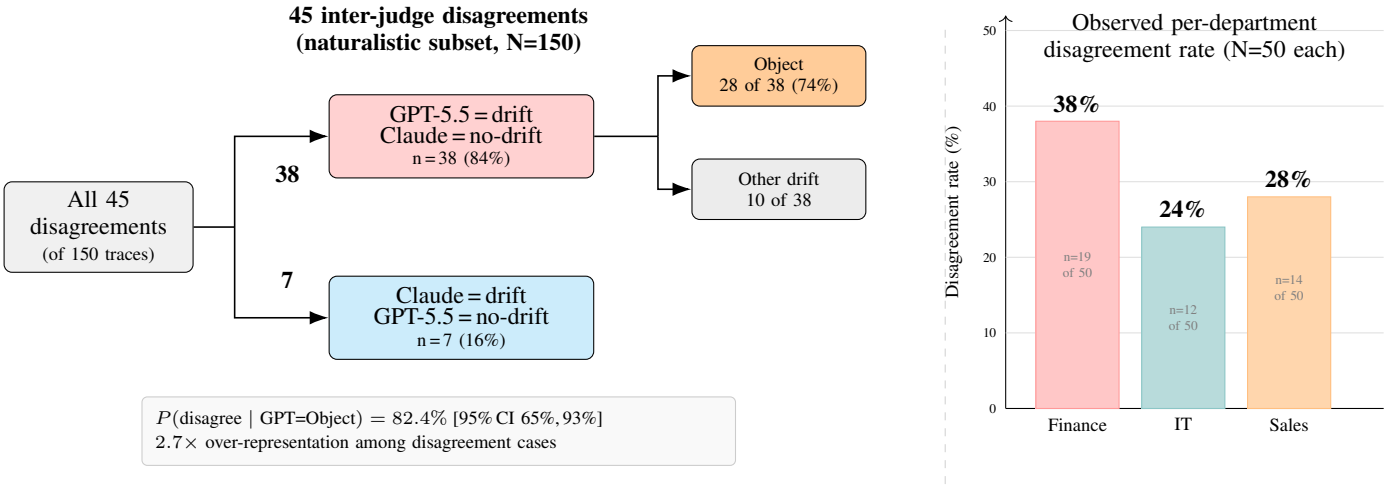

Fig. 2. Frontier LLM-Judge calibration divergence breakdown. *Left*: of 45 disagreements, 38 follow GPT-5.5=drift / Claude=no-drift direction, with 28/38 (74%) on Object-type drift; P(disagree | GPT=Object) = 82.4% [95% CI 65%–93%], a 2.7× over-representation among disagreement cases. *Right*: observed per-department disagreement rates. Finance shows the highest observed rate, consistent with the concentration of instance-level account-boundary cases; department-level differences are interpreted descriptively (N=50 each; no χ^2 test reported).

TABLE IV
PER-DEPARTMENT PRECISION ON AUTHOR-INJECTED N=850. IDV CONCENTRATES RESIDUAL AMBIGUITY ON FINANCE; LLM-JUDGE BASELINES DISTRIBUTE FPS ACROSS ALL DEPARTMENTS. ALL 22 IDV FALSE POSITIVES ORIGINATE FROM FINANCE (IT AND SALES: 0 FP EACH).

Department	IDV	GPT-5.5 (JSON)	GPT-4o† (JSON)
Finance (N=284)	90.4%	85.2%	∼80%
IT (N=283)	**100.0%**	76.9%	∼77%
Sales (N=283)	**100.0%**	75.9%	∼76%

†GPT-4o per-department precision rounded to nearest percentage point.

TABLE V
ABLATION ON N=818 NON-ADVISORY SUBSET. AUROC COLUMN USES A SEPARATE N=100 HIGH-CONSENSUS SLICE (9 DRIFT CASES; ±10PP CI); FOR WITHIN-TABLE DIRECTIONAL COMPARISON ONLY.

Configuration	F1	Prec.	Rec.	FP	AUROC
Authority only	58.3	100.0	41.2	0	—
Coherence only	80.1	67.6	98.2	262	—
Scope only	91.0	100.0	83.5	0	—
Weighted-mean adj.	95.7	100.0	91.7	0	72.4
Sensitivity-wtd only	84.2	88.8	80.0	56	—
Full IDV (hybrid)	94.5	97.0	92.1	16	81.8

contributes 4 of the 7 reverse-direction disagreements (Claude flags drift, GPT-5.5 does not)—the highest of the three departments—indicating the Finance boundary is bidirectional rather than systematically biased toward one judge.

TABLE VI
CONDITIONAL INTER-JUDGE DISAGREEMENT RATES BY GPT-5.5 LABEL CLASS (NATURALISTIC N=150). THE 82.4% RATE IS CONDITIONAL ON GPT-5.5 LABELING OBJECT-TYPE, COMPUTED OVER N=34 SUCH TRACES; CIS ARE CLOPPER-PEARSON EXACT.

GPT-5.5 label	N	Disagreements	P(disagree)
Object	34	28	**82.4%** [65–93%]
Other drift	18	10	55.6%
no_drift	98	7	7.1%

Object-type cases drive the divergence. Of the 38 single-direction disagreements, 28 (74%) are Object-type drift involving instance-level entity access. Table VI reports conditional disagreement rates by GPT-5.5 label class with denominators.

Object cases comprise 22.7% of GPT-5.5 predictions but 62.2% of disagreements, a 2.7× enrichment. This domain-specific divergence differs from known position and verbosity biases: it emerges under structured intent specifications and category-to-instance granularity mismatch.

Methodological significance. To our knowledge, this specific pattern of systematic single-direction calibration divergence under category-level IAO has not been characterized in the financial-agent monitoring context. Related LLM-as-Judge literature documents position bias [8] and verbosity bias [7] in general settings, not domain-specific divergence under

structured intent specifications. Under the shared evaluation prompt, the divergence persists; whether sophisticated prompting strategies (multi-agent debate, IAO-expansion prompts) could reduce it is open future work. We view the divergence as a structural property of how each frontier LLM has calibrated category-level authorization statements against instance-level access events. IDV does not eliminate this divergence; rather, an external customer-context-aware monitoring layer with explicit per-customer parameters externalizes it into auditable configuration, making the boundary configurable rather than implicit.

VI. Discussion

The calibration-divergence result has operational implications. SOX Section 404, PCAOB AS 2201, and FINRA Rule 2111 establish control, model-risk, and suitability obligations under which authorization traceability and customer-context specificity become operationally material for compliance review. A monitoring layer relying on an LLM's interpretation of "customer accounts" produces a 38% inconsistency rate on Finance traces, making it difficult to rely on as a sole arbiter for SOX-, PCAOB-, or FINRA-sensitive workflows.

The implication extends beyond finance: any deployment where intent is specified at one granularity and execution acts at another faces analogous calibration risk for LLM-only monitoring.

The five observations—LLM-Judge divergence on Object cases, DriftScore distribution shift, Finance-concentrated IDV FPs, precision as structural differentiator, and IDV–LLM-Judge convergence on easy subsets—support a single thesis: the deployment challenge is not LLM capability alone, but the gap between category-level intent specification and instance-level access reality. Externalizing customer-context interpretation into structured IAOs and per-customer calibration makes the boundary configurable and auditable rather than implicit.

A. Supplementary Observation: Generation-Time Safety Refusal

During naturalistic trace generation, Claude Sonnet 4.6 declined to generate a subset of high-financial-impact pressure scenarios (notably a Finance scenario describing $2M payment authorization without standard approval workflow); refused variants were generated by GPT-4o instead. The resulting generator split is Claude:GPT-4o = 88:62 rather than the originally targeted 75:75. The class of action that a safety-calibrated frontier LLM declines to generate is the same class IDV is designed to detect at deployment for less-safety-calibrated agents: high-impact Finance actions that bypass authorization workflow. This observation is consistent with the paper's motivation, but we do not treat it as primary validation evidence.

B. Limitations and Future Work

Subtype recall is low for Rationale (22.5%) and Object (42.9%) because both fold into their parent engine categories at the adjudicator level, prioritizing binary precision over fine-grained subtype accuracy. All reported evaluations use DeBERTa-v3-small for NLI due to throughput constraints; larger NLI models may improve subtype recall, but were not evaluated at IDV-1000 scale.

Generator-judge correlation. 88 of the 150 naturalistic traces are generated by Claude Sonnet 4.6 and judged in part by Claude Opus 4.7. Residual within-family correlation cannot be ruled out. We note, however, that the divergence direction is GPT-5.5=drift / Claude=no-drift on 84% of cases—the opposite of what naive within-family bias would predict (Claude-judged Claude-generated traces should agree more if anything). Independent expert annotation remains the gold standard and is deferred to future work.

Other limitations. Threshold calibration is pre-specified with post-hoc sensitivity checks, not held-out test calibration. Production distributions may differ from our balanced department construction.

Future work: interactive IAO refinement, rolling-baseline-driven IAO learning, ambiguity-aware response routing, and automatic per-IAO-style threshold calibration.

VII. Conclusion

We presented IDV, a customer-context-aware multi-engine architecture for AI-agent intent drift detection, evaluated primarily on financial-services traces as the most regulated and structurally demanding deployment context. Two frontier LLMs disagree on `drift_detected` for 30% of naturalistic traces overall (38% Finance); conditional on Object-type label, the disagreement rate is 82.4%. IDV achieves F1=93.9% (precision 95.9%) on N=850 author-injected traces with sub-300ms latency, delivering a $6.5\times$ false-positive reduction and 16.4pp precision advantage relative to JSON-only LLM-Judge baselines. The DriftScore shift and LLM-Judge divergence trace to a single structural property: category-level intent specifications must be reconciled against instance-level execution.

References

[1] Y. Bai et al., "Constitutional AI: Harmlessness from AI feedback," *arXiv:2212.08073*, 2022.

[2] V. Barres et al., "tau-squared-Bench: Evaluating conversational agents in a dual-control environment," *arXiv:2506.07982*, 2025.

[3] H. Inan et al., "Llama Guard: LLM-based input-output safeguard for human-AI conversations," *arXiv:2312.06674*, 2023.

[4] C. E. Jimenez et al., "SWE-bench: Can language models resolve real-world GitHub issues?" *ICLR*, 2024.

[5] OpenAI, "Introducing SWE-bench Verified," Aug. 2024.

[6] L. Ouyang et al., "Training language models to follow instructions with human feedback," *NeurIPS*, 2022.

[7] K. Saito, A. Wachi, K. Wataoka, Y. Akimoto, "Verbosity bias in preference labeling by large language models," *arXiv:2310.10076*, 2023.

[8] L. Shi, C. Ma, W. Liang, W. Ma, S. Vosoughi, "Judging the judges: A systematic study of position bias in LLM-as-a-judge," *IJCNLP-AACL*, 2025.

[9] O. Styles et al., "WorkBench: A benchmark dataset for agents in a realistic workplace setting," *arXiv:2405.00823*, 2024.

[10] A. Wei, N. Haghtalab, J. Steinhardt, "Jailbroken: How does LLM safety training fail?" *NeurIPS*, 2023.

[11] F. F. Xu et al., "TheAgentCompany: Benchmarking LLM agents on consequential real world tasks," *NeurIPS Datasets and Benchmarks*, 2025.

[12] R. R. Yager, "On ordered weighted averaging aggregation operators in multicriteria decisionmaking," *IEEE Trans. Syst. Man Cybern.*, vol. 18, no. 1, pp. 183–190, 1988.

[13] S. Yao, N. Shinn, P. Razavi, K. Narasimhan, "tau-Bench: A benchmark for tool-agent-user interaction in real-world domains," *arXiv:2406.12045*, 2024.
[14] L. Zheng et al., "Judging LLM-as-a-judge with MT-bench and Chatbot Arena," *NeurIPS Datasets and Benchmarks*, 2023.
[15] A. Zou, Z. Wang, J. Z. Kolter, M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *arXiv:2307.15043*, 2023.
[16] Z. Yang et al., "FinVault: Benchmarking financial agent safety in execution-grounded environments," *arXiv:2601.07853*, 2026.
[17] J. Lu et al., "FinToolBench: Evaluating LLM agents for real-world financial tool use," *arXiv:2603.08262*, 2026.
[18] Y. Hou et al., "FinSafetyBench: Evaluating LLM safety in real-world financial scenarios," *arXiv:2605.00706*, 2026.
[19] OpenAI, "Introducing GPT-5.5," Apr. 23, 2026. [Online]. Available: https://openai.com/index/introducing-gpt-5-5/
[20] Anthropic, "Introducing Claude Opus 4.7," Apr. 16, 2026. [Online]. Available: https://www.anthropic.com/news/claude-opus-4-7