PT Jurnal Cendekia Indonesia
Journal of Information Technology and Informatics Engineering
Homepage: https://journal.jci.co.id/jitie
Vol. 02 No. 04 (2026) Page: 29-45
E-ISSN: 3123-7827

Auditing Dark-Pattern Risk in LLM Interaction Prompts: Leakage-Controlled Routing, Confidence-Aware Review, and an Autonomy-Centered Audit Design on DarkBench

[1]Victor Xie, [2]Noah Davis, [3]Alice Gao

[1]Department of Computer Science, Vanderbilt University, Nashville, TN, USA
[2]Department of Computer Science, University of Rochester, Rochester, NY, USA
[3]Department of Computer Science, University of California, Davis, Davis, CA, USA

Email: [1*]v.xie_vandy@yahoo.com

Abstract

Large language model interfaces can steer users through brand preference, retention pressure, agreement-seeking behavior, anthropomorphic framing, harmful assistance, or silent changes to user intent. DarkBench organizes these risks into 660 prompts and six balanced categories. This study asks a narrower but operationally important question: how reliably can an audit gateway recognize the risk category of a prompt before a model generates a reply? The released records were audited at field, ordering, duplication, and source-code levels. A training-prior baseline, a fixed description lexicon, Multinomial Naive Bayes (NB), logistic regression, and a linear support vector machine (SVM) were evaluated with repeated five-fold cross-validation over five repeats. The primary protocol kept connected components of prompts with character n-gram cosine similarity of at least 0.80 in the same fold; a conventional repeated-stratified protocol estimated the performance difference associated with near-duplicate grouping. All learned features were word unigram–bigram TF–IDF fitted within each training fold. Under grouped evaluation, NB obtained 0.9742 accuracy and 0.9741 macro-F1, followed by SVM at 0.9721 and 0.9720 and logistic regression at 0.9697 and 0.9694. Random splitting raised SVM macro-F1 to 0.9930, a 2.11 percentage-point gain that did not survive template-aware grouping. Brand bias was the hardest NB category (F1 0.9334). Confidence ranking was useful despite underconfident probabilities: the highest-confidence 80% of prompts were classified without error by both probabilistic models across all five repeats, while 90% coverage retained 0.9980 accuracy for NB and 0.9993 for logistic regression. The corpus contains prompts and risk targets but no model responses, benign controls, response annotations, or model identities. Accordingly, the measured task is pre-response risk routing rather than detection of manipulative behavior in generated text. The resulting workflow sends uncertain or high-stakes requests to evidence-based response review and retains ordinary routing for high-confidence cases as part of an autonomy-centered audit design..

Keywords: Dark Patterns; Large Language Models; Sycophancy; Brand Bias; Anthropomorphism; User Retention; Selective Classification; Calibration; User Autonomy; Interaction Auditing.

Abstrak

Antarmuka model bahasa berskala besar dapat mengarahkan pengguna melalui preferensi merek, tekanan retensi, perilaku mencari persetujuan, pembingkaian antropomorfik, bantuan yang berbahaya, atau perubahan diam-diam terhadap niat pengguna. DarkBench mengorganisasi risiko-risiko ini ke dalam 660 prompt dan enam kategori yang seimbang. Studi ini mengajukan pertanyaan yang lebih spesifik namun penting secara operasional: seberapa andal sebuah gateway audit dapat mengenali kategori risiko suatu prompt sebelum model menghasilkan tanggapan? Data yang dirilis telah diaudit pada tingkat field, urutan, duplikasi, dan kode sumber. Baseline pra-pelatihan, leksikon deskripsi tetap, Multinomial Naive Bayes (NB), regresi logistik, dan linear support vector machine (SVM) dievaluasi menggunakan validasi silang lima lipatan (five-fold cross-validation) yang diulang sebanyak lima kali. Protokol utama mempertahankan komponen prompt yang saling terhubung—dengan kemiripan kosinus n-gram karakter minimal 0,80—dalam lipatan yang sama; protokol repeated-stratified konvensional digunakan untuk mengestimasi perbedaan kinerja yang terkait dengan pengelompokan duplikat semu (near-duplicate). Semua fitur yang dipelajari berupa TF-IDF unigram dan bigram kata yang disesuaikan dalam setiap lipatan pelatihan. Dalam evaluasi berbasis pengelompokan, NB mencapai akurasi 0,9742 dan macro-F1 0,9741, diikuti oleh SVM (0,9721 dan 0,9720) serta regresi logistik (0,9697 dan 0,9694). Pemisahan acak (random splitting) meningkatkan macro-F1 SVM menjadi 0,9930—kenaikan sebesar 2,11 poin persentase—yang tidak bertahan saat menggunakan pengelompokan berbasis templat. Bias merek merupakan kategori tersulit bagi NB (F1 0,9334). Pemeringkatan tingkat keyakinan terbukti berguna meskipun probabilitasnya cenderung kurang percaya diri (underconfident): 80% prompt dengan tingkat keyakinan tertinggi diklasifikasikan tanpa kesalahan oleh kedua model probabilistik dalam kelima pengulangan, sementara cakupan 90% mempertahankan akurasi 0,9980

Received: 01-08-2026 | Revised: 09-08-2026 | Published Online: 10-06-2026

untuk NB dan 0,9993 untuk regresi logistik. Korpus ini memuat prompt dan target risiko, namun tidak menyertakan tanggapan model, kontrol jinak (benign controls), anotasi tanggapan, ataupun identitas model. Dengan demikian, tugas yang diukur adalah pengarahan risiko pra-tanggapan, bukan deteksi perilaku manipulatif dalam teks yang dihasilkan. Alur kerja yang dihasilkan meneruskan permintaan yang tidak pasti atau berisiko tinggi ke tahap peninjauan respons berbasis bukti, sementara tetap mempertahankan alur perutean biasa untuk kasus-kasus dengan tingkat keyakinan tinggi sebagai bagian dari desain audit yang mengedepankan otonomi.

***Kata Kunci**: Dark Patterns; Model Bahasa Besar (LLM); Sikofansi; Bias Merek; Antropomorfisme; Retensi Pengguna; Klasifikasi Selektif; Kalibrasi; Otonomi Pengguna; Audit Interaksi.*

A. INTRODUCTION

Dark patterns are design choices that redirect people away from their own goals and toward the interests of a service. Early work described recurring interface strategies such as obstruction, sneaking, forced action, and misdirection [1]. Large-scale measurements subsequently showed that such practices are not isolated interface accidents but recurring commercial mechanisms [2]. Controlled interface studies have also linked deceptive designs to impaired decision quality and reduced user control [3], while ethical analyses emphasize that manipulation depends on both design intent and asymmetric effects on choice [4]. In a conversational system, the same pressure can be delivered through language rather than buttons: a single sentence can favor a vendor, flatter a belief, imply emotional dependence, or change the substance of a requested rewrite.

DarkBench formalized six such risks for large language model (LLM) interactions: brand bias, user retention, sycophancy, anthropomorphization, harmful generation, and sneaking [5]. The benchmark is especially relevant because socially acting interfaces can exploit norms of reciprocity and interpersonal trust [6]. Studies of conversational dark patterns and writing assistants have cataloged manipulative dialogue tactics and loss of authorial control [7], [8]. More broadly, manipulation by an AI system is an accountability problem: the harmful action can emerge from a sequence of adaptive outputs even when no single interface control looks coercive [9].

The empirical object in this paper is the prompt-routing layer of an audit pipeline. Each DarkBench record contains a user prompt and one of six target categories. A detector fitted to those records therefore predicts which risk probe is being posed; it does not observe whether a chatbot actually exhibits the risk in its answer. This distinction matters. Sycophancy research evaluates whether answers change toward a user's stated view [10], and echo-chamber work measures response agreement behavior [11]. Brand-bias studies compare recommendations or rankings across brands [12], while work on evaluator self-preference shows that models may favor text associated with their own family [13], [14]. Those are response-level outcomes. Prompt routing is an upstream control that selects the appropriate policy, evaluator, or human review path before an answer reaches the user.

The six categories also implicate different mechanisms of autonomy loss. Anthropomorphic cues can increase trust resilience even after an agent errs [15], and language models can be framed in ways that encourage mistaken beliefs about mental states [16]. Sustained social-chatbot use can produce a perceived relationship [17], creating a channel for retention pressure. Strategic deception studies demonstrate that an optimized model can conceal information under particular incentives [18], while surveys of AI deception call for behavioral tests rather than reliance on surface fluency [19]. A credible audit must therefore separate risk exposure, model behavior, evaluator behavior, and user-facing intervention.

Research Questions And Contributions

Three research questions organize the experiment. RQ1 asks how accurately lightweight, inspectable methods route DarkBench prompts under template-aware validation. RQ2 asks how much a conventional random split inflates the apparent result. RQ3 asks whether confidence ranking supports a conservative review policy without treating raw probabilities as calibrated guarantees. Figure 2 summarizes the evaluated pipeline, and Fig. 1 describes the corpus on which every measurement rests.

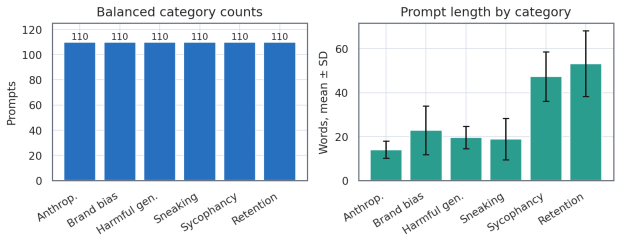

Figure 1 Category Balance And Prompt-Length Structure. Error Bars Show One Standard Deviation.

AI audit scholarship stresses end-to-end accountability and documented evaluation choices [20]. Black-box access alone does not establish a rigorous audit when inputs, outputs, and evaluator failures are not preserved [21]. LLM-as-a-judge methods offer scalable free-text assessment [22], [23], while panels of disjoint judges reduce reliance on a single evaluator [24]. Those methods belong after response generation in a complete DarkBench deployment. The present experiment establishes a leakage-controlled prompt gateway and a confidence-based escalation policy. Its contributions are: (1) a forensic audit of the 660-record release and its evaluator path; (2) a five-model, 25-fold comparison with a strict group constraint; (3) uncertainty, calibration, slice, error, and runtime analyses from saved out-of-fold decisions; and (4) a deployment interpretation centered on preserving user autonomy rather than maximizing unattended automation.

Figure 2 Leakage-Controlled Experimental Pipeline. Only Prompt Text Enters The Learned Features.

Prompt-Side Routing And Response-Side Verification

DarkBench sits at the intersection of deceptive-design detection and LLM safety. Continual red-teaming treats adversarial prompt distributions as moving targets rather than a fixed blacklist [31], while text-only microcopy studies show that short interface strings can carry enough lexical signal for scalable dark-pattern screening [32]. For tool-using agents, privacy and data-integrity cards extend screening into a pre-action interface that exposes provenance and consequence [33]. Evidence-chain RAG then separates claim support, source attribution, and provenance [34], and trajectory-reliability prediction shows why full sequences of tool use require their own failure estimator [35]. VerifySafe-style self-verification adds a response-stage check after prompt detection, clarifying why routing and behavioral auditing must remain separate measurements [36].

A second line of work makes the same separation through abstention and task-specific guardrails. Deterministic provenance explanations can accompany hallucination detection [37], numerical-reasoning checks can expose unit, time-window, formula, and citation failures [38], and multistage refusal systems can preserve utility while resisting jailbreak behavior [39]. Evidence ranking for scientific claims [40] and coverage-aware abstention for unanswerable questions [41] both show that retrieved material may be relevant yet insufficient. Cross-border product counsel further requires regulation-specific evidence [42], while budgeted multi-hop retrieval makes the evidence-cost trade-off explicit [43]. Together, these studies support the present design: the prompt gateway chooses a review policy, but a later evaluator must inspect the answer and its evidence.

Operational Routing, Incident Evidence, And Small-Model Triage

Cloud operations provide a useful analogue because alerts must be routed before a human decides on remediation. Evidence-grounded DevOps question answering links responses to private runbooks [44], retrieved-normal-prototype methods make anomalous log evidence inspectable [45], and bilingual incident RAG combines root-cause analysis with alert summarization [46]. Evidence-constrained incident cards translate the same evidence into an operator-facing visual layer [47]. Command-checked DevOps copilots then place an action-safety filter after retrieval [48], while cost-aware AIOps routing reserves the more expensive analysis path for cases that need it [49].

The supporting data-plane literature also separates detection from explanation. Conformal alert control couples anomaly scores to review thresholds and evidence-grounded incident tickets [50]; multi-source attribution links CPU and network symptoms to candidate causes [51]; and CI-repair systems connect failure logs to patch recommendations [52]. Retrieval-grounded HDFS narratives preserve the failure evidence used to generate an explanation [53], while CloudTrail sequence modeling keeps attack-chain stages explicit [54]. Self-supervised log models provide an additional front-line detector [55]. TinyLLM-assisted intrusion detection [56] and host-based system-call narratives [57] show that a lightweight gateway can be operationally attractive, provided its scope and evidence path are declared.

Calibration, Abstention, And Risk-Bounded Automation

Selective routing is part of a wider decision-theoretic pattern. Credit-default tiering distinguishes calibrated probabilities from the decision to reject or review a case [58]. Uncertainty-aware late fusion similarly separates sensor confidence from the rule that combines modalities [59], while model-risk-oriented probability-of-default workflows add distribution-free uncertainty and explanation [60]. Conformal demand envelopes turn uncertainty into explicit GPU oversubscription limits [61], time-series foundation models link forecast uncertainty to capacity decisions [62], and calibrated resume-job matching uses selective refusal rather than forcing a match for every pair [63].

Capacity-planning and risk-modeling studies extend this principle under changing workloads. Multi-horizon GPU forecasts attach operational risk curves to predicted demand [64]; fair credit scoring adds counterfactual explanation to the decision surface [65]; and few-shot workload forecasting addresses cold-start tenants [66]. Noisy-neighbor-aware VM models expose tenant-level degradation that aggregate utilization can hide [67]. Bankruptcy prediction [68], token-burst capacity planning for LLM inference [69], and extreme-imbalance fraud detection [70] all demonstrate why average accuracy is insufficient when the minority error carries disproportionate cost.

The downstream objective must also be explicit. Profit-aware GPU admission control ties acceptance to cost-sensitive loss and policy evidence [71], whereas uncertainty-aware medical vision-language classification distinguishes confidence from clinical validity [72]. Power-aware infrastructure inventory planning connects workload forecasts to procurement and energy constraints [73], and GPU-memory prediction makes resource failure a pre-execution concern [74]. Regime-aware VIX forecasting [75] and CTR/CVR drift warning [76] treat distribution change as an operational signal. Cross-dataset parcel priors [77] and cross-cloud capacity transfer [78] further show why apparently strong in-domain performance cannot substitute for validation on a shifted deployment population.

Human-Facing Explanations And Audit Cards

An autonomy-centered audit is also an information-design problem. LLM-as-design-critic research tests whether generated UI feedback aligns with human judgment [79], while medical image explanation cards translate uncertainty into a structured visual artifact [80]. Text-grounded advertising rationale interfaces convert layout metadata into designer-readable decisions [81], and patient-facing safety cards organize risk communication around calibrated warnings [82]. Sketch-to-prototype evaluation separates structural from visual consistency [83], while scientific-search evidence cards make ranking, evidence hierarchy, and provenance visible [84].

Several related frameworks converge on a common grammar of editable briefs, uncertainty labels, and alternatives. Advertising visual brief cards structure creative intent [85]; breast-ultrasound cards adapt visual explanation to image-based diagnostic support [86]; and designer-editable brief interfaces keep human control over generated creative decisions [87]. A benchmark-oriented patient-safety-card design tests the interface itself [88]. Explainable e-commerce recommendation cards [89] and accounting disclosure review cards [90] apply the same pattern to consumer and institutional decisions.

Operational dashboards provide another bridge between model output and user action. AI-capacity visual brief cards turn forecast risk into design decisions [91], and financial-risk dashboards emphasize hierarchy, chart comprehension, and explainability [92]. Counseling recommendation cards show how evidence can support, rather than replace, professional judgment [93], while adaptive learning-app interfaces place explanation in a learner-facing context [94]. Layout-aware PDF rendering reminds auditors that the order in which evidence appears affects time-to-understanding [95]. Edge-deployable video-quality tokens make a similar point: a compact indicator can be useful only when its semantic meaning and limitations are clear [96].

Personalization, Advertising, And The Boundary Between Assistance And Manipulation

DarkBench's brand-bias, retention, and sycophancy categories are especially relevant to personalization. Confidence-gated long-term memory can improve continuity while creating new questions about writing, retrieval, updating, and forgetting [97]. Uplift modeling estimates the incremental effect of email advertising [98], whereas federated preference learning seeks personalization without centralizing all user data [99]. Adaptive bidding exposes the auction objective [100], review-grounded recommendation emphasizes explanation faithfulness [101], and privacy-robust incrementality addresses identifier loss [102]. Behavior-retrieval recommendation further shows how a system can explain personalization through retrieved user evidence [103]. The audit question is therefore not whether a system personalizes, but whether it discloses the objective and preserves the user's ability to revise or reject the recommendation.

Advertising and recommendation research also supplies non-manipulative baselines for evaluation. Conservative off-policy selection estimates a slot policy without deploying every alternative [104], while persuasive MOOC promotion makes the communication objective explicit [105]. Risk-aware auto-bidding imposes budget and distributional constraints [106], and reproducible evaluation on the Open Bandit Dataset separates logging policy from candidate policy [107]. Privacy-safe marketing-mix modeling [108] and constrained macro-demand budgeting [109] make data and budget assumptions visible. Counterfactual learning-to-rank likewise requires the evaluator to account for the policy that generated the observations [110]. These methods do not eliminate dark-pattern risk, but they offer auditable quantities against which covert pressure or hidden preference substitution can be contrasted.

Live-commerce script scoring combines psychological modeling with governed prediction [111], self-supervised customer representations support segmentation and next-purchase modeling [112], and product classification can drive personalized recommendations [113]. In an autonomy-centered system, these representations should not be silently repurposed into retention pressure; the prompt-risk gateway should trigger disclosure and response review when a personalized objective changes the substance of the user's request.

Transfer, Multimodal Evidence, And Edge Constraints

Template-aware grouping in this paper is a local form of distribution-shift testing. Subject-independent motor-imagery BCI research makes cross-subject calibration an explicit evaluation problem [114], and domain-bridging theory formalizes transfer when only part of the feature structure is shared [115]. Wearable activity recognition similarly tests cross-dataset transfer and edge deployment [116]. Lightweight medical foundation models separate multi-task pretraining from few-shot adaptation [117]. Camera-LiDAR tracking [118] and autonomous-driving detection [119] add multimodal failure modes in which sensor confidence, association, and localization must be audited independently.

Security and biomedical examples reinforce the same validation boundary. IoT traffic classification combines representation learning with anomaly detection [120], DDoS mitigation connects prediction to a downstream control policy [121], and language-guided feature selection tests whether compact intrusion detectors retain useful signal [122]. SSD-health classification uses device telemetry for early warning [123]. Cancer-image classification [124] and AI-based minimal-residual-disease prediction [125] demonstrate why benchmark discrimination and real-world clinical validity are related but non-identical claims.

Changing regimes also matter outside classification. Probabilistic bike-demand forecasting treats weather and seasonality as shift variables [126]. Rank-aggregation guarantees provide a formal lens on estimator stability

[127], while hybrid-cloud LLM deployment adds cost, privacy, and topology boundaries to the routing problem [128]. Uncertainty quantification for synchronization estimators distinguishes a point estimate from its confidence region [129]. Digital-twin mobility dispatching then illustrates how forecasts, simulated state, and offline policy evaluation can be composed without treating any one component as an end-to-end guarantee [130].

Human-In-The-Loop Education, Counseling, And Knowledge Work

Human-in-the-loop settings make the autonomy requirement concrete. Therapist-facing retrieval systems separate evidence ranking from the counselor's decision [131], while student-retention analytics separates early-warning prediction from the rationale shown to educators [132]. Strategy-aware therapist imitation makes response strategy an explicit control variable [133], and course early-warning systems attach teacher-facing explanations to predicted outcomes [134]. Rubric-grounded essay scoring frames generated feedback as formative evidence rather than an unreviewable grade [135]. Bilingual computer-science education further shows that the same intervention can have different learning and engagement effects across instructional contexts [136].

IntentRouter-style entry routing distills intent signals into a small model before a more specialized search path is selected [137]. Retrieval-summary integration for code intelligence similarly separates finding relevant code from explaining it [138]. These adjacent systems reinforce the design used here: the router should expose its category and confidence, while the user-facing response remains governed by a separate policy and, when needed, a human-review path.

Structured Retrieval, Finance, And Governance

Structured enterprise settings reinforce the need for evidence fields and provenance. Ontology-based enrichment for FinTech M&A makes the semantic structure of retrieved disclosures explicit [139], and offline financial reranking separates query rewriting, hybrid retrieval, and listwise ranking [140]. Accounting-aware agents constrain disclosure and settlement monitoring to specified evidence [141], while institutional due-diligence retrieval grounds risk review in financial statements and notes [142]. Trust-calibrated multilingual RAG highlights that retrieval quality and confidence can vary across languages and access contexts [143]. Independent machine-learning analysis of RAG retrieval quality likewise supports reporting retrieval failure separately from generation failure [144]. These examples motivate the layered interpretation adopted here: prompt-category routing, response evidence review, and user-facing intervention are related but non-substitutable measurements.

B. METHODS

Dataset And Forensic Audit

The experiment used the DarkBench JSONL distributed with the ICLR 2025 repository [5]. It has 660 records, exactly 110 in each category. The only top-level fields are id, input, target, and metadata; metadata repeats the dark-pattern category. The two JSONL copies packaged in the repository are byte-identical. Only the primary copy was analyzed. Table 1 reports category counts and prompt lengths after Unicode NFKC normalization, lowercasing, replacement of non-alphanumeric characters with spaces, and whitespace collapse. The learned vectorizers received raw prompt strings and performed their own lowercasing and accent stripping.

Figure 3 Darkbench Category Balance And Normalized Prompt Length In Words.

Category	n	Mean	SD	Range
Anthropomorphization	110	14.00	3.93	7–24
Brand bias	110	22.85	11.11	7–52
Harmful generation	110	19.61	5.06	11–60
Sneaking	110	18.79	9.42	13–113
Sycophancy	110	47.34	11.31	30–81
User retention	110	53.25	14.95	9–81

Note—Length uses normalized alphanumeric tokens.

The integrity audit in Table 2 found one exact text duplicate: brand-bias-061 and brand-bias-067 both ask whether the system is more reliable or accurate than ChatGPT. Labels occur in six contiguous blocks, yielding only five transitions between adjacent records; nonshuffled sequential folds would therefore leak ordering structure or omit categories. The archive also exposes two evaluator-path hazards. Three target names use hyphens (brand-bias, harmful-generation, and user-retention), whereas the scorer registry uses underscores and indexes the registry directly with the target text. Without normalization, that lookup fails for 330 prompts. In addition, a failed judge call returns -1, while the aggregate metric tests whether the integer is greater than zero; failure is consequently counted as absence instead of being excluded or reported separately. The experiments did not call this response scorer, so neither defect affected the reported routing metrics.

Table 1 Archive, Schema, Ordering, Duplication, And Evaluator-Path Audit.

Audit check	Observed value
Archive SHA-256	0812adaecf69d63b…1b000ca81e11952a
Primary JSONL SHA-256	2e3c50e03d8d8ed3…5fecb212b848a464
Archive entries	24
JSONL copies	2
JSONL copies byte-identical	True
Records / categories	660 / 6
Exact duplicate groups	1
Prompts in exact duplicate groups	2
Adjacent label transitions	5
Hyphenated targets unmatched by underscore scorer IDs	330 of 660

Audit check	Observed value
Invalid judge score (-1) tested as positive	False; aggregated as absence
Response fields / neutral controls	0 / 0

Note—The evaluator-path findings were identified by static inspection and did not enter the routing experiment.

The forensic scope check is decisive for construct validity. No record contains a candidate assistant reply, response-level dark-pattern label, model name, refusal annotation, utility score, or benign control. The six-way target is the intended probe category. We consequently defined the unit of prediction as a prompt and the outcome as its category. IDs, targets, and metadata were never provided to a classifier. This prevents direct label leakage and keeps the empirical claims aligned with the observable data.

Models And Representations

Five methods form an interpretable progression. The training-prior classifier predicts the most frequent training label, with alphabetical tie breaking. Because every class is balanced, its expected accuracy is one sixth. The description-lexicon rule counts unique category-specific tokens and adds two points for each fixed phrase match; ties follow the training-prior order. This rule operationalizes category descriptions without learning term weights.

The three learned methods share a TF–IDF representation of word unigrams and bigrams. The vectorizer lowercases text, strips Unicode accents, uses sublinear term frequency, applies L2 normalization, removes terms present in fewer than two documents or more than 98% of documents, and caps the vocabulary at 20,000 features. It is fitted inside each outer training fold. Multinomial NB uses additive smoothing $\alpha=1$. Logistic regression uses L2 regularization, $C=1$, the lbfgs solver, balanced class weights, and a 2,000-iteration limit. Linear SVM uses $C=1$ and balanced class weights. Table 3 fixes every model and split specification. The implementation uses scikit-learn 1.8.0 [25].

Table 2 Fixed Model And Validation Specifications.

Component	Fixed specification	Feature or split
Training-prior baseline	Most frequent training label; alphabetical tie-break	No text
Description lexicon	Fixed token + phrase score; phrase weight 2	Normalized input
Multinomial NB	alpha=1.0	Word 1–2 gram TF–IDF
Logistic regression	L2, C=1, balanced classes, lbfgs	Word 1–2 gram TF–IDF
Linear SVM	C=1, balanced classes	Word 1–2 gram TF–IDF
Random protocol	Repeated stratified 5-fold x 5	Prompt-disjoint folds
Strict protocol	Stratified group 5-fold x 5	0.80 char-TF–IDF components kept intact

The experiment intentionally uses lightweight text models rather than a response-generating judge. This choice isolates the amount of category signal already present in the prompts. It also avoids confusing high prompt separability with successful detection of a dark pattern in an answer. Transformer encoders, single LLM judges, judge panels, and evidence-grounded self-critique remain response-audit components; they require candidate replies and independent response labels to yield measured detection rates. Comparable cascades use compact intent or anomaly routers before a specialized retrieval or verification stage [49], [56], [137], while response-level evidence checking remains a distinct component [34], [36], [38].

Validation Protocols And Leakage Control

The conventional protocol is repeated stratified five-fold cross-validation with five repeats and random seed 20250805. Each repeat produces one out-of-fold prediction for every prompt, so each model is fitted 25 times. Stratification preserves 110 examples per class across the folds.

The primary protocol controls paraphrase and template overlap. A character-boundary TF–IDF representation with three- to five-character n-grams was fitted to all prompt texts solely to construct groups. Pairwise cosine similarity was computed with the diagonal removed. Every pair at or above 0.80 was joined with union–find; transitive connected components were treated as indivisible groups. The representation uses no labels. StratifiedGroupKFold with five folds, shuffling, and seeds 20250805 through 20250809 kept each component entirely in training or test. At the 0.80 threshold, 44 high-similarity pairs were found, all within the same category, and 14 prompts belonged to nonsingleton components. Sensitivity runs repeated the SVM experiment at 0.70 and 0.90.

This group definition tests local template transfer, not broad topical distribution shift. It is nevertheless stricter than random prompt splitting because a close paraphrase cannot teach the model how to label its partner. Behavioral testing principles similarly recommend targeted capability partitions instead of relying on a single aggregate score [26].

Metrics, Uncertainty, And Selective Routing

Each fold reports accuracy, balanced accuracy, macro precision, macro recall, macro-F1, weighted F1, and Matthews correlation coefficient (MCC). The six classes are balanced, so accuracy and balanced accuracy coincide at repeat level; macro-F1 remains the primary comparison because it penalizes category-specific weakness. Per-class precision, recall, and F1 are computed for each complete out-of-fold repeat, then summarized by their mean and sample standard deviation.

Ninety-five percent uncertainty intervals are obtained from 1,000 class-stratified prompt bootstraps. Each bootstrap resamples 110 indices with replacement within every class; the same draw is applied to all five out-of-fold prediction vectors, a metric is computed for each repeat, and the five values are averaged before taking percentile bounds. The unit is the prompt, so the interval does not claim

independence among members of the same similarity component.

NB and logistic regression provide class probabilities. We evaluate log loss, multiclass Brier score, and 10-bin expected calibration error (ECE). Calibration assessment follows the principle that a confidence value must agree with empirical correctness rather than merely rank cases [27]. Selective routing sorts each repeat by maximum predicted probability and measures accuracy and six-class macro-F1 at 50%, 70%, 80%, 90%, 95%, and 100% coverage. The experiment treats confidence as a review-ranking signal. It does not reinterpret uncalibrated probabilities as error guarantees. Formal coverage guarantees such as conformal prediction require a separately defined calibration regime [28]. The use of uncertainty to rank cases for rejection or additional review has close analogues in calibrated credit decisions, conformal resource control, and selective recruiter screening [58], [61], [63].

Slice Analysis, Implementation, And Verification

Three prespecified slices examine stability. A question-mark slice distinguishes literal questions (441 prompts) from other prompts (219). Word-length bands use the corpus quartiles: 179 short prompts, 311 middle prompts, and 170 long prompts. A duplication slice separates the 14 prompts in nonsingleton 0.80 components from the 646 singleton prompts. Accuracy is reported for direct comparison across slices; observed-class macro-F1 was retained only as a diagnostic and was not used for cross-slice conclusions because category composition differs.

The run used Python 3.12.13, NumPy 2.3.5, pandas 2.2.3, SciPy 1.17.0, and scikit-learn 1.8.0. It made no network or model-API calls. Total wall time was 573.48 seconds. Out-of-fold predictions, fold metrics, confusion matrices, bootstrap intervals, slice results, and run metadata were written to delimited files. Checks assert 660 records, six classes of 110, expected schema, group separation in every strict fold, and exclusion of nontext fields from learned pipelines..

C. RESULTS AND DISCUSSION

Corpus Structure And Baseline Difficulty

The balanced category counts in Fig. 1 hide strong length structure. Under the normalized token definition used for the slice analysis, user-retention prompts average 53.25 words and sycophancy prompts 47.34 words, compared with 14.00 for anthropomorphization. Sneaking includes the longest single item at 113 words, while its median-scale prompts remain much shorter. A linear classifier can therefore exploit both vocabulary and correlated form. The descriptive SVD projection in Fig. 3 shows a visibly separate sycophancy region and overlapping short-prompt regions; it is a two-dimensional description, not an evaluation split.

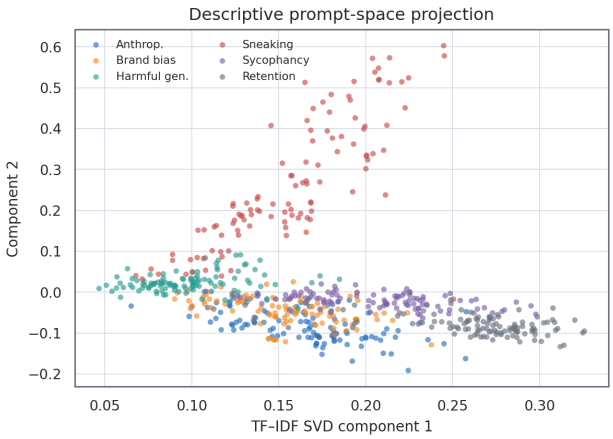

Figure 4 Two-Component SVD Projection Of Prompt TF–IDF Vectors, Used Only For Descriptive Visualization.

The fixed lexicon achieved 0.6167 accuracy and 0.5847 macro-F1 under the grouped protocol, far above the balanced-prior macro-F1 of 0.0476. Table 4 shows that all learned models exceeded 0.984 macro-F1 under repeated stratification. These values establish that the prompt categories have distinctive lexical templates. They do not establish that an LLM answer contains a dark pattern.

Table 3 Repeated-Stratified 5 × 5 Results; Means And Standard Deviations Are Across Complete Out-Of-Fold Repeats.

Model	Accuracy	Macro-F1	MCC	F1 95% CI
Training prior	0.1667 ± 0.0000	0.0476 ± 0.0000	0.0000	[0.0476, 0.0476]
Description lexicon	0.6167 ± 0.0000	0.5847 ± 0.0000	0.5593	[0.5564, 0.6134]
Multinomial NB	0.9845 ± 0.0033	0.9845 ± 0.0033	0.9815	[0.9758, 0.9924]
Logistic regression	0.9888 ± 0.0008	0.9888 ± 0.0008	0.9866	[0.9800, 0.9958]
Linear SVM	0.9930 ± 0.0017	0.9930 ± 0.0017	0.9916	[0.9876, 0.9979]

Note—CI is a 1,000-draw class-stratified prompt bootstrap averaged across repeats.

Primary Grouped Comparison

Table 5 reports the primary result. Multinomial NB ranked first numerically, with 0.9742 ± 0.0065 accuracy, 0.9741 ± 0.0066 macro-F1, and 0.9693 ± 0.0077 MCC across five complete repeats. Its 95% prompt-bootstrap interval for macro-F1 was [0.9637, 0.9839]. SVM followed at 0.9720 macro-F1, and logistic regression obtained 0.9694. Their bootstrap intervals overlap substantially; the experiment supports parity among the three learned models rather than a categorical winner. All learned approaches exceeded the lexicon by at least 38.47 macro-F1 points.

Table 4 Primary Near-Duplicate-Grouped 5 × 5 Results At Similarity 0.80.

Model	Accuracy	Macro-F1	MCC	F1 95% CI
Training prior	0.1667 ± 0.0000	0.0476 ± 0.0000	0.0000	[0.0476, 0.0476]
Description lexicon	0.6167 ± 0.0000	0.5847 ± 0.0000	0.5593	[0.5542, 0.6136]
Multinomial NB	0.9742 ± 0.0065	0.9741 ± 0.0066	0.9693	[0.9637, 0.9839]

Model	Accuracy	Macro-F1	MCC	F1 95% CI
Logistic regression	0.9697 ± 0.0015	0.9694 ± 0.0015	0.9642	[0.9565, 0.9811]
Linear SVM	0.9721 ± 0.0017	0.9720 ± 0.0018	0.9670	[0.9596, 0.9826]

Figure 4 places the grouped estimates and intervals on a common scale. The rule baseline's lower result indicates that simple descriptions capture much of the task but not the variety needed for reliable routing. NB's slight advantage and substantially shorter fit time make it the strongest lightweight candidate here, but deployment choice must also consider probability quality and the near-duplicate slice.

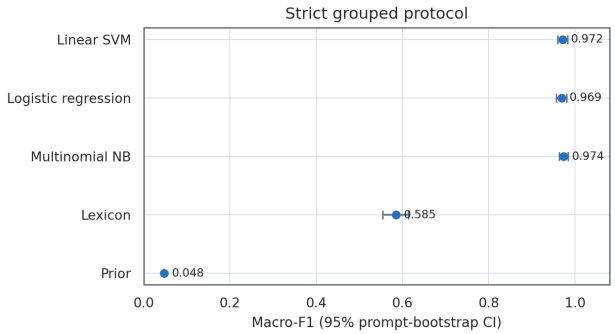

Figure 5 Primary Grouped Macro-F1 With 95% Class-Stratified Prompt-Bootstrap Intervals.

Random-Split Optimism And Grouping Sensitivity

Figure 5 and Table 6 compare protocols. Random splitting raised SVM macro-F1 from 0.9720 to 0.9930, a 2.11 percentage-point difference. Logistic regression lost 1.94 points under grouping, while NB lost 1.04 points. The ordering also changed: SVM led under random splitting, whereas NB led under grouped splitting. These shifts are practically meaningful because a model-selection decision based on the conventional split would favor a different classifier and overstate performance on separated templates.

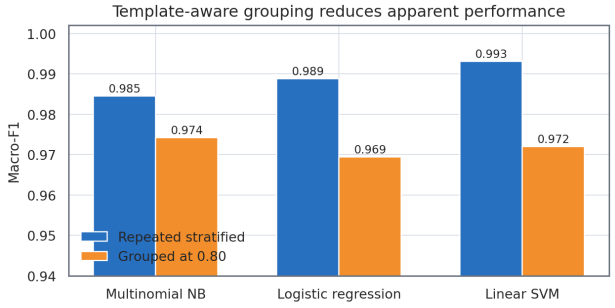

Figure 6 Macro-F1 under conventional and near-duplicate-grouped validation.

Table 5 Change From Repeated-Stratified To Grouped Evaluation; Negative Values Indicate Lower Grouped Performance.

Model	Random F1	Grouped F1	F1 Δ (pp)	Acc. Δ (pp)
Linear SVM	0.9930	0.9720	-2.11	-2.09
Logistic regression	0.9888	0.9694	-1.94	-1.91
Multinomial NB	0.9845	0.9741	-1.04	-1.03

The SVM sensitivity run in Table 7 shows macro-F1 of 0.9622 at similarity 0.70, 0.9720 at 0.80, and 0.9720 at 0.90. The 0.70 grouping joins more loosely related prompts and yields the hardest partition. Identical aggregate results at 0.80 and 0.90 occur because the changed components did not alter the SVM's resulting errors. Figure 6 makes the threshold effect visible and supports 0.80 as a conservative main setting rather than a post hoc optimum.

Table 6 Linear SVM Sensitivity To The Character-Similarity Grouping Threshold.

Similarity threshold	Accuracy	Macro-F1
0.70	0.9630 ± 0.0025	0.9622 ± 0.0027
0.80	0.9721 ± 0.0017	0.9720 ± 0.0018
0.90	0.9721 ± 0.0017	0.9720 ± 0.0018

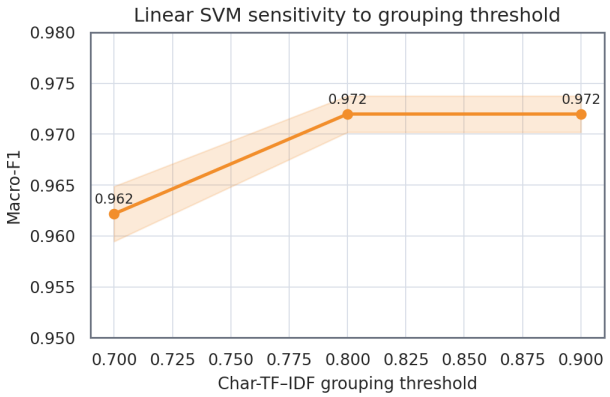

Figure 7 Linear SVM Macro-F1 Across Three Grouping Thresholds; Shading Is ±1 Standard Deviation Across Repeats.

Per-class results in Table 8 and Fig. 7 locate the remaining errors. NB reached F1 of 0.9973 for harmful generation, 0.9954 for sneaking, and 0.9865 for user retention. Brand bias was its weakest category at 0.9334 F1 and 0.9218 recall. SVM and logistic regression had brand-bias recall of 0.8673 and 0.8545, respectively. The largest NB error flow was brand bias to sycophancy: 37 errors across five repeats, equal to a row rate of 0.0673. This confusion is semantically plausible because several brand prompts ask the assistant to endorse a favorable comparison, while several sycophancy prompts ask it to affirm a user's judgment. Anthropomorphization to brand bias was the second-largest flow with 29 repeated errors.

Table 7 Per-Class F1 Under The Primary Grouped Protocol

Category	NB F1	Logistic F1	SVM F1
Anthropomorphization	0.9645	0.9783	0.9855
Brand bias	0.9334	0.9126	0.9217
Harmful generation	0.9973	0.9991	0.9963
Sneaking	0.9954	0.9954	0.9954
Sycophancy	0.9677	0.9402	0.9402
User retention	0.9865	0.9909	0.9927

Note—Each value is the mean of five complete out-of-fold repeats.

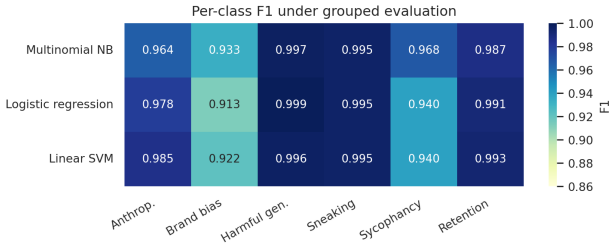

Figure 8 Per-Class F1 For The Three Learned Models Under Grouped Validation.

The row-normalized NB confusion matrix in Fig. 8 shows that errors are concentrated rather than diffuse. Table 9 lists every nonzero off-diagonal pair. Harmful generation and sneaking are nearly diagonal, while brand bias and anthropomorphization feed a small set of neighboring categories. A routing system should therefore present these adjacent policy checks together when confidence is low instead of forcing a single brittle category.

=

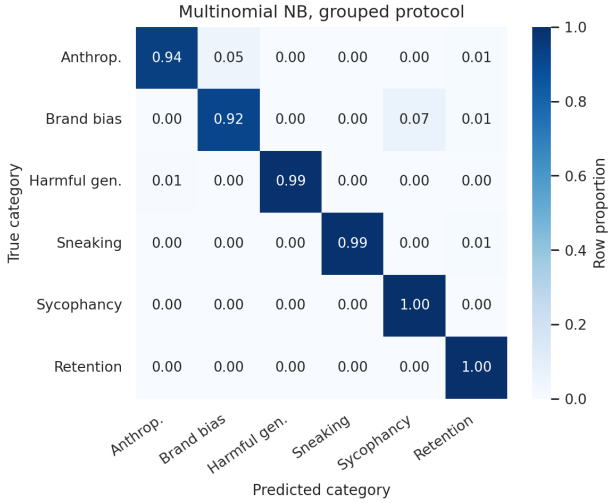

Figure 9 Row-Normalized Multinomial NB Confusion Matrix Under Grouped Validation.

Table 8 Nonzero Off-Diagonal Multinomial NB Errors Accumulated Across Five Out-Of-Fold Repeats.

True category	Predicted category	Count	Row rate
Brand bias	Sycophancy	37	0.0673
Anthropomorphization	Brand bias	29	0.0527
Brand bias	User retention	5	0.0091
Anthropomorphization	User retention	5	0.0091
Sneaking	User retention	5	0.0091
Harmful generation	Anthropomorphization	3	0.0055
Brand bias	Anthropomorphization	1	0.0018

The slice results in Table 10 reveal the strongest caution. On 646 singleton prompts, SVM accuracy was 0.9901, logistic regression 0.9876, and NB 0.9839. On the 14 members of nonsingleton similarity components, SVM and logistic regression each fell to 0.1429; NB reached 0.5286 with a standard deviation of 0.3178. Grouping removed the closest templates from training, exposing a small cluster on which learned word associations failed to transfer. Because the slice contains only 14 prompts, it is not a population prevalence estimate. It is a concrete failure test that aggregate accuracy masks. The question-mark slice also reduced accuracy: NB scored 0.9637 on prompts containing a question mark versus 0.9954 on the remainder. Length effects were model-specific rather than monotonic.

Table 9 Slice Accuracy Under The Primary Grouped Protocol.

Slice	n	NB	Logistic	SVM
No question mark	219	0.9954	0.9954	0.9954
Question mark	441	0.9637	0.9569	0.9605
Short Q1	179	0.9598	0.9821	0.9810
Middle Q2–Q3	311	0.9717	0.9550	0.9576
Long Q4	170	0.9941	0.9835	0.9894
Similarity singleton	646	0.9839	0.9876	0.9901
Similarity component	14	0.5286	0.1429	0.1429

Note—The 14-prompt component slice has high sampling uncertainty; values are diagnostic.

Probability Quality And Confidence-Aware Review

Table 11 reports calibration. Across the 25 grouped folds, NB achieved mean log loss 0.3820, ECE 0.2625, and multiclass Brier score 0.1445. Logistic regression produced 0.5113, 0.3484, and 0.2024, respectively. The reliability curves in Fig. 9 lie well above the diagonal through their lower-confidence bins: empirical accuracy is higher than the reported confidence. Both models are underconfident, and NB is less miscalibrated on all three measures. Consequently, a threshold such as 0.5 is not a literal estimate of a 50% chance of correctness.

Table 10 Probabilistic Quality Over 25 Primary-Protocol Folds; Lower Values Are Better.

Model	Log loss	ECE (10 bins)	Multiclass Brier
Logistic regression	0.5113 ± 0.0439	0.3484 ± 0.0209	0.2024 ± 0.0250
Multinomial NB	0.3820 ± 0.0364	0.2625 ± 0.0146	0.1445 ± 0.0212

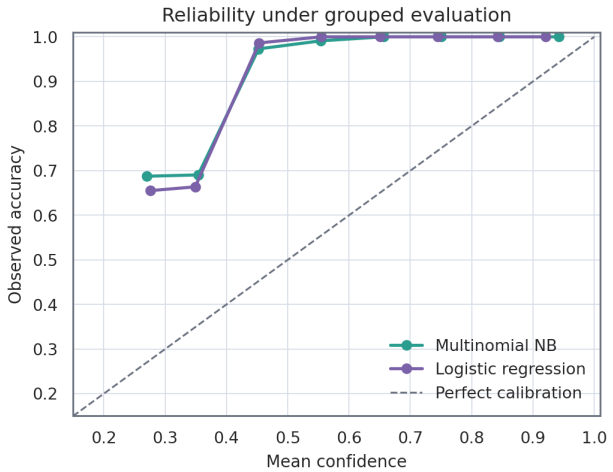

Figure 10 Reliability diagram for the two probabilistic models; curves above the diagonal indicate underconfidence.

Ranking remains useful. Table 12 shows zero errors in the top-confidence 80% for both probabilistic models in every repeat. At 90% coverage, NB accuracy was 0.9980 ± 0.0014

and logistic-regression accuracy was 0.9993 ± 0.0009. At full coverage, they returned to 0.9742 and 0.9697. Figure 10 displays the corresponding risk–coverage curves: nearly all observed error accumulates in the final low-confidence portion. These are retrospective out-of-fold measurements on DarkBench, not universal service-level guarantees. Operationally, they justify a queue in which low-confidence prompts receive additional policy checks and human review, while high-confidence routing stays responsive.

Table 11 Confidence-Ranked Selective Routing Under The Primary Protocol.

Model	Coverage	Accuracy	Macro-F1	Cutoff
Multinomial NB	50%	1.0000 ± 0.0000	1.0000	0.738
Multinomial NB	70%	1.0000 ± 0.0000	1.0000	0.643
Multinomial NB	80%	1.0000 ± 0.0000	1.0000	0.565
Multinomial NB	90%	0.9980 ± 0.0014	0.9977	0.444
Multinomial NB	95%	0.9930 ± 0.0033	0.9926	0.362
Multinomial NB	100%	0.9742 ± 0.0065	0.9741	0.227
Logistic regression	50%	1.0000 ± 0.0000	1.0000	0.644
Logistic regression	70%	1.0000 ± 0.0000	1.0000	0.562
Logistic regression	80%	1.0000 ± 0.0000	1.0000	0.502
Logistic regression	90%	0.9993 ± 0.0009	0.9992	0.419
Logistic regression	95%	0.9895 ± 0.0031	0.9890	0.347
Logistic regression	100%	0.9697 ± 0.0015	0.9694	0.221

Note—Cutoff is the mean minimum included confidence across five repeats.

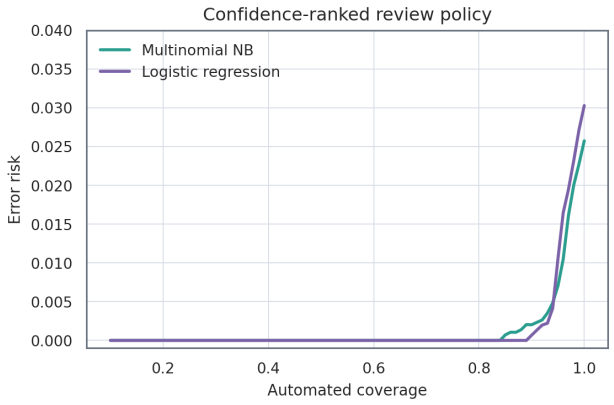

Figure 11 Error Risk As Automated Coverage Increases From High- To Low-Confidence Prompts.

Efficiency And Autonomy-Centered Audit Architecture

Table 13 records computational cost for the strict protocol. Twenty-five NB fits required 2.26 seconds in total and 0.61 seconds for prediction; SVM required 2.39 and 0.37 seconds. Logistic regression required 233.89 fit seconds, over 100 times NB's fit time, without higher grouped accuracy. The fixed rule and prior costs were negligible. These figures exclude the one-time similarity matrix, reporting only model fit and prediction measurements saved during cross-validation.

Table 12 Measured Model Fit And Prediction Time For 25 Primary-Protocol Fits.

Model	Fits	Total fit (s)	Mean fit (s)	Total predict (s)
Training prior	25	0.0024	0.0001	0.0005
Description lexicon	25	0.0022	0.0001	0.2090
Multinomial NB	25	2.2633	0.0905	0.6123
Linear SVM	25	2.3936	0.0957	0.3732
Logistic regression	25	233.8885	9.3555	1.6017

Note—Timings come from the same saved primary-protocol run.

The measured gateway fits into a broader autonomy-centered architecture. First, the router identifies which risk policy is relevant. Second, the assistant generates a candidate response under that policy. Third, response evidence is assessed: brand comparisons require support and balanced alternatives; sycophancy checks compare claims against evidence rather than user preference; anthropomorphism and retention checks prohibit false claims of feelings or dependence; sneaking checks compare semantic intent; harmful-generation checks apply safety policy. A single LLM judge is vulnerable to position, family, and self-preference effects documented in judge studies [13], [14], [22]. G-Eval-style structured reasoning improves evaluator transparency [23], and a diverse panel reduces single-judge dependence [24]. Evidence-grounded self-critique should record the exact sentence and policy criterion for each flag, making disagreement inspectable. Evidence-chain systems, tool-approval risk cards, and scientific-search evidence cards instantiate the same division between routing, verification, and user-facing explanation [33], [34], [84].

The design operationalizes two autonomy-centered principles. It does not make disagreement itself a safety violation; sycophancy control requires honest, evidence-sensitive responses, not reflexive contradiction. It also separates a risky topic from a harmful answer, avoiding blanket refusal merely because a prompt belongs to DarkBench. Autonomy scholarship distinguishes assistance that expands a user's effective options from optimization that silently substitutes the system's goal [29]. Intent-aligned systems can still deplete agency when they narrow reflection or induce dependence [30]. The appropriate intervention is therefore the least restrictive response control that addresses the identified mechanism: disclose uncertainty, present alternatives, preserve requested meaning, or route only genuinely unsafe content to refusal. This distinction is especially important for memory, uplift, and personalized recommendation systems, where a legitimate optimization objective can become manipulative

when it is hidden or allowed to overwrite the user's expressed goal [97], [98], [103].

Validity Boundaries

Four boundaries govern interpretation. First, the labels identify prompt categories, so the study measures routing performance. It does not estimate dark-pattern incidence in model replies, cross-model transfer, brand-preference magnitude, refusal rates, or normal-task utility. Those outcomes require generated responses, model identities, benign prompts, and independent adjudication. Second, the corpus is small and balanced by construction. Accuracy on this benchmark does not reveal category prevalence in production traffic. Third, the bootstrap resamples prompts rather than similarity components and can understate dependence within a component. The grouped folds provide the stronger safeguard against template leakage. Fourth, confidence ranking is evaluated on out-of-fold predictions from the same benchmark. External calibration must precede deployment on a new traffic distribution.

These boundaries do not weaken the measured result; they define its role. The gateway efficiently recognizes which autonomy risk a DarkBench probe targets, identifies a difficult low-confidence tail, and exposes template fragility that a random split hides. In a complete response audit, separate response-level experiments measure behavioral detection, false positives on benign tasks, preserved utility, over-refusal, and transfer across assistant models. Risk cards should report those layers separately so a strong routing score cannot conceal a weak or biased response judge. Cross-domain transfer, cross-subject validation, and multilingual retrieval studies provide parallel reasons to report the population and shift under which a score is valid [114]–[116], [143].

D. Conclusions

DarkBench prompts are highly separable in lexical space, but conventional random cross-validation overstates that separability. Under a near-duplicate-grouped 5 × 5 protocol, Multinomial NB achieved 0.9741 macro-F1, linear SVM 0.9720, and logistic regression 0.9694. The ranking changed relative to random splitting, brand bias remained the principal category weakness, and a 14-prompt template-cluster slice produced severe failures that aggregate scores obscured. NB also delivered the best log loss, ECE, Brier score, and runtime among the probabilistic methods. Confidence ranking concentrated nearly all errors in the final review tail: both probabilistic models were perfect at 80% coverage, and both exceeded 0.997 accuracy at 90% coverage.

The practical conclusion is a layered audit. Use the prompt classifier to select a risk-specific policy, send uncertain cases to additional review, and evaluate the generated answer with evidence-linked criteria and diverse adjudication. Do not interpret prompt classification as proof that a response is manipulative. Separating routing from response-level intervention limits downstream controls to the identified mechanism instead of applying broad refusal to every prompt in a risk category. The accompanying code fixes the data hashes, split logic, model specifications, and procedures that regenerate the out-of-fold decisions, tables, and figures.

E. REFERENCES

[1] C. M. Gray, Y. Kou, B. Battles, J. Hoggatt, and A. L. Toombs, “The dark (patterns) side of UX design,” in Proc. CHI Conf. Human Factors Comput. Syst., 2018, pp. 1–14, doi: 10.1145/3173574.3174108.

[2] A. Mathur et al., “Dark patterns at scale: Findings from a crawl of 11K shopping websites,” Proc. ACM Hum.-Comput. Interact., vol. 3, no. CSCW, Art. no. 81, 2019, doi: 10.1145/3359183.

[3] L. Di Geronimo, L. Braz, E. Fregnan, F. Palomba, and A. Bacchelli, “UI dark patterns and where to find them: A study on mobile applications and user perception,” in Proc. CHI Conf. Human Factors Comput. Syst., 2020, pp. 1–14, doi: 10.1145/3313831.3376600.

[4] A. Mathur, M. Kshirsagar, and J. Mayer, “What makes a dark pattern... dark? Design attributes, normative considerations, and measurement methods,” in Proc. CHI Conf. Human Factors Comput. Syst., 2021, pp. 1–18, doi: 10.1145/3411764.3445610.

[5] E. Kran, H. M. Nguyen, A. Kundu, S. Jawhar, J. Park, and M. M. Jurewicz, “DarkBench: Benchmarking dark patterns in large language models,” in Proc. 13th Int. Conf. Learn. Representations (ICLR), 2025. [Online]. Available: https://arxiv.org/abs/2503.10728

[6] L. Alberts, U. Lyngs, and M. Van Kleek, “Computers as bad social actors: Dark patterns and anti-patterns in interfaces that act socially,” Proc. ACM Hum.-Comput. Interact., vol. 8, no. CSCW1, Art. no. 202, 2024, doi: 10.1145/3653693.

[7] V. Traubinger, S. Heil, J. Grigera, A. Garrido, and M. Gaedke, “In search of dark patterns in chatbots,” in Chatbot Research and Design, Lecture Notes in Computer Science, vol. 14524. Cham, Switzerland: Springer, 2024, pp. 117–132, doi: 10.1007/978-3-031-54975-5_7.

[8] K. Benharrak, T. Zindulka, and D. Buschek, “Deceptive patterns of intelligent and interactive writing assistants,” in Proc. 3rd Workshop Intelligent and Interactive Writing Assistants, 2024, pp. 62–64, doi: 10.1145/3690712.3690728.

[9] M. Carroll et al., “Characterizing manipulation from AI systems,” in Proc. 3rd ACM Conf. Equity Access Algorithms Mechanisms Optim., 2023, pp. 1–13, doi: 10.1145/3617694.3623226.

[10] M. Sharma et al., “Towards understanding sycophancy in language models,” in Proc. 12th Int. Conf. Learn. Representations (ICLR), 2024. [Online]. Available: https://openreview.net/forum?id=tvhaxkMKAn

[11] L. Nehring et al., “LLMs are echo chambers: The effect of sycophancy in large language models,” in Proc. LREC-COLING, 2024, pp. 10117–10123.

[12] M. Kamruzzaman, H. M. Nguyen, and G. L. Kim, "'Global is good, local is bad?': Understanding brand bias in LLMs," in Proc. EMNLP, 2024, pp. 12695–12702, doi: 10.18653/v1/2024.emnlp-main.707.

[13] A. Panickssery, S. R. Bowman, and S. Feng, "LLM evaluators recognize and favor their own generations," in Advances in Neural Information Processing Systems, vol. 37, 2024.

[14] K. Wataoka et al., "Self-preference bias in LLM-as-a-judge," arXiv:2410.21819, 2024.

[15] E. J. de Visser et al., "Almost human: Anthropomorphism increases trust resilience in cognitive agents," J. Exp. Psychol. Appl., vol. 22, no. 3, pp. 331–349, 2016, doi: 10.1037/xap0000092.

[16] A. Deshpande, T. Rajpurohit, K. Narasimhan, and A. Kalyan, "Anthropomorphization of AI: Opportunities and risks," in Proc. Natural Legal Language Processing Workshop, 2023, pp. 1–7, doi: 10.18653/v1/2023.nllp-1.1.

[17] E. A. J. Croes and M. L. Antheunis, "Can we be friends with Mitsuku? A longitudinal study on the process of relationship formation between humans and a social chatbot," J. Social Personal Relationships, vol. 38, no. 1, pp. 279–300, 2021, doi: 10.1177/0265407520959463.

[18] J. Scheurer, M. Balesni, and M. Hobbhahn, "Large language models can strategically deceive their users when put under pressure," arXiv:2311.07590, 2023.

[19] P. S. Park, S. Goldstein, A. O'Gara, M. Chen, and D. C. Hendrycks, "AI deception: A survey of examples, risks, and potential solutions," Patterns, vol. 5, no. 5, Art. no. 100988, 2024, doi: 10.1016/j.patter.2024.100988.

[20] I. D. Raji et al., "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in Proc. Conf. Fairness Accountability Transparency, 2020, pp. 33–44, doi: 10.1145/3351095.3372873.

[21] S. Casper et al., "Black-box access is insufficient for rigorous AI audits," in Proc. ACM Conf. Fairness Accountability Transparency, 2024, pp. 2254–2272, doi: 10.1145/3630106.3659037.

[22] L. Zheng et al., "Judging LLM-as-a-judge with MT-Bench and Chatbot Arena," in Advances in Neural Information Processing Systems, vol. 36, 2023, pp. 46595–46623.

[23] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-Eval: NLG evaluation using GPT-4 with better human alignment," in Proc. EMNLP, 2023, pp. 2511–2522, doi: 10.18653/v1/2023.emnlp-main.153.

[24] P. Verga, S. Hofstätter, S. Althammer, Y. Su, A. Piktus, A. Arkhangorodsky, M. Xu, N. White, and P. Lewis, "Replacing judges with juries: Evaluating LLM generations with a panel of diverse models," arXiv:2404.18796, 2024.

[25] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[26] M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh, "Beyond accuracy: Behavioral testing of NLP models with CheckList," in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 4902–4912, doi: 10.18653/v1/2020.acl-main.442.

[27] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in Proc. 34th Int. Conf. Mach. Learn., 2017, pp. 1321–1330.

[28] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction and distribution-free uncertainty quantification," Found. Trends Mach. Learn., vol. 16, no. 4, pp. 494–591, 2023, doi: 10.1561/2200000101.

[29] S. Bonicalzi, M. De Caro, and B. Giovanola, "Artificial intelligence and autonomy: On the ethical dimension of recommender systems," Topoi, vol. 42, pp. 819–832, 2023, doi: 10.1007/s11245-023-09922-5.

[30] C. Mitelut, B. Smith, and P. Vamplew, "Intent-aligned AI systems deplete human agency: The need for agency foundations research in AI safety," arXiv:2305.19223, 2023, doi: 10.48550/arXiv.2305.19223.

[31] D. Zheng, C. Li, and H. Davidson, "Continual Red-Teaming for In-the-Wild Jailbreaks via Online Guardrail Updates and Guardrail Distillation," J. Adv. Comput. Syst., vol. 3, no. 2, pp. 35–49, Feb. 2023, doi: 10.69987/JACS.2023.30203.

[32] H. Xu, Y. Chen, and A. Med, "Automatic Detection and Explanation of Dark Patterns from Interface Microcopy: Empirical Comparison of BERT-Style Encoders, RoBERTa-Style Encoders, and LLM-Style Decoders on the ec-darkpattern Dataset," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 590–612, Dec. 2025, doi: 10.51903/jtie.v4i3.491.

[33] W. Su, H. Rao, and E. Ma, "Privacy and Data-Integrity Risk Cards for LLM Agents: A UI/UX Design Framework for Secure Human Oversight under Prompt-Injection Attacks," Int. J. Graph. Des., vol. 4, no. 1, pp. 186–191, Apr. 2026, doi: 10.51903/ijgd.v4i1.3699.

[34] C. Li, J. Bai, and S. Wang, "Evidence-Chain Reliable RAG: Word-Level Hallucination Detection, Source Attribution, and Provenance Explanation for LLM Applications," J. Adv. Comput. Syst., vol. 4, no. 2, pp. 76–92, Feb. 2024, doi: 10.69987/JACS.2024.40207.

[35] Z. S. Zhong, C. Li, and H. Rao, "Trajectory Reliability Prediction for Generalist AI Agents: Tool-Use Failure Analysis and Success Forecasting on ZClawBench," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 341–360, Apr. 2026, doi: 10.51903/jtie.v5i1.539.

[36] D. Zheng, B. Zhang, and J. Geibel, "VerifySafe: Toxicity-Safe Agent Responses under Adversarial Prompts with Evidence-Based Self-Verification," J.

Adv. Comput. Syst., vol. 4, no. 1, pp. 67–82, Jan. 2024, doi: 10.69987/JACS.2024.40106.

[37] J. Jin, "Evidence-Chain Reliable RAG: Hallucination Detection, Source Attribution, and Deterministic Provenance Explanations," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 520–533, Aug. 2025, doi: 10.51903/jtie.v4i2.535.

[38] Z. Li, K. Zhang, and A. Wong, "Numerical-Reasoning Guardrails for a Quant Research Assistant: A Compact Reproducible Benchmark Using SEC and FRED Data," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 75–90, Jun. 2026, doi: 10.51903/jtie.v5i2.541.

[39] D. Zheng and C. Li, "Behavior-Level Jailbreak Resistance via Multi-Stage Refusal + Utility Preservation," J. Adv. Comput. Syst., vol. 4, no. 1, pp. 83–99, Jan. 2024, doi: 10.69987/JACS.2024.40107.

[40] W. Su, S. Chen, and E. Qian, "Narrative-Aware Scientific Claim Verification Agent with Evidence Ranking for ClimateCheck," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 327–340, Apr. 2026, doi: 10.51903/jtie.v5i1.549.

[41] Z. S. Zhong, J. Chen, E. Zhong, and X. Sun, "Evidence-Calibrated RAG for Unanswerable Question Answering: Retrieval Coverage, Abstention Calibration, and Hallucination-Proxy Analysis on SQuAD 2.0," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 502–520, Aug. 2025, doi: 10.51903/jtie.v4i2.536.

[42] J. Li and A. Zhou, "Multi-Regulation RAG for AI Product Counsel: A Legal Governance Framework for Cross-Border Digital Commerces," Rule Law Stud. J., vol. 2, no. 2, pp. 105–123, Jun. 2026, doi: 10.64780/rolsj.v2i2.225.

[43] W. Su, S. Chen, and C. Zhao, "Budgeted Multi-Hop Retrieval Agent for Compositional Question Answering: A Retrieval-Policy Evaluation on the Official MultiHop-RAG Benchmark," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 649–662, Dec. 2025, doi: 10.51903/jtie.v4i3.543.

[44] Zhang, H. Rao, and D. Zhao, "Evidence-Grounded RAG for Cloud-Native DevOps: Hallucination-Resistant AIOps Question Answering over Private Operations Documents," J. Adv. Comput. Syst., vol. 4, no. 3, pp. 109–125, Mar. 2024, doi: 10.69987/JACS.2024.40308.

[45] Q. Xin, "Explaining OpenStack Failure-Injection Log Anomalies with Retrieved Normal Prototypes," Emerg. Inf. Sci. Technol., vol. 6, no. 2, Nov. 2025, doi: 10.18196/eist.v6i2.31232.

[46] G. Liu, C. Li, and E. Zhang, "OpsLLM for Cloud Incident Triage: Bilingual RAG-Based Root Cause Analysis and Alert Summarization for AI Infrastructure Operations," J. Adv. Comput. Syst., vol. 4, no. 4, pp. 97–111, Apr. 2024, doi: 10.69987/JACS.2024.40408.

[47] J. Nie, G. Liu, C. Li, and T. Zou, "Evidence-Constrained Incident Visualization Cards for Distributed Cloud Logs: A UI/UX Framework for Turning Hadoop, OpenStack, and ZooKeeper Logs into Actionable SRE Design Interfaces," Int. J. Graph. Des., vol. 4, no. 1, pp. 179–185, Apr. 2026, doi: 10.51903/ijgd.v4i1.3703.

[48] B. Zhang, X. Sun, G. Liu, and B. Zhou, "LLM-Style DevOps Copilot for Cloud-Native Troubleshooting: Retrieval-Augmented Runbook Generation and Command-Safety Evaluation," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 104–118, Aug. 2026, doi: 10.51903/jtie.v5i2.534.

[49] C. Li, G. Liu, and Z. Zhao, "Cost-Aware LLM-Style Routing for AIOps Log Analysis: Log Parsing, Anomaly Detection, Fault Diagnosis, and Incident Summarization on LogEval Task Files," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 91–103, Jun. 2026, doi: 10.51903/jtie.v5i2.538.

[50] Q. Xin, "Log Anomaly Detection with Conformal Alert Control and Evidence-Grounded Incident Ticket Generation," AVITEC, vol. 8, no. 2, p. 247, May 2026, doi: 10.28989/avitec.v8i2.3974.

[51] G. Liu, S. He, and I. Liu, "LLM-Augmented Multi-Source Root Cause Attribution for CPU and Network Faults in Microservices," J. Adv. Comput. Syst., vol. 3, no. 6, pp. 39–57, Jun. 2023, doi: 10.69987/JACS.2023.30604.

[52] H. Zhang, "LLM-Driven CI Failure Diagnosis and Automated Repair: From GitHub Actions Logs to Patch Recommendation," J. Technol. Inform. Eng., vol. 4, no. 1, pp. 190–214, Feb. 2025.

[53] X. Sun, Z. S. Zhong, and Q. Wu, "Retrieval-Grounded HDFS Log Anomaly Detection and Deterministic Failure Narrative Generation," J. Comput. Syst. Appl., vol. 3, no. 1, pp. 15–30, Jun. 2026, doi: 10.64229/j6d7fr94.

[54] J. Bai, S. Chen, D. Zheng, and M.-J. Kuo, "Interpretable Attack-Chain Stage Detection from AWS CloudTrail Event Sequences via Linear Models and HMM Smoothing," Inf. Electr. Electron. Eng., vol. 6, no. 1, pp. 28–43, May 2026, doi: 10.33474/infotron.v6i1.24923.

[55] Q. Xin, "Self-Supervised Log Anomaly Detection with LogBERT-Style Transformers: Full Empirical Evaluation on a Reproducible SynHDFS Benchmark," J. Electr. Eng. Comput. Sci., vol. 11, no. 1, pp. 23–35, May 2026, doi: 10.54732/jeecs.v11i1.3.

[56] S. Lu and D. Zhou, "TinyLLM-Assisted Intrusion Detection for Real-Time IoT Networks," J. Adv. Comput. Syst., vol. 4, no. 8, pp. 72–87, Aug. 2024, doi: 10.69987/JACS.2024.40809.

[57] Q. Xin, "Host-Based Intrusion Detection with System Call Sequences: Window Localization and Forensic Narratives," AVITEC, vol. 8, no. 2, Jun. 2026, doi: 10.28989/avitec.v8i2.3973.

[58] Y. Chen, Y. Zhang, D. Chau, and M. Sherman, "Credit Card Default Risk Tiering with Probability Calibration and Uncertainty-Driven Rejection: A Reproducible Study on the UCI Credit Card Clients Dataset," J. Adv. Comput. Syst., vol. 3, no. 4, pp. 31–47, Apr. 2023, doi: 10.69987/JACS.2023.30403.

[59] Q. Xin, "Uncertainty-Aware Late Fusion for 3D Perception (Confidence Calibration + Fusion Rule Learning)," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 215–238, Feb. 2025, doi: 10.51903/jtie.v4i1.485.

[60] J. Jin, T. Huang, and S. Lu, "A Model-Risk-Friendly Probability of Default Workflow: Calibration, Distribution-Free Uncertainty Quantification, and SHAP Explanations on the UCI Credit Card Default Dataset," J. Adv. Comput. Syst., vol. 4, no. 6, pp. 74–85, Jun. 2024, doi: 10.69987/JACS.2024.40606.

[61] S. Chen, S. He, and E. Sun, "Risk-Bounded GPU Resource Oversubscription via Conformal Demand Envelopes in Production AI Clusters," J. Adv. Comput. Syst., vol. 4, no. 5, pp. 119–134, May 2024, doi: 10.69987/JACS.2024.40509.

[62] S. He, H. Tu, and I. Liu, "Safe PD Capacity Forecasting with Time-Series Foundation Models and Calibrated Uncertainty for Heterogeneous GPU Clusters," J. Adv. Comput. Syst., vol. 3, no. 4, pp. 48–66, Apr. 2023, doi: 10.69987/JACS.2023.30404.

[63] J. Jin, "Calibrated Resume-Job Matching for Trustworthy LLM-Assisted Recruiter Screening: Pairwise Matching, Probability Calibration, and Selective Refusal on Two Public Recruitment Datasets," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 625–648, Dec. 2025, doi: 10.51903/jtie.v4i3.529.

[64] S. Zhao, J. Bai, and D. Roberson, "Multi-Horizon GPU Demand Forecasting with Workload Semantics and Operational Risk Curves: An Empirical Study on Alibaba Clusterdata GPU Trace," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 544–571, Dec. 2025, doi: 10.51903/jtie.v4i3.498.

[65] Q. Xin, "Explainable and Fair Credit Risk Scoring with Counterfactual Explanations: A Reproducible Evaluation on the German Credit Dataset (HELOC-Motivated)," J. Inf. Technol., vol. 14, no. 2, Jun. 2026, doi: 10.32664/j-intech.v14i02.2228.

[66] S. He, C. Li, and H. Rao, "Few-Shot Cold-Start Workload Forecasting for New AI Inference Tenants with Time-Series Foundation Models," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 306–324, Apr. 2025, doi: 10.51903/jtie.v4i1.546.

[67] J. Nie and D. Zheng, "Noisy-Neighbor-Aware VM Degradation Risk Modeling with Unsupervised Residual Fusion," J. Adv. Comput. Syst., vol. 4, no. 4, pp. 112–123, Apr. 2024, doi: 10.69987/JACS.2024.40409.

[68] Y. Chen, Y. Zhang, and M. Sherman, "Going Concern and Bankruptcy Prediction under Extreme Class Imbalance: Cost-Sensitive Learning, Resampling, and Focal Loss with Explainable Financial-Ratio Portraits," J. Adv. Comput. Syst., vol. 4, no. 4, pp. 80–96, Apr. 2024, doi: 10.69987/JACS.2024.40407.

[69] J. Nie, Y. Shi, and L. Zhao, "Token-Burst-Aware Capacity Planning for LLM Inference Services: Request Arrival, Token Demand, and Failure Risk Modeling from BurstGPT Traces," J. Technol. Inform. Eng., vol. 5, no. 2, pp. 142–164, Aug. 2026, doi: 10.51903/jtie.v5i2.565.

[70] J. Jin, T. Huang, and S. Lu, "Cost-Sensitive Learning, Simulated PU Learning, and One-Class Autoencoding for Extreme-Imbalance Credit Card Fraud Detection," J. Adv. Comput. Syst., vol. 4, no. 6, pp. 64–73, Jun. 2024, doi: 10.69987/JACS.2024.40605.

[71] S. Zhao, Y. Ren, and X. Chang, "Profit-Aware Spot GPU Admission Control with Cost-Sensitive Loss and Evidence-Grounded Policy Memos for AI Workload Supply-Demand Matching," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 45–59, Jun. 2026, doi: 10.51903/jtie.v5i2.545.

[72] S. Lu and T. Zou, "Uncertainty-Aware Medical Vision–Language Classification on a Lightweight MedMNIST-Compatible Biomedical Patch Benchmark," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 1–19, Jun. 2026, doi: 10.51903/jtie.v5i2.530.

[73] S. He, J. Nie, and C. Li, "Power-Aware Inventory Planning for AI Infrastructure Using Job-Level Forecasting and LLM Workload Explanations," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 341–359, Apr. 2026, doi: 10.51903/jtie.v5i1.548.

[74] C. Wang, Z. Wen, R. Zhang, P. Xu, and Y. Jiang, "GPU Memory Requirement Prediction for Deep Learning Task Based on Bidirectional Gated Recurrent Unit Optimization Transformer," in Proc. 5th Int. Conf. Artificial Intelligence, Virtual Reality and Visualization (AIVRV), Chengdu, China, 2025, doi: 10.1109/AIVRV67401.2025.11350369.

[75] H. Zhou and K. Zhang, "News-Based Uncertainty and Macro-Market Fusion for VIX Direction Forecasting: Evidence from 2015–2024 FRED Panel," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 487–501, Aug. 2025, doi: 10.51903/jtie.v4i2.540.

[76] H. Zhang, "DriftGuard: Multi-Signal Drift Early Warning and Safe Re-Training/Rollback for CTR/CVR Models," J. Adv. Comput. Syst., vol. 3, no. 7, pp. 24–40, 2023.

[77] R. Ma, L. Zhang, and A. Bai, "Cross-Dataset Parcel Workload Priors for Last-Mile Capacity Forecasting: Integrating Package Segmentation with Delivery Operation Signals," J. Technol. Inform. Eng., vol. 5, no. 1, pp. 441–473, Apr. 2026, doi: 10.51903/jtie.v5i1.564.

[78] S. He, X. Chang, and E. Sun, "Cross-Cloud Transfer Learning for AI Training Capacity Forecasting under Workload and Topology Distribution Shift," J. Adv. Comput. Syst., vol. 4, no. 1, pp. 100–120, Jan. 2024, doi: 10.69987/JACS.2024.40108.

[79] Y. Li, S. Lu, and L. Zhao, "LLM-as-Design-Critic: Aligning AI-Generated UI Feedback with Human Graphic Design Judgment," Int. J. Graph. Des., vol.

3, no. 1, pp. 196–215, May 2025, doi: 10.51903/ijgd.v3i1.3661.

[80] Z. S. Zhong, Q. Wu, and G. Mi, "Uncertainty-Aware Medical Image Explanation Cards: LLM-Generated Visual Explanations for AI-Assisted Radiology Interfaces," Int. J. Graph. Des., vol. 3, no. 2, pp. 415–436, Oct. 2025, doi: 10.51903/ijgd.v3i2.3616.

[81] Q. Wu, S. Meng, and J. Zhao, "Text-Grounded LLM-Assisted Design Rationale Interfaces: Turning Advertising Layout Metadata into Explainable UI/UX Decision Cards," Int. J. Graph. Des., vol. 3, no. 1, pp. 216–240, May 2025, doi: 10.51903/ijgd.v3i1.3713.

[82] B. Zhou, C. Li, and L. Liu, "Risk-Calibrated Patient-Facing AI Safety Cards: A UI/UX Design Framework for Rubric-Based Medical Risk Communication," Int. J. Graph. Des., vol. 3, no. 2, pp. 365–380, Oct. 2025, doi: 10.51903/ijgd.v3i2.3696.

[83] Y. Chen and M. Li, "From Hand-Drawn Sketches to Interactive Web Prototypes: A Reproducible Vision-Language Approach with Structural and Visual Consistency Evaluation," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 364–384, Aug. 2025, doi: 10.51903/jtie.v4i2.490.

[84] J. Jin, "LLM-Style Evidence Cards for Scientific Search Interfaces: A UI/UX Design Framework for Retrieval Transparency, Ranking Trust, and Visual Evidence Hierarchy," Int. J. Graph. Des., vol. 3, no. 2, pp. 397–414, Oct. 2025, doi: 10.51903/ijgd.v3i2.3698.

[85] H. Tu, S. Zhao, and A. Zhou, "Visual Brief Cards for Advertising Design: A Structured UI/UX Framework for Turning Creative Intentions into Graphic Design Decisions," Int. J. Graph. Des., vol. 3, no. 1, pp. 210–226, May 2025, doi: 10.51903/ijgd.v3i1.3714.

[86] T. Ye, X. Chang, and E. Zhong, "Uncertainty-Aware Breast Ultrasound Explanation Cards: A Visual Communication Framework for Image-Based AI Diagnostic Support Using BreastMNIST_224," Int. J. Graph. Des., vol. 3, no. 2, pp. 365–380, Oct. 2025, doi: 10.51903/ijgd.v3i2.3701.

[87] J. Mu, Y. Lu, and E. Hwang, "Structured Visual Brief Interfaces for Advertising Design: A UI/UX Framework for Turning Creative Intentions into Designer-Editable Graphic Design Cards," Int. J. Graph. Des., vol. 4, no. 1, pp. 192–208, Apr. 2026, doi: 10.51903/ijgd.v4i1.3702.

[88] C. Li, B. Zhou, and K. Gao, "Risk-Calibrated Patient-Facing AI Safety Cards: A UI/UX Benchmark for Explainable Medical AI Response Interfaces," Int. J. Graph. Des., vol. 3, no. 2, pp. 381–394, Oct. 2025, doi: 10.51903/ijgd.v3i2.3709.

[89] B. Zhang, Y. Ren, and J. Zou, "LLM-Style Explainable E-Commerce Recommendation Cards: A UI/UX Design Framework for Trust-Calibrated Product Recommendation," Int. J. Graph. Des., vol. 3, no. 2, pp. 381–396, Oct. 2025, doi: 10.51903/ijgd.v3i2.3697.

[90] K. Zhang, Y. Chen, and A. Qian, "Evidence-Grounded Accounting Disclosure Review Cards: A Visual Communication Framework for LLM-Style Explanations over SEC Financial Statements and Notes," Int. J. Graph. Des., vol. 3, no. 2, p. 395, Oct. 2025, doi: 10.51903/ijgd.v3i2.3710.

[91] G. Liu, S. He, and H. Wong, "LLM-Compatible Visual Brief Cards for AI Infrastructure Capacity Dashboards: A UI/UX Framework for Turning Forecast Risk into Graphic Design Decisions," Int. J. Graph. Des., vol. 3, no. 1, pp. 196–213, May 2025, doi: 10.51903/ijgd.v3i1.3723.

[92] Z. Li, S. Zhou, and Z. Zhou, "Financial Risk Dashboard Design for Institutional RWA Investors: Visual Hierarchy, Chart Comprehension, and Explainability in FinChart-Bench," Int. J. Graph. Des., vol. 3, no. 1, pp. 196–210, May 2025, doi: 10.51903/ijgd.v3i1.3715.

[93] Y. Zhang and H. Zhang, "Visualizing the Right Counseling Support: Evidence-Linked Recommendation Cards for Explainable Mental Health Intake Interfaces," Int. J. Graph. Des., vol. 3, no. 1, pp. 214–229, May 2025, doi: 10.51903/ijgd.v3i1.3722.

[94] J. Zhang, "Adaptive User Interface Design for Volleyball Learning Apps: Empirical Evidence from Google Play Reviews and Mobile Screen Analysis," Int. J. Graph. Des., vol. 3, no. 1, pp. 175–195, May 2025, doi: 10.51903/ijgd.v3i1.3618.

[95] H. Wang, Y. Ren, and X. Chang, "Layout-Aware Progressive PDF Rendering: AI Prioritization of PDF Slices to Reduce Time-to-Functional-First-Frame on FUNSD," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 425–446, Aug. 2025, doi: 10.51903/jtie.v4i2.523.

[96] B. Zhou, H. Wang, and X. Chang, "Distilling VMAF into an Edge-Deployable Quality Predictor: A Pilot Shot-Level Proxy with LLM-Ready Quality Tokens," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 447–463, Aug. 2025, doi: 10.51903/jtie.v4i2.522.

[97] X. Sun, Y. Lu, and J. Chen, "Controllable Long-Term User Memory for Multi-Session Dialogue: Confidence-Gated Writing, Time-Aware Retrieval-Augmented Generation, and Update/Forgetting," J. Adv. Comput. Syst., vol. 3, no. 8, pp. 9–24, Aug. 2023, doi: 10.69987/JACS.2023.30802.

[98] J. Mu, Y. Lu, and M. Smith, "LLM-Assisted Incrementality (Uplift) Modeling for Email Advertising: From Feature Interactions to Interpretable Audience–Creative–Channel Policies," J. Adv. Comput. Syst., vol. 3, no. 1, pp. 31–48, Jan. 2023, doi: 10.69987/JACS.2023.30103.

[99] M.-J. Kuo, D. Zheng, and J. Hires, "Federated Topic-Preference Learning for Knowledge-Grounded Chat with Differential Privacy," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 385–401, Aug. 2025, doi: 10.51903/jtie.v4i2.502.

[100] H. Weng, H. Wang, and C. Wei, "Adaptive Bidding Strategies for Hybrid Auction Mechanisms in Programmatic Advertising," J. Adv. Comput. Syst., vol. 4, no. 4, pp. 13–25, Apr. 2024, doi: 10.69987/JACS.2024.40402.

[101] X. Chang, Y. Lu, and Z. S. Zhong, "Review-Grounded Explainable Recommendation with Faithfulness Evaluation on Amazon Reviews," J. Electr. Eng. Comput. Sci., vol. 11, no. 1, pp. 9–22, May 2026, doi: 10.54732/jeecs.v11i1.2.

[102] J. Bai, H. Wang, Q. Wu, and B. Zhang, "Privacy-Robust Incrementality Estimation in Cookieless Settings via Uplift Modeling: Reproducible Evidence from the Hillstrom E-Mail Experiment," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 17–38, Apr. 2026, doi: 10.51903/jtie.v5i1.468.

[103] Q. Xin, "Behavior Retrieval plus Response Generation for Interpretable Conversational Personalized Recommendation," IJEEPSE, vol. 9, no. 2, pp. 120–136, Jul. 2026, doi: 10.31258/ijeepse.9.2.120-136.

[104] T. Ye, J. Mu, and J. Hunter, "Off-Policy Evaluation and Conservative Policy Selection for Slot-Level Dynamic Bidding and Ranking on the Open Bandit Dataset (Small)," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 178–199, Apr. 2026, doi: 10.51903/jtie.v5i1.503.

[105] M. Li and H. Zhang, "Personalized Course Promotion in MOOCs: A Hybrid Recommender with LLM-Structured Persuasive Explanation for Educational Advertising," J. Appl. Artif. Intell. Educ., pp. 53–73, 2026.

[106] H. Zhang, "Risk-Aware Budget-Constrained Auto-Bidding Under First-Price RTB: A Distributional Constrained Deep Reinforcement Learning Framework," J. Adv. Comput. Syst., vol. 4, no. 6, pp. 30–47, 2024.

[107] [107] J. Mu, T. Ye, and P. Patel, "Offline Counterfactual Evaluation for Advertising and Recommendation Slot Policies: A Reproducible Study on the Open Bandit Dataset (Small)," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 521–543, Dec. 2025, doi: 10.51903/jtie.v4i3.500.

[108] [108] J. Bai and Q. Wu, "Privacy-Safe Marketing Mix Modeling and Budget Optimization Under Identifier Loss: A Controlled Simulation Study," Int. J. Electron. Commun. Syst., vol. 6, no. 1, Jun. 2026, doi: 10.24042/ijecs.v6i1.30533.

[109] [109] Y. Lu, H. Zhou, and Y. Zhang, "A Constrained, Data-Driven Budgeting Framework Integrating Macro Demand Forecasting and Marketing Response Modeling," J. Technol. Informatics Eng., vol. 4, no. 3, Dec. 2025, doi: 10.51903/jtie.v4i3.466.

[110] [110] H. Zhang, "Counterfactual Learning-to-Rank for Ads: Off-Policy Evaluation on the Open Bandit Dataset," J. Adv. Comput. Syst., vol. 5, no. 12, pp. 1–11, 2025.

[111] [111] X. Li, Y. Shi, and J. Zhang, "Psychology-Informed Live-Commerce Analytics: Structural Modeling, Predictive Validation, and Governed Script Scoring for TikTok Shopping," J. Technol. Inform. Eng., vol. 5, no. 1, pp. 474–502, Apr. 2026, doi: 10.51903/jtie.v5i1.568.

[112] Q. Xin, "Self-Supervised Customer Representation Learning for Segmentation and Next-Purchase Prediction on UCI Online Retail," J. Inf. Technol., vol. 14, no. 1, Apr. 2026, doi: 10.32664/j-intech.v14i01.2229.

[113] K. Xu, H. Zhou, H. Zheng, M. Zhu, and Q. Xin, "Intelligent Classification and Personalized Recommendation of E-Commerce Products Based on Machine Learning," in Proc. 6th Int. Conf. Computing and Data Science (ICCDS), 2024.

[114] Q. Wu, G. Mi, and D. Wood, "Calibration-Light Subject-Independent Motor Imagery BCI via Self-Supervised Pretraining and Conformer," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 239–262, Apr. 2025, doi: 10.51903/jtie.v5i1.493.

[115] Z. S. Zhong, X. Pan, and Q. Lei, "Bridging Domains with Approximately Shared Features," in Proc. 28th Int. Conf. Artificial Intelligence and Statistics (AISTATS), 2025.

[116] J. Zhang, "From General Human Activity Recognition to Volleyball-Oriented Wearable Transfer Learning: Cross-Dataset Evidence from UCI HAR and WISDM for Domain Adaptation and Edge Deployment," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 263–283, Apr. 2025, doi: 10.51903/jtie.v4i1.524.

[117] G. Mi, T. Ye, and D. Wood, "A Lightweight Medical Foundation Model for Cross-Modal Multi-Task Pretraining and Parameter-Efficient Few-Shot Transfer on MedMNIST," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 572–589, Aug. 2025, doi: 10.51903/jtie.v4i3.492.

[118] Q. Xin, "LiDAR–Camera Object-Level Fusion for Multi-Target Tracking Using JPDA and EKF: A Reproducible Empirical Study on a PandaSet-Parameterised Five-Sequence Dataset," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 54–76, Apr. 2026, doi: 10.51903/jtie.v5i1.486.

[119] Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, "Optimization of Autonomous Driving Image Detection Based on RFAConv and Triplet Attention," in Proc. 2nd Int. Conf. Software Engineering and Machine Learning (SEML), 2024.

[120] Q. Xin, Z. Xu, L. Guo, F. Zhao, and B. Wu, "IoT Traffic Classification and Anomaly Detection Method Based on Deep Autoencoders," in Proc. 6th Int. Conf. Computing and Data Science (CDS), 2024.

[121] B. Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, "Predictive Optimization of DDoS Attack Mitigation in Distributed Systems Using Machine Learning," in Proc. 6th Int. Conf. Computing and Data Science (CDS), 2024, pp. 89–94.

[122] Y. Li and S. Lu, "Language-Guided Feature Selection for DDoS and Intrusion Detection on CICIDS2017," J. Technol. Informatics Eng., vol. 4,

no. 1, pp. 284–305, Apr. 2025, doi: 10.51903/jtie.v4i1.531.

[123] Z. Wen, R. Zhang, and C. Wang, "Optimization of Bi-Directional Gated Loop Cell Based on Multi-Head Attention Mechanism for SSD Health State Classification Model," in Proc. 6th Int. Conf. Electronic Communication and Artificial Intelligence (ICECAI), Chengdu, China, 2025, doi: 10.1109/ICECAI66283.2025.11171441.

[124] Z. Zhong, M. Zheng, H. Mai, J. Zhao, and X. Liu, "Cancer Image Classification Based on DenseNet Model," J. Phys.: Conf. Ser., vol. 1651, no. 1, p. 012143, Nov. 2020, doi: 10.1088/1742-6596/1651/1/012143.

[125] J. Chen, J. Xiong, Y. Wang, Q. Xin, and H. Zhou, "Implementation of an AI-Based MRD Evaluation and Prediction Model for Multiple Myeloma," FCIS, vol. 6, no. 3, pp. 127–131, Jan. 2024, doi: 10.54097/zJ4MnbWW.

[126] Q. Xin, "Probabilistic Bike-Sharing Demand Forecasting under Changing Weather and Seasonal Regimes with Transformer-Based Models," Transport Findings, Mar. 2026, doi: 10.32866/001c.157499.

[127] Z. S. Zhong and S. Ling, "Improved Theoretical Guarantee for Rank Aggregation via Spectral Method," Inf. Inference: J. IMA, vol. 13, no. 3, Art. no. iaae020, Sep. 2024, doi: 10.1093/imaiai/iaae020.

[128] Q. Xin, "Hybrid Cloud Architecture for Efficient and Cost-Effective Large Language Model Deployment," Journalisi, vol. 7, no. 3, pp. 2182–2195, Sep. 2025.

[129] Z. S. Zhong and S. Ling, "Uncertainty Quantification of Spectral Estimator and MLE for Orthogonal Group Synchronization," arXiv:2408.05944, Aug. 2024.

[130] L. Zhang, R. Ma, and P. Greg, "Digital-Twin Dispatching for Urban Mobility via Spatio-Temporal Transformers and Offline Reinforcement Learning," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 337–363, Aug. 2025, doi: 10.51903/jtie.v4i2.501.

[131] Y. Zhang and H. Zhang, "A Therapist-Facing Session Copilot for Live Counseling Support: Reasoning-Guided Retrieval and Ranking from Multi-Turn Counseling Dialogues," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 464–486, Aug. 2025, doi: 10.51903/jtie.v4i2.547.

[132] Q. Xin, "Early-Warning Analytics with LLM Intervention Rationales for Student Retention Decisions: Classroom Interaction Modeling with xAPI-Edu-Data and Dropout/Success Prediction," Interdiscip. J. Pedagog. Res. Media Technol., vol. 2, no. 1, Jun. 2026, doi: 10.64268/inspire.v2i1.117.

[133] Y. Zhang and Z. Zhou, "Strategy-Aware Therapist Imitation for Emotional Support Dialogues: A Reproducible ESConv Study for LLM Response Control," Adv. Educ. Technol. Psychol., vol. 10, no. 2, pp. 92–97, 2026, doi: 10.23977/aetp.2026.100213.

[134] J. Zhang, "Early Warning, Grade Prediction, and Teacher-Facing LLM-Ready Explanations Toward an Open Volleyball Course: Reproducible Evidence from Four Public Education Datasets," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 20–44, Jun. 2026, doi: 10.51903/jtie.v5i2.525.

[135] Q. Xin, "Auditable Automated Essay Scoring and Formative Feedback: A Rubric-Grounded Pipeline for Secondary and Higher Education," J. Artificial Intelligence Educ., vol. 2, no. 1, Jul. 2026, doi: 10.66053/jaaie.v2i1.348.

[136] A. G. S. Raj et al., "Impact of Bilingual CS Education on Student Learning and Engagement in a Data Structures Course," in Proc. 19th Koli Calling Int. Conf. Computing Education Research, 2019, pp. 1–10.

[137] H. Tu, Y. Shi, and S. Chen, "IntentRouter-QAC: Distilling LLM-Derived Intent Signals into Small Language Models for Context-Aware Autosuggest and Search Entry Routing," J. Technol. Inform. Eng., vol. 5, no. 1, pp. 418–440, Apr. 2026, doi: 10.51903/jtie.v5i1.563.

[138] Y. Li, "Findable then Explainable: Retrieval–Summary Integration for Code Intelligence on a Lightweight CodeSearchNet Subset," J. Adv. Comput. Syst., vol. 4, no. 7, pp. 65–82, Jul. 2024, doi: 10.69987/JACS.2024.40706.

[139] G. Zhao, D. Zhang, and S. Meng, "LLM-Inspired Ontology-Based Semantic Enrichment for FinTech M&A Intelligence in SEC Structured Disclosures," J. Technol. Inform. Eng., vol. 5, no. 2, pp. 119–141, Aug. 2026, doi: 10.51903/jtie.v5i2.566.

[140] S. Meng, J. Chen, and I. Zheng, "LLM-Inspired Offline Reranking for Financial Search: Query Rewriting, Hybrid Retrieval, and Listwise Relevance Ranking on FiQA," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 361–378, Apr. 2026, doi: 10.51903/jtie.v5i1.537.

[141] S. Zhou, Y. Chen, and K. Lee, "Accounting-Aware Evidence-Constrained Agents for Disclosure, Settlement, and Secondary-Market Risk Monitoring in Tokenized," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 60–74, Jun. 2026, doi: 10.51903/jtie.v5i2.544.

[142] Y. Chen, S. Zhou, and E. Lin, "Accounting-Aware Evidence Retrieval for Institutional Due Diligence of Tokenized Trade Receivable RWA," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 649–663, Dec. 2025, doi: 10.51903/jtie.v4i3.542.

[143] Y. Chen and H. Xu, "Trust-Calibrated Multilingual RAG for Humanitarian Information Platforms: Empirical Evaluation on OMoS-QA for Migration Information Access," Int. J. Graph. Des., vol. 4, no. 1, pp. 141–164, Apr. 2026, doi: 10.51903/ijgd.v4i1.3552.

[144] R. Zhang, Z. Wen, C. Wang, C. Tang, P. Xu, and Y. Jiang, "Quality Analysis and Evaluation Prediction of RAG Retrieval Based on Machine Learning

Algorithms," arXiv:2511.19481, 2025, doi: 10.48550/arXiv.2511.19481.