Adaptive Domain Intelligence: An Honest LLM-as-Feature-Engineer Architecture for Compressing Qualitative Records into Calibrated Empirical Predictions

David C. Liu
Independent Research
daliu.github.io

June 2026 *(working paper)*

Abstract

A recurring class of high-stakes problems asks us to compress a qualitative record—a court opinion, a clinical note, a field dispatch—into an empirical verdict: violation or not, cured or not, escalation or not. Large language models read such records fluently, but two failures block their use as trustworthy predictors here: they fabricate, and they degrade when allowed to judge their own improvements. We present **metis**, an architecture that confronts both. An LLM is confined to two roles—proposing a feature schema and performing *verbatim-quote-grounded extraction*—while a small, *calibrated* classical model makes the actual prediction; any extracted value lacking a verbatim evidence span is dropped and counted against a measured hallucination rate. The system improves itself only when a domain expert *and* a frozen gold set the loop never fits to both endorse a change, adjudicated by a Bonferroni-corrected permutation test with a hard subgroup-fairness veto—an external gate answering the now-robust finding that LLMs cannot reliably self-correct without ground truth. We evaluate not by chasing a benchmark but as a **known-answer test**: across four real legal corpora spanning six decades of empirical judicial-politics findings, metis independently re-derives the established signals—judge party predicts decision *direction* (ΔAUC +0.034, p=0.002), court ideology adds incremental lift, litigant "repeat-player" status predicts *settlement* (+0.016, p=0.002)—and returns the established *nulls*: judge attributes do not predict whether parties settle (≈ 0, $p > 0.4$) and appellate panel composition washes out ($p \geq 0.37$). It also exhibits the honesty its design targets: it vetoes one of its own expert-approved escalations on held-out evidence, self-rejects features that do not help, and on real European Court of Human Rights texts extracts ∼1,200 features with a measured-zero hallucination rate yet lands at chance on the outcome—the correct answer to a documented leakage trap in which prior text-only work reported ∼79% by reading court-authored facts written *after* the decision. We offer the composition, and the known-answer methodology that tests it, as a general recipe for adaptive domain intelligence.

1 Introduction

Consider three questions from three professions. A human-rights lawyer asks whether a set of facts will be found to violate the Convention. An oncologist asks whether a patient described across pages of notes will be readmitted. An analyst asks whether a region's dispatches portend escalation. Each compresses a rich qualitative record into a small empirical verdict, under conditions that make naive machine learning treacherous: the data are scarce (hundreds to thousands of labeled cases), the

stakes are high, the output must be *auditable* by an accountable domain expert, and that expert's judgment is the scarcest signal available.

LLMs are obvious candidates but two failures stand in the way. First, they **fabricate**. Second, **they cannot reliably improve themselves by judging their own work**: intrinsic self-correction without an external signal does not help and often degrades performance [12]; apparent self-improvement is usually an external verifier recognizing an already-present answer [29]; a critical survey finds no success from prompted-LLM feedback alone [14]; and LLM judges favor their own generations [24]. A system that rewrites its own logic and grades the result with the same model family builds an echo chamber.

The legal-prediction literature supplies a sharp cautionary instance. Text-only prediction of European Court of Human Rights (ECtHR) outcomes reported ∼79% accuracy [1], with the case "Facts" section the strongest predictor—but those facts are written by the Court *after* it decides, so a survey reframed much of the field as retrospective *identification* rather than *forecasting* [21]. The more honest benchmark—predicting U.S. Supreme Court outcomes from pre-decision features—sits near 70% [15]. We take these failures as design constraints. Our contributions: (i) an architecture combining LLM-*proposed* grounded extraction, a calibrated classical decider, and an automated held-out statistical + fairness gate on the system's own self-modifications; (ii) the *known-answer test*, a validation methodology using domains with established empirical regularities as a behavioral test bed; and (iii) honesty results the benchmark framing cannot produce.

2 The architecture: four disciplines

The LLM as an untrusted feature-engineer. The LLM proposes a schema—quantifiable features, each tagged with the record section it may legally be read from—and extracts each feature, returning a value, calibrated confidence, and one or more verbatim evidence spans. The validator verifies each span is an exact substring of the source; unsupported values are dropped and counted against a per-run hallucination rate. The LLM never sees the label and never makes the call. This is stricter than CHiLL [20] and FeatLLM [11]: the verbatim-quote requirement makes grounding a hard structural precondition, not a post-hoc citation or an LLM-judged faithfulness score [25].

The calibrated classical decider. The prediction is a small calibrated model over named features—sparse logistic regression by default (signed coefficients we can show the user), with evidence-gated escalation to gradient boosting. Calibration is a gate: expected calibration error above 0.10 is treated as not-deployable [10]. Abstention is first-class via a conformal layer [2]; high-stakes and uncertain cases escalate to a human [18]. That a small model on a few features matches an opaque one is not new [4]; the contribution is the provenance of the features and the gating.

The dual gate. Given the evidence that LLMs cannot reliably self-evaluate, every accept/reject verdict moves to two non-LLM arbiters. A *frozen gold set* the loop never fits to: a change must improve held-out performance by a margin, with a paired-bootstrap CI above zero, *and* survive a Bonferroni-corrected permutation test; a parsimony drop need only be non-inferior. A *domain expert* who adjudicates clustered, named-feature failure cards. A repair applies only when both agree. A hard subgroup-fairness veto sits over both [16, 3]. Crucially the evaluator is not in the loop's edit set: the system improves the engine, never rewrites the judge to agree with itself—the configuration the self-correction literature endorses (external signal: 31, 27).

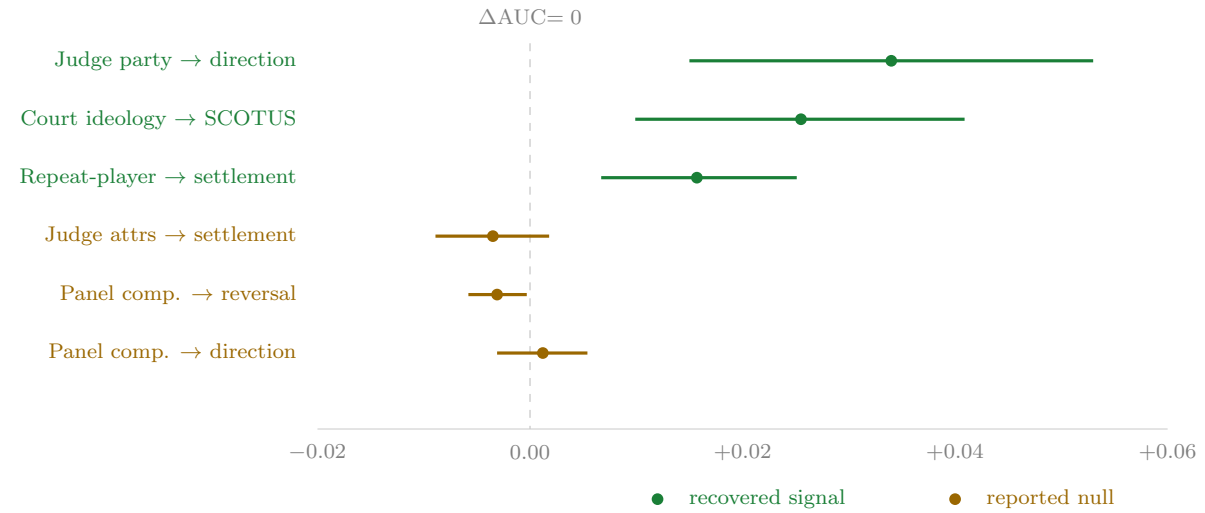

Figure 1: Paired ΔAUC with 95% confidence intervals, every estimate under a leakage-safe split. The three signals sit cleanly above zero; the three nulls straddle it. A trustworthy adaptive agent must produce *both*—recover the real effects and decline the absent ones. (Stakes×type→duration is omitted as it is measured in error, not AUC; see Table 1.)

Table 1: Known-answer results. All under leakage-safe CV with paired significance tests.

Relationship (corpus)	CV split	Effect	Verdict
Judge party → direction (Carp–Manning, civil lib.)	unseen-judge	+0.034 [.015,.053], p=.002	signal
Court ideology → SCOTUS, incremental (SCDB+MQ)	chronological	+0.026/+0.012, p=.002	signal
Repeat-player status → settlement (FJC)	district-grouped	+0.016 [.007,.025], p=.002	signal
Stakes × type → duration (FJC)	district-grouped	+0.076 err↓, p=.002	signal
Judge attributes → settlement (FJC)	judge-grouped	≈ 0, $p > .4$	null
Panel composition → reversal (Songer)	lead-judge-grp.	−0.003, p=.84	null
Panel composition → direction (Songer)	lead-judge-grp.	+0.001, p=.37	null

Leakage discipline. A provenance firewall quarantines post-outcome features; a univariate audit flags any single feature that alone separates the classes; a grounding-span scan drops outcome-revealing evidence; and real data is cross-validated chronologically or grouped by entity (judge, court, party), never naively. Every result below uses a leakage-safe split.

3 Evaluation as a known-answer test

Six decades of empirical judicial politics supply ground truth: party shapes how judges *decide* [26, 19], capability shapes who *wins and settles* [6, 28], and panel effects are real but domain-specific [30]. A trustworthy agent must recover the real effects, decline the absent ones, and be honest about its machinery. We use four real corpora—the FJC Integrated Database, the Carp–Manning U.S. District Court Database (∼111k decisions), the Supreme Court Database joined to Martin–Quinn ideology, and the Songer Courts of Appeals Database—plus real ECtHR texts. Every number is re-derived deterministically by an automated verifier (8/8 reproduce, zero drift).

The pattern (Figure 1, Table 1) is coherent judicial politics: ideology shows up where a judge decides alone, attenuates on three-judge panels, aggregates at the court level, and never reaches

Table 2: Honest self-behavior and extraction faithfulness.

Behavior	Result
Frozen gold vetoes expert-approved escalation (Carp–Manning)	Δ=+0.000, vetoed
Capacity probe flips as features accrue (FJC duration)	boost≈linear → +0.076 gain
Financial-status features self-rejected (SEC-EDGAR)	3.6% match, marginal −0.013
Session-LLM hallucination rate (real ECtHR, ∼1,200 values)	0.0000
ECtHR outcome prediction (honest answer to leakage trap)	AUC 0.504, chance
Conformal interval coverage on sealed real data (FJC duration)	80.7% vs. 80% target

into whether litigants settle. The nulls match Sunstein et al.'s finding that panel effects concentrate in a few hot-button domains and wash out across a broad corpus.

Table 2 carries the spine. On real ECtHR text the extraction was faithful by measurement—zero fabricated quotes across ∼1,200 values—yet the prediction landed at chance. This is the *correct* result: the architecture separates *did the model extract honestly?* (yes) from *do those features predict?* (no, on the pre-judgment framing), and refuses to manufacture the second from the first. Where prior text-only work reported ∼79% by reading court-authored facts, an honest, leakage-careful pipeline lands at chance—and says so.

4 Related work and novelty

The pieces exist separately: LLM-as-feature-engineer for a small interpretable model [20, 11]; verbatim grounding and fine-grained faithfulness [32, 25, 22, 7]; the finding that self-correction needs an external signal [12, 29, 14]; calibration, conformal abstention, and the fairness impossibility results [10, 2, 16, 3]; and the LLM-as-annotator validity literature [8, 23, 5, 9]. To our knowledge no prior system combines LLM-*proposed* grounded extraction, a calibrated classical decider, and an automated held-out statistical + fairness gate on its own self-modifications, validated as a known-answer test across real corpora. The contribution is an honesty-and-calibration layer the benchmark branch has lacked, and the external-gate answer the self-correction literature demands.

5 Limitations and ethics

Outcome prediction reproduces historical patterns, including injustice; the fairness veto, abstention, and escalation reduce but do not remove this. The system is decision-*support* only—an architectural refusal layer blocks prohibited contexts (bail, sentencing, recidivism, immigration enforcement) regardless of confidence. Real-data samples are in the thousands; some source party names are truncated, capping one financial join; the expert gate is only as good as the expert, and authoring the frozen gold set is the binding human cost. A human audit of a sample of machine extractions is pending; this is a working paper.

6 Conclusion

The valuable and dangerous problems—violation or not, cured or not, escalation or not—ask us to compress a qualitative record into an empirical verdict under scarcity and high stakes. Pointing a large model at the text and reporting the accuracy yields a number that is often leakage, from a system that drifts when it grades itself. metis confines the LLM to grounded extraction, hands the decision to a small calibrated model, and changes itself only when an expert and a frozen gold

set agree. Validated where the answers are known, it recovers the real signals, reports the real nulls, vetoes its own bad ideas, and measures its own hallucination at zero while admitting when its features do not predict—a general recipe for adaptive domain intelligence wherever qualitative evidence must become an auditable empirical claim.

References

[1] Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016). Predicting Judicial Decisions of the European Court of Human Rights. *PeerJ Computer Science*, 2:e93.

[2] Angelopoulos, A. N., & Bates, S. (2021). A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *arXiv:2107.07511*.

[3] Chouldechova, A. (2017). Fair Prediction with Disparate Impact. *Big Data*, 5(2), 153–163.

[4] Dressel, J., & Farid, H. (2018). The Accuracy, Fairness, and Limits of Predicting Recidivism. *Science Advances*, 4(1):eaao5580.

[5] Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2024). RAGAS: Automated Evaluation of Retrieval Augmented Generation. *EACL 2024, System Demonstrations*, 150–158.

[6] Galanter, M. (1974). Why the 'Haves' Come Out Ahead. *Law & Society Review*, 9(1), 95–160.

[7] Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling Large Language Models to Generate Text with Citations. *EMNLP 2023*.

[8] Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT Outperforms Crowd Workers for Text-Annotation Tasks. *PNAS*, 120(30):e2305016120.

[9] Gu, J., et al. (2025). A Survey on LLM-as-a-Judge.

[10] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. *ICML 2017*.

[11] Han, S., Yoon, J., Arik, S. O., & Pfister, T. (2024). Large Language Models Can Automatically Engineer Features for Few-Shot Tabular Learning (FeatLLM). *ICML 2024*.

[12] Huang, J., et al. (2023). Large Language Models Cannot Self-Correct Reasoning Yet. *ICLR 2024* (arXiv:2310.01798).

[13] Kamath, A., Jia, R., & Liang, P. (2020). Selective Question Answering under Domain Shift. *ACL 2020*.

[14] Kamoi, R., Zhang, Y., Zhang, N., Han, J., & Zhang, R. (2024). When Can LLMs Actually Correct Their Own Mistakes? *TACL*, 12, 1417–1440.

[15] Katz, D. M., Bommarito, M. J., & Blackman, J. (2017). A General Approach for Predicting the Behavior of the Supreme Court. *PLOS ONE*, 12(4):e0174698.

[16] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent Trade-Offs in the Fair Determination of Risk Scores. *ITCS 2017*.

[17] Madaan, A., et al. (2023). Self-Refine: Iterative Refinement with Self-Feedback. *NeurIPS 2023*.

[18] Madras, D., Pitassi, T., & Zemel, R. (2018). Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. *NeurIPS 2018*.

[19] Martin, A. D., & Quinn, K. M. (2002). Dynamic Ideal Point Estimation via MCMC for the U.S. Supreme Court. *Political Analysis*, 10(2), 134–153.

[20] McInerney, D. J., Young, G., van de Meent, J.-W., & Wallace, B. C. (2023). CHiLL: Zero-shot Custom Interpretable Feature Extraction from Clinical Notes. *Findings of ACL 2023*.

[21] Medvedeva, M., Wieling, M., & Vols, M. (2023). Rethinking the Field of Automatic Prediction of Court Decisions. *Artificial Intelligence and Law*, 31, 195–212.

[22] Min, S., et al. (2023). FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. *EMNLP 2023*.

[23] Mohta, J., Ak, K. E., Xu, Y., & Shen, M. (2023). Are Large Language Models Good Annotators? *PMLR*, 239, 38–48 (NeurIPS 2023 Workshop).

[24] Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM Evaluators Recognize and Favor Their Own Generations. *NeurIPS 2024*.

[25] Rashkin, H., et al. (2023). Measuring Attribution in Natural Language Generation Models. *Computational Linguistics*, 49(4).

[26] Segal, J. A., & Spaeth, H. J. (2002). *The Supreme Court and the Attitudinal Model Revisited*. Cambridge University Press.

[27] Shinn, N., et al. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. *NeurIPS 2023*.

[28] Songer, D. R., Sheehan, R. S., & Haire, S. B. (1999). Do the 'Haves' Come Out Ahead Over Time? *Law & Society Review*, 33(4), 811–832.

[29] Stechly, K., Valmeekam, K., & Kambhampati, S. (2024). On the Self-Verification Limitations of LLMs on Reasoning and Planning Tasks. *arXiv:2402.08115*.

[30] Sunstein, C. R., Schkade, D., & Ellman, L. M. (2004). Ideological Voting on Federal Courts of Appeals. *Virginia Law Review*, 90(1), 301–354.

[31] Zelikman, E., Wu, Y., Mu, J., & Goodman, N. D. (2022). STaR: Bootstrapping Reasoning With Reasoning. *NeurIPS 2022*.

[32] Zhang, J., Marone, M., Li, T., Van Durme, B., & Khashabi, D. (2025). Verifiable by Design: Aligning Language Models to Quote from Pre-Training Data. *NAACL 2025*.