Stout in Humanities and Social Sciences (SHSS)
ISSN: 2979-2797
DOI: https://doi.org/10.61424/q3a0yd34
Journal Homepage: https://stoutjournals.org/index.php/shss

Who Defines a Good Answer? Demographic-Conditioned Pluralistic LLM Alignment with Preference Uncertainty and Privacy-Aware Personalization

Gavin Guo
Computer Science, University of Arizona, Tucson, AZ, USA
Corresponding Author: Gavin Guo, **E-mail**: g_guo_arizona@yahoo.com

ARTICLE INFORMATION

Submitted: March 10, 2026
Accepted: July 24, 2026
Published: August 01, 2026
Volume: 2
Issue: 1

KEYWORDS

Pluralistic alignment; large language models; human feedback; preference learning; demographic conditioning; personalization; uncertainty calibration; local differential privacy; algorithmic fairness.

ABSTRACT

Large language model alignment usually treats human feedback as if it expressed one stable population objective, although judgments of helpfulness, style, safety, and values differ across people and contexts. This study tested whether demographic conditioning and warm-start personalization improved prediction of pairwise response preferences in PRISM Alignment 2024. PRISM links detailed surveys from 1,500 participants to 8,011 live conversations, 68,371 scored model responses, and text-level language, privacy, and moderation annotations. After English-language and personally identifiable information filters, the analysis formed 54,293 unequal-score comparisons within the same interaction. Whole conversations were assigned within participant to training, validation, and test sets, yielding 36,041, 8,966, and 9,286 pairs. A global logistic preference model was compared with partial-pooled demographic residual models and a personalized stack based on prior preference history and nine stated-preference dimensions. Temperature scaling quantified preference uncertainty, subgroup audits measured worst-group predictive performance, and randomized response traced an event-level local privacy-utility frontier. On held-out conversations, the global model achieved AUC 0.693, negative log-likelihood 0.625, and expected calibration error 0.014. Demographic conditioning produced comparable discrimination (AUC 0.692) and slightly lower calibration error (0.013), but personalization reduced AUC to 0.682 and increased negative log-likelihood to 0.634. The personalized-global AUC difference was -0.0106, with a 95% conversation-cluster bootstrap interval of [-0.0163, -0.0052]. Privacy perturbation changed AUC by less than 0.001 across epsilon 0.5-8 because observed-history personalization contributed little test utility. Preference uncertainty depended strongly on rating margin: global AUC rose from 0.547 for 1-9-point differences to 0.855 for 50-99-point differences. The results show that access to demographic and individual data does not automatically produce better alignment. Population conditioning requires strong regularization, independent calibration, subgroup evaluation, and explicit privacy scope.

1. Introduction

Human preference data has become a central control signal for language-model behavior. Reward learning from comparisons established a practical route from qualitative judgments to trainable objectives, and contemporary instruction-following systems extended that route to large-scale language models (Christiano et al., 2017; Ouyang et al., 2022; Stiennon et al., 2020). Direct preference optimization later showed that a policy can be optimized from the same pairwise structure without an explicit reward-model training stage (Rafailov et al., 2023). These methods are powerful, but their usual statistical target is an average annotator. Averaging is appropriate only when disagreement behaves like noise around a shared latent preference. On moral, political, cultural, and stylistic questions, disagreement instead carries information about distinct conceptions of a good answer.

Pluralistic alignment treats that heterogeneity as part of the target. Sorensen, Moore, et al. (2024) organized pluralism into complementary goals: a model can expose a range of defensible positions, steer behavior to a specified viewpoint, or represent

the distribution of views in a population. Sorensen, Jiang, et al. (2024) demonstrated how values, rights, and duties can conflict inside one apparently simple judgment. Conitzer et al. (2024) further argued that combining diverse feedback is a social-choice problem because the aggregation rule decides whose welfare and judgments count. The technical question is therefore inseparable from a governance question: who defines the labels that a preference model learns?

PRISM Alignment makes this question empirically tractable. The resource connects ratings to participant demographics, geography, stated behavioral preferences, and live multi-turn conversations with 21 language models. Kirk, Whitefield, et al. (2024b) designed the data around subjective and multicultural prompts for which disagreement is expected, while the released files preserve participant and interaction identifiers needed for disaggregated analysis (Kirk, Whitefield, et al., 2024a). This linkage permits direct comparison of three statistical views of alignment. A global model estimates one preference function. A demographic-conditioned model allows broad groups to depart from that function. A personalized model uses an individual's prior choices and stated preferences to estimate a more specific function.

Personalization is not equivalent to retaining more history. Confidence-gated conversational memory distinguishes durable constraints from transient turns, while self-supervised customer representations compress earlier transactions into reusable profiles (Sun et al., 2023; Xin, 2026e). These approaches sharpen the hypothesis tested here: a warm-start profile should improve a new conversation only when it captures preferences that persist across topics rather than the contingent path of earlier prompts and model exposure.

More information does not guarantee a better estimate. Demographic categories can be coarse proxies for lived experience, group-specific models can overfit small samples, and personal histories can encode transient prompts rather than stable values. Santurkar et al. (2023) found that model opinions often misrepresented demographic groups, and Durmus et al. (2024) showed that country-conditioned representations of global opinions require careful measurement. Kirk, Vidgen, et al. (2024) identified privacy loss, profiling, stereotyping, and manipulation as central risks of personalized alignment. A useful empirical design must therefore evaluate average predictive utility, probability calibration, worst-group performance, and the cost of protecting preference histories.

In other pairwise and risk-sensitive settings, probability estimates are treated as decision inputs rather than rankings alone. Credit-risk tiering and multimodal fusion couple discrimination with calibration, while recruiter screening and unanswerable-question answering add selective refusal when confidence or evidence is insufficient (Y. Chen et al., 2023; Jin, 2025a; Xin, 2025b; Zhong, Chen, et al., 2025). This broader evidence motivates evaluating NLL, ECE, Brier score, and risk-coverage behavior alongside AUC.

Conditioning also creates governance obligations because the same attribute can support access, audit, profiling, or exclusion. Multilingual humanitarian retrieval, explainable credit scoring, and prompt-injection oversight interfaces treat trust as a joint property of predictive performance, disclosed uncertainty, and control over sensitive information (Y. Chen & Xu, 2026; Su, Rao, et al., 2026; Xin, 2026c). The present study therefore treats demographic and personal data as features whose incremental value must be demonstrated, not as intrinsically beneficial signals.

This study addressed four research questions. First, how accurately did a shared global model rank responses that the same participant scored differently? Second, did partial-pooled demographic conditioning improve average or worst-group predictive performance? Third, did a warm-start personal profile derived from prior conversations and survey preferences improve held-out choices? Fourth, how did randomized-response privacy budgets change the utility of that history component? Every model used one conversation-disjoint test set, which enabled direct comparison without target or response leakage. The analysis reported pairwise AUC, negative log-likelihood (NLL), expected calibration error (ECE), Brier score, accuracy, subgroup gaps, a privacy-utility frontier, and uncertainty-aware risk-coverage curves.

The central result is cautionary. The global model was the strongest average predictor. Demographic conditioning was nearly neutral in AUC and NLL and modestly improved aggregate ECE and accuracy. The tested personalized stack performed worse overall, even though it slightly improved AUC among participants with only 1-9 training interactions and raised the minimum AUC when all supported single and intersectional categories were pooled. Strong local privacy changed the personalized result very little because the private history signal was already weak. These findings do not reject pluralistic alignment. They show that demographic labels and personal traces are not interchangeable with stable preference structure, and that richer conditioning must earn its complexity through held-out evidence.

2. Literature Review

2.1 Preference learning and aggregation

Pairwise preference learning models the probability that one item is preferred to another from their relative utility. Bradley and Terry (1952) supplied the logistic comparison likelihood that underlies many modern reward models. Christiano et al. (2017) used comparisons to train reward functions for reinforcement learning, and Stiennon et al. (2020) adapted the approach to language summarization. Ouyang et al. (2022) combined supervised instruction tuning, reward modeling, and policy optimization in a widely adopted alignment pipeline. Rafailov et al. (2023) then derived a direct classification objective over preferred and rejected responses. Across these approaches, the pair is the fundamental observational unit, but the annotator distribution determines what the learned score represents.

Pairwise learning also sits within a wider ranking and transfer problem. Spectral rank-aggregation theory studies recovery when comparisons are sparse or corrupted, while approximately shared-feature methods ask which structure survives a domain change (Zhong & Ling, 2024a; Zhong, Pan, et al., 2025). Cross-cloud forecasting and macro-market fusion make the same practical distinction between fitting heterogeneity and transferring it under changing workloads or regimes (S. He et al., 2024; H. Zhou & K. Zhang, 2025). These perspectives support conversation-disjoint testing rather than judging a conditioned model by in-sample fit.

Bakker et al. (2022) trained language models to find statements that received broad approval from people with diverse views. That work made disagreement a design constraint but still pursued an agreement-oriented collective outcome. Conitzer et al. (2024) showed why any such aggregation embeds choices familiar from social choice, including how to trade majority welfare against minority protection. Chakraborty et al. (2024) addressed diverse preferences with max-min RLHF, giving explicit weight to the least-served reward component. Ramesh et al. (2024) developed group-robust preference optimization without a separate reward model. These methods motivate evaluation that reports worst-group loss beside average loss; a single pooled AUC cannot show whether one group bears most errors.

A preference score becomes a policy once it determines which response, creative, or action is shown. Uplift modeling estimates heterogeneous treatment effects rather than average response, and offline bandit evaluation tests a proposed policy without deploying it (Mu et al., 2023; Mu et al., 2025; Ye et al., 2026). Privacy-robust incrementality further shows that heterogeneity estimates must be evaluated under the information regime available at decision time (Bai, Wang, et al., 2026). This policy perspective reinforces the use of one held-out test set for every conditioning strategy.

The same separation between prediction and decision appears in constrained budgeting, privacy-safe marketing-mix optimization, and personalized product recommendation (Bai & Wu, 2026; Y. Lu et al., 2025; K. Xu et al., 2024). These applications differ from dialogue alignment, but they share a relevant lesson: a richer segmentation scheme is useful only when it improves a declared downstream objective under realistic information and resource constraints. The present study accordingly reports probability quality and subgroup behavior instead of treating segmentation itself as success.

Pluralistic systems also differ in architecture. Feng et al. (2024) preserved multiple viewpoints through modular collaboration among language models rather than compressing them into one policy. Sorensen, Moore, et al. (2024) emphasized that no single architecture implements every form of pluralism. A distributional system that mirrors population frequencies serves a different purpose from a steerable system that honors a user's explicit request. Sorensen, Jiang, et al. (2024) represented plural values, rights, and duties in a structured resource, making conflicts visible rather than treating them as annotation defects. The present comparison concerns predictive pluralism: it asks whether observed population and individual attributes explain held-out response choices.

The definition of a good answer also depends on whether its claims can be traced to supporting material. Evidence-grounded DevOps RAG, word-level provenance models, scientific evidence chains, and retrieval-quality predictors operationalize this requirement through coverage, attribution, or faithfulness checks (C. Li et al., 2024; Jin, 2025b; B. Zhang et al., 2024; R. Zhang et al., 2025). These systems address different domains, yet they show why a preference label should not be interpreted as a complete measure of factual reliability.

Evidence requirements are themselves contextual. Bilingual incident triage prioritizes operational grounding, code intelligence integrates retrieval with explanation, institutional due diligence requires accounting-aware support, and compositional question answering must allocate a finite retrieval budget (Y. Chen et al., 2025; Y. Li, 2024; G. Liu et al., 2024; Su et al., 2025). This variety is consistent with pluralistic alignment: response quality can include common criteria such as attribution while still varying with task, user, and consequence.

Retrieved evidence must also constrain the narrative or action that follows. Prototype-based OpenStack explanations, evidence-constrained financial agents, narrative-aware scientific verification, and retrieval-grounded HDFS failure accounts all couple an output to inspectable support (Su, Chen, et al., 2026; X. Sun et al., 2026; Xin, 2025a; S. Zhou et al., 2026). This design pattern is relevant to alignment because a highly preferred answer can still be unsafe if its rationale cannot be checked.

Reliability extends beyond answer text when systems rewrite queries, route tasks, or use tools. Financial-search reranking, trajectory success prediction, cost-aware AIOps routing, and multi-regulation legal retrieval each expose failure modes that an aggregate quality score can conceal (C. Li et al., 2026; J. Li & Zhou, 2026; Meng et al., 2026; Z. S. Zhong et al., 2026). These studies motivate separating predictive preference, evidential reliability, and operational safety rather than collapsing them into one alignment claim.

2.2 Disagreement, demographics, and personalization

Subjective labels are often collapsed through majority vote before modeling. Mostafazadeh Davani et al. (2022) showed that this practice erases systematic disagreement and that annotator-specific models retain useful information. Fleisig et al. (2023) likewise demonstrated that demographic and attitudinal variables can explain parts of annotator variance, while also showing that majority labels can misstate subjective tasks. Gordon et al. (2022) proposed jury learning, in which users configure the composition of an annotator population whose judgments drive a model. These studies support disaggregation, but they do not imply that every demographic split improves generalization. A variable can be socially meaningful and still be a weak predictive feature for a particular response pair.

Language-model studies confirm this distinction. Santurkar et al. (2023) measured which public opinions language models reflected and found uneven alignment across demographic groups. Durmus et al. (2024) evaluated subjective global opinions and documented gaps across countries and topics. PRISM was constructed to connect these population questions to individual behavior. Kirk, Whitefield, et al. (2024b) collected fine-grained surveys and contextual ratings from participants born in 75 countries and residing in 38 countries. Its conversations center values-guided, controversy-guided, and unguided interactions, giving demographic analysis a natural but observational setting. Because participants wrote their own prompts, group differences can reflect topic selection, model exposure, and linguistic style as well as preference.

Personalization operates at a finer level. X. Li et al. (2024) conditioned language modeling on personalized human feedback and demonstrated the value of compact user representations. Poddar et al. (2024) modeled heterogeneous rewards through variational preference learning, allowing a user-specific latent variable and uncertainty over preference modes. These approaches motivate personal adapters or embeddings, but their success depends on repeated, informative observations per user. Theoretical flexibility cannot recover stable individual structure when histories are sparse, contexts shift, or earlier choices determine later model exposure.

Recent personalization architectures separate several functions that a single user vector can blur. Federated topic-preference learning keeps individual signals local, review-grounded recommendation retrieves behaviorally relevant evidence, and counseling systems retrieve or control strategies appropriate to the current dialogue (Chang et al., 2026; Kuo et al., 2025; Y. Zhang & H. Zhang, 2025a; Y. Zhang & Z. Zhou, 2026). Their common implication is not that personalization always helps, but that storage, retrieval, representation, and response control are distinct design decisions.

Safety preferences illustrate why disagreement cannot be reduced to a single permissiveness threshold. Continual red-teaming updates defenses as attacks change, evidence-based self-verification checks a proposed response, and behavior-level refusal attempts to preserve utility while blocking harmful actions (Zheng et al., 2023; Zheng et al., 2024; Zheng & Li, 2024). Dark-pattern detection adds a complementary concern: an apparently fluent interface can manipulate choice even when its content is not conventionally unsafe (H. Xu et al., 2025).

The ethical boundary is equally important. Kirk, Vidgen, et al. (2024) described benefits such as accessibility and user control alongside risks of manipulation, filter bubbles, stereotyping, and intimate inference. Demographic conditioning can essentialize group averages; personal conditioning can expose beliefs or vulnerabilities. Hashimoto et al. (2018) showed that average loss can hide performance on latent minority groups, which motivates a worst-group audit even when average utility does not improve. Fairness in this study therefore means predictive-performance parity across supported categories, not a claim that demographic groups possess fixed or causally distinct preferences.

Calibrated predictions also need a communication layer. Medical explanation cards, patient-facing safety cards, scientific-search evidence cards, and breast-ultrasound uncertainty cards translate confidence, source status, or risk into visual structures

intended for human review (Jin, 2025c; Ye et al., 2025; Zhong, Wu, et al., 2025; B. Zhou et al., 2025). These works do not establish that a card improves the underlying model; they show that uncertainty and evidence must remain legible when a person decides whether to rely on an output.

Similar interface patterns appear in recommendation, accounting review, medical-risk benchmarking, and institutional dashboards (B. Zhang et al., 2025; C. Li et al., 2025; K. Zhang et al., 2025; Z. Li et al., 2025). Across these domains, the recurring design problem is selective disclosure: enough evidence and uncertainty must be visible to support scrutiny without presenting a dense model trace as an explanation. For pluralistic alignment, this means a personalized probability should be accompanied by the information needed to contest or override it.

Human-facing rationales make conditioning inspectable. Advertising rationale interfaces, brief cards, design critics, and editable creative cards externalize the criteria behind a recommendation (Mu et al., 2026; Tu et al., 2025; Q. Wu et al., 2025; Y. Li et al., 2025). Work on vision-language prototyping adds a related requirement: structural and visual consistency should survive the transformation from an input representation to an interactive output (Y. Chen & Li, 2025). This is especially relevant when user attributes influence an answer because the system should communicate the operative preference signal rather than silently infer that a demographic average defines the user.

The same principle spans infrastructure, counseling, education, adaptive learning interfaces, and incident response. Capacity dashboards and incident cards expose risk evidence for operators; counseling cards link recommendations to dialogue evidence; and rubric-grounded educational systems tie feedback or intervention rationales to explicit criteria (G. Liu et al., 2025; Nie et al., 2026; Xin, 2026a; Xin, 2026b; Y. Zhang & H. Zhang, 2025b). Adaptive interface analysis further illustrates that observed usage and screen evidence should inform, rather than be conflated with, stable user preference (J. Zhang, 2025). These examples support a narrow claim for the present study: calibrated uncertainty and subgroup audits should be available to human overseers even when personalization is not deployed.

2.3 Calibration and privacy

Preference probabilities support decisions only when their numerical confidence is meaningful. A model with AUC above chance can still be overconfident about ambiguous comparisons. Guo et al. (2017) formalized reliability diagrams, ECE, and temperature scaling as practical calibration tools. Scaling a logit by one validation-selected temperature preserves ranking while adjusting probability sharpness. In pluralistic alignment, calibration has an additional interpretation: probability near 0.5 represents unresolved preference uncertainty, whether it arises from similar response quality, limited features, or genuine heterogeneity.

Calibration has become a cross-domain requirement wherever probabilities trigger consequential action. Capacity forecasting combines uncertainty with operational planning, default prediction couples calibration with distribution-free bounds, conformal envelopes limit oversubscription risk, and workload forecasts express risk across horizons (S. Chen et al., 2024; S. He et al., 2023; Jin et al., 2024a; Zhao et al., 2025). Extreme-imbalance fraud and bankruptcy studies add that class prevalence and asymmetric error costs can make apparently strong aggregate metrics misleading (Y. Chen et al., 2024; Jin et al., 2024b). The domains differ from preference learning, but the statistical lesson transfers: ranking quality alone cannot determine whether a probability is safe to threshold.

Generalization risk also changes with data availability and deployment conditions. Few-shot tenant forecasting addresses cold starts, medical vision-language classification evaluates uncertainty under limited transfer, and power-aware inventory planning connects forecasts and explanations to downstream commitments (S. He et al., 2025; S. He et al., 2026; S. Lu & Zou, 2026). These studies motivate the manuscript's separation of short histories from longer histories and its refusal to infer benefit merely from having more recorded interactions.

Protecting individual histories introduces a second tradeoff. Warner (1965) introduced randomized response, which masks a sensitive binary answer by probabilistically reporting its opposite. Duchi et al. (2013) formalized local privacy and its statistical efficiency limits. For a binary preference with privacy budget epsilon, retaining the true direction with probability exp(epsilon)/(1 + exp(epsilon)) gives epsilon-local differential privacy for that event. This mechanism fits the present history-profile design because each training comparison can be privatized before it contributes to a participant vector. The guarantee is event-level rather than user-history-level; repeated events compose. That limited but exact scope is preferable to labeling an ordinary noisy model as private.

The literature therefore yields three evaluation principles. First, heterogeneous feedback must be modeled and audited without assuming that conditioning improves accuracy. Second, calibrated probabilities and score-margin analysis must distinguish

confident preferences from close judgments. Third, privacy claims must identify the protected object, mechanism, and budget. The following methodology implements those principles on one consistent PRISM comparison set.

3. Methodology

3.1 Data, joins, and analytical sample

The analysis used the four PRISM Alignment 2024 JSONL files. The survey contained 1,500 participants; 1,396 continued to the conversation stage. The conversation file contained 8,011 conversation trees, and the long-format utterance file contained 68,371 scored prompt-response observations. These three primary files sum to the published 77,882 records. The metadata file added 106,554 language, personally identifiable information (PII), and moderation records for prompts, responses, feedback, self-descriptions, and system strings. Conversations and utterances are two representations of the same interactions and were not counted as independent datasets.

Table 1. PRISM file inventory and analytical units.

Component	Rows	Unique IDs	Analytical unit
Survey	1,500	1,500	participants
Conversations	8,011	8,011	conversation trees
Utterances	68,371	68,371	scored candidate responses
Metadata	106,554	106,554	text-instance safety/language records
Pairwise analysis set	54,293	24,698	unequal-score comparisons

Records were joined at their documented grain. Survey rows joined through user_id, conversations through conversation_id, candidate responses through utterance_id, and prompt metadata through interaction_id. The analysis did not join metadata on user_id alone because that would multiply text-instance records. It also excluded post-outcome fields from prediction: numeric score, score gap, if_chosen, conversation-level performance attributes, choice explanations, open feedback, duration, and completion counts. Scores defined the target but never entered the feature matrix.

The survey composition supplied the conditioning variables. Age was represented as 18-24, 25-34, 35-44, 45-54, and 55 or older. Gender retained female and male categories and combined the small nonbinary and withheld categories for model fitting. Education was collapsed to secondary or less, some tertiary or vocational education, bachelor's degree, graduate or professional degree, and withheld. Residence region retained Africa, the Americas, Asia, Europe, Oceania, and withheld. English proficiency was grouped as native, fluent, or advanced-or-lower. Nine 0-100 stated-preference sliders captured values, creativity, fluency, factuality, diversity, safety, personalisation, helpfulness, and other.

Table 2. Participant composition in the 1,500-record survey.

Attribute	Category	Participants	Percent
Age	25-34	454	30.3%
Age	55+	303	20.2%
Age	18-24	297	19.8%
Age	35-44	237	15.8%
Age	45-54	208	13.9%
Age	Prefer not to say	1	0.1%
Gender	Male	757	50.5%
Gender	Female	718	47.9%
Gender	Gender diverse/prefer not	25	1.7%
Education	Bachelor's	637	42.5%
Education	Some tertiary/vocational	361	24.1%
Education	Secondary or less	252	16.8%
Education	Graduate/professional	241	16.1%
Education	Prefer not to say	9	0.6%
Residence region	Europe	610	40.7%
Residence region	Americas	573	38.2%
Residence region	Oceania	149	9.9%
Residence region	Africa	88	5.9%
Residence region	Asia	79	5.3%
Residence region	Prefer not to say	1	0.1%
English proficiency	Native speaker	886	59.1%
English proficiency	Fluent	405	27.0%
English proficiency	Advanced or lower	209	13.9%

PRISM contains three conversation modes. Unguided conversations accounted for 3,113 records, values-guided conversations for 2,460, and controversy-guided conversations for 2,438. Opening interactions showed as many as four responses from different models; later interactions typically showed two generations from the model selected earlier in the conversation. Figure 1 presents these structural distributions.

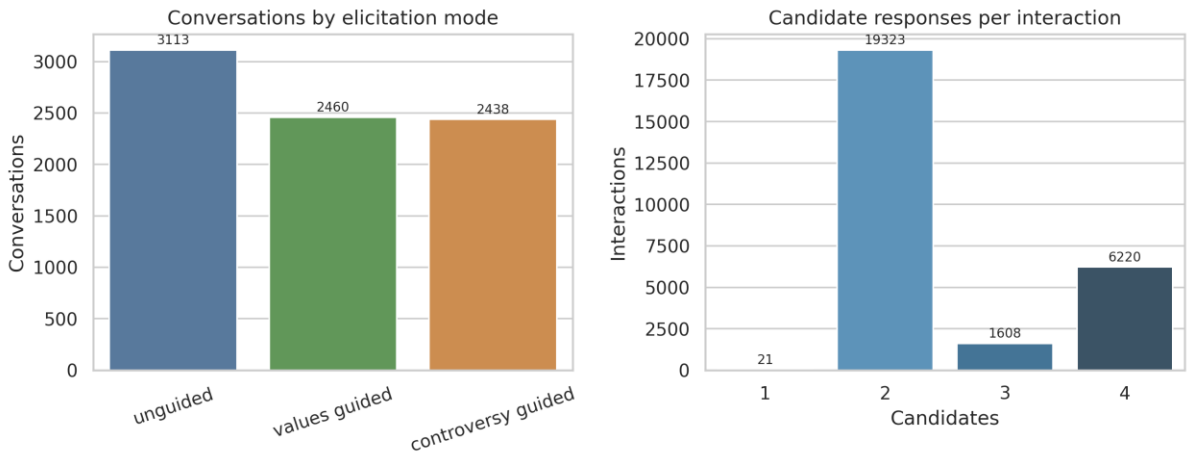

Figure 1. Dataset composition. The three elicitation modes were broadly balanced, while most later interactions contained two candidate generations.

Automated quality filtering retained candidates whose prompt and response metadata indicated English and did not flag PII. Moderation flags were retained as a response feature rather than used as an exclusion criterion, preserving controversial content that belongs to the study design. This process retained 65,261 of 68,371 candidate rows. A total of 25,785 interactions retained at least two eligible candidates. Within each interaction, the procedure enumerated every unordered pair of eligible candidates. Equal numeric scores were removed from the binary task, eliminating 3,259 eligible ties. The remaining 54,293 comparisons came from 24,698 interactions with at least one strict pair.

For each unequal-score pair, the higher-scored response defined the preferred label. Candidate orientation was determined by a SHA-256 hash of the interaction and utterance identifiers, so the left response won in approximately half the cases. The label rate was 0.500 in training, 0.499 in validation, and 0.502 in testing. Four-candidate opening turns generated multiple correlated pairs; the split and bootstrap procedures therefore operated above the pair level.

3.2 Conversation-disjoint splitting

Whole conversations were assigned within participant with fixed seed 20260729. Participants with one conversation contributed it to training. Those with two contributed one to training and one to testing. Those with at least three contributed at least one conversation to validation, one to testing, and the remainder to training. This warm-user design provided prior choices for personalization while preventing prompts, alternative responses, and later turns from the same conversation from crossing partitions.

The resulting training set contained 36,041 pairs from 5,236 conversations and 1,390 participants. Validation contained 8,966 pairs from 1,325 conversations, and testing contained 9,286 pairs from 1,351 conversations. The difference in participant counts reflects conversations without two quality-eligible, unequal-score candidates. The global and demographic models were evaluated on the same warm-user partition as the personalized model; the experiment therefore estimates generalization to new conversations from previously observed participants, not cold-start generalization to unseen people.

Table 3. Conversation-disjoint warm-user experimental splits after language and PII filtering.

Split	Users	Conversations	Interactions	Eligible candidates	Pairs	Left-win rate
train	1,390	5,236	16,415	43,342	36,041	50.0%
val	1,325	1,325	4,074	10,821	8,966	49.9%
test	1,351	1,351	4,209	11,098	9,286	50.2%

Figure 2 summarizes the flow from linked files to pair construction, shared partitions, model comparison, and evaluation.

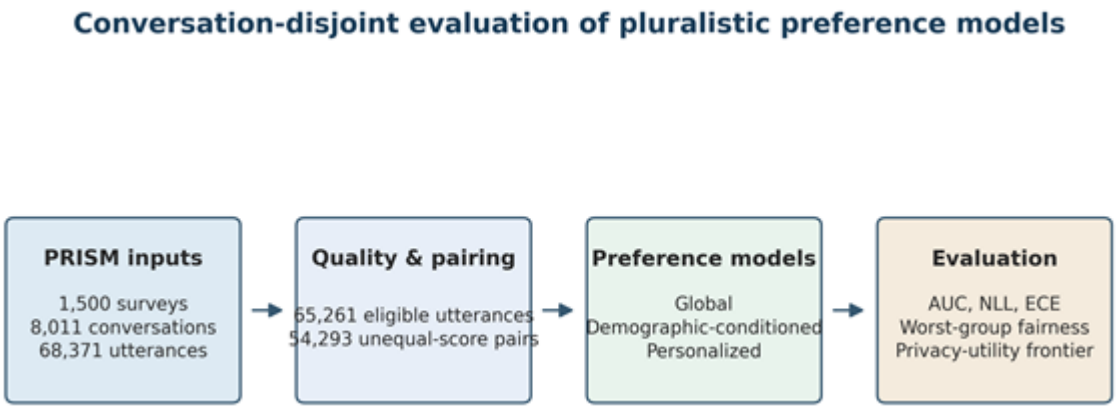

Figure 2. Study design. Whole conversations remained intact across training, validation, and test partitions; all models used the same strict-score comparison set.

3.3 Shared representation

All three models used the same 147-dimensional pair representation. A word unigram-bigram TF-IDF vocabulary was fitted on training responses and prompts only, capped at 25,000 terms with minimum document frequency three. Truncated singular value decomposition compressed training response vectors to 64 latent dimensions; the fitted transform was then applied to validation and test text. These components explained 9.76% of TF-IDF variance, which is expected for a broad multilingual-adjacent conversational vocabulary but limits semantic resolution.

The representation was deliberately shared and low-complexity so that any gain from conditioning could not be attributed to unequal encoders. Multi-source fault attribution and attack-sequence analysis similarly separate representation, diagnosis, and decision layers to determine which component contributes to performance (Bai, Chen, et al., 2026; G. Liu et al., 2023). Here, fixing the 147-dimensional pair vector isolates the incremental contribution of demographic residuals and personal history.

Twelve bounded style features measured log word count, log character count, sentence count, lexical diversity, question and exclamation marks, URLs, list items, refusal markers, hedging markers, empathy markers, and citation-like markers. Twenty-one model-identity indicators represented the left-minus-right difference between candidate generators. Two additional features recorded the difference in prompt-response latent similarity and automated moderation status. Finally, the first 16 semantic dimensions interacted with each of the three conversation modes, adding 48 context-sensitive dimensions. Every candidate-level feature entered the pair as left minus right. A training-fitted standard scaler normalized dimensions.

Table 4. Shared pairwise feature representation.

Feature component	Dimensions	Construction
Latent response semantics	64	TF-IDF (1-2 grams) followed by train-only truncated SVD
Response style	12	Length, diversity, punctuation, lists, URLs, and lexical markers
Model identity	21	Left-minus-right one-hot model identity
Prompt relevance and moderation	2	Latent prompt-response similarity difference and moderation flag difference
Conversation-mode interactions	48	First 16 semantic dimensions interacted with three elicitation modes

3.4 Global, demographic-conditioned, and personalized models

The global model used an L2-regularized logistic Bradley-Terry likelihood. For pair vector x, the probability that the left response was preferred was sigmoid(beta-transpose x). Validation NLL selected regularization C from 0.03, 0.10, 0.30, 1.00, and 3.00; C = 0.03 was selected. This shared function represents the population-average preference surface.

The demographic-conditioned model added partial-pooled residual predictions. Separate logistic models were trained for every supported category of age, gender, education, and residence region with at least 300 training pairs, producing 17 local models. For a test participant, each available category model supplied a logit deviation from the global model. The deviation was shrunk by n/(n + tau), where n was category training support, and deviations were averaged across attributes. Validation tested tau values 100, 300, 1,000, and 3,000; tau = 3,000 minimized NLL. Strong shrinkage kept small-category estimates close to the global function.

The personalized model summarized observed training history. Each training pair vector was normalized and oriented toward the preferred response. Vectors were averaged within interaction so that a four-candidate opening did not dominate a two-candidate turn, then averaged and normalized within participant. For a validation or test pair, cosine similarity with that participant profile measured agreement with earlier preferred directions. A stacked logistic model combined the demographic-conditioned logit, profile similarity, similarity multiplied by log history size, history size, and interactions between nine standardized stated-preference sliders and the first 12 semantic pair dimensions. Conversation-grouped five-fold validation selected C = 0.01 before the stack was fitted on all validation pairs and evaluated once on test.

3.5 Calibration, fairness, uncertainty, and privacy

Temperature scaling used validation logits for the global and demographic-conditioned models. The fitted temperatures were 1.058 and 1.075. Probabilities were evaluated with ROC-AUC, NLL, 15 equal-width-bin ECE, Brier score, and accuracy. AUC measured pair ordering; NLL and Brier score evaluated probability quality; ECE measured the weighted gap between predicted probability and empirical left-win frequency. Two hundred conversation-cluster bootstrap replicates produced 95% intervals and paired model differences.

The metric suite treats confidence as an empirical object rather than a rhetorical qualifier. Numerical-reasoning guardrails use explicit checks before action, conformal alert control couples uncertainty to an operational threshold, and spectral estimation work quantifies uncertainty around recovered latent structure (Z. Li et al., 2026; Xin, 2026d; Zhong & Ling, 2024b). Although these applications differ from PRISM, they support the same evaluation discipline: report ranking, probability quality, and abstention behavior separately.

Predictive-performance disparities were calculated for age, gender, education, residence region, English proficiency, and gender-by-region intersections. A category entered the worst-group summary when it contained at least 150 test pairs. The audit reported minimum AUC, maximum NLL, maximum ECE, and the best-minus-worst AUC gap. These quantities describe model behavior in the observed sample. They do not establish that demographic identity caused a preference or error.

Preference uncertainty was examined in two ways. First, test pairs were divided by absolute human score difference: 1-9, 10-24, 25-49, and 50-99 points. Score gap remained an outcome-only stratification variable. Second, predictions were ranked by absolute distance from 0.5. Selective error was recomputed while coverage increased from the most confident 10% to all pairs.

The privacy experiment protected each observed training-history comparison with binary randomized response. At local budget epsilon, the true direction was retained with probability exp(epsilon)/(1 + exp(epsilon)) and flipped otherwise. Budgets 0.5, 1, 2, 4, and 8 were compared with unperturbed history. Each finite budget was evaluated over ten fixed seeds. The personalized stack was refitted on validation features derived from the privatized profile and evaluated on the unchanged test set.

This mechanism gives event-level epsilon-local privacy for one binary history record. It does not give the same epsilon to an entire multi-event user history because privacy losses compose. It also does not privatize survey attributes, the shared text representation, the global model, or inference-time transmission. The experiment therefore measures the utility of protecting the observed-choice history component, not a complete private training system.

Nearby approaches protect different objects: federated preference learning decentralizes updates; privacy-robust uplift and marketing-mix modeling address identifier loss; and agent risk cards expose sensitive data flows (Bai & Wu, 2026; Bai, Wang, et al., 2026; Kuo et al., 2025; Su, Rao, et al., 2026). Their guarantees do not transfer to the present randomized-response experiment.

4. Results and Discussion

4.1 Overall predictive performance

The global model produced the strongest held-out average. Its AUC was 0.6931 with a conversation-cluster 95% interval of [0.6817, 0.7056], NLL was 0.6247, ECE was 0.0144, Brier score was 0.2188, and accuracy was 0.6264. These values indicate useful but limited discrimination across the diverse prompts and models in PRISM.

Demographic conditioning closely matched the global model. AUC decreased by 0.0007 to 0.6925 and NLL increased by 0.0007 to 0.6254. ECE improved from 0.0144 to 0.0133, and accuracy rose from 0.6264 to 0.6295. The selected tau of 3,000 confirms that validation favored substantial pooling. Broad demographic groups added small probability corrections rather than separate preference functions.

Personalization performed worse overall. AUC fell to 0.6821, NLL increased to 0.6337, ECE increased to 0.0183, and accuracy fell to 0.6195. Across 200 paired conversation bootstraps, the personalized-minus-global AUC difference averaged -0.0106 with interval [-0.0163, -0.0052]. The NLL difference was 0.0088 with interval [0.0051, 0.0121]. The independent test result therefore rejected the proposition that this history-and-survey stack improved average prediction.

Table 5. Held-out pairwise preference performance. AUC intervals are 95% conversation-cluster bootstrap intervals.

Model	AUC (95% CI)	NLL	ECE	Brier	Accuracy
Global	0.693 [0.682, 0.706]	0.625	0.014	0.219	0.626
Demographic-conditioned	0.692 [0.681, 0.705]	0.625	0.013	0.219	0.630
Personalized	0.682 [0.670, 0.695]	0.634	0.018	0.222	0.620

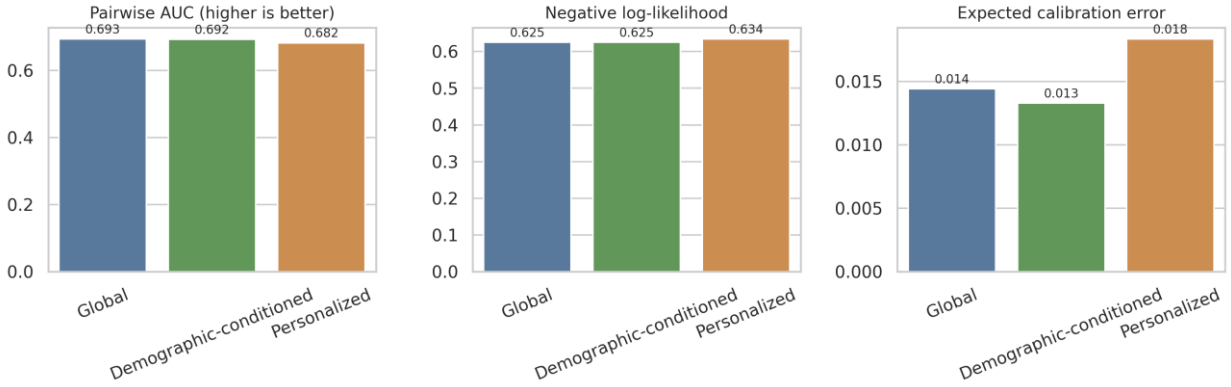

Figure 3. Held-out performance. Demographic conditioning matched the global model closely, while the tested personalization stack reduced discrimination and probability quality.

This negative result is substantively informative. PRISM participants generally authored different prompts in each conversation, and later candidate models depended on earlier selections. A participant profile can therefore encode topic and exposure patterns that do not repeat in the test conversation. The profile also averaged high-dimensional directions across a modest number of interactions. More personal data added degrees of freedom, but validation regularization did not convert those degrees into stable test utility. Poddar et al. (2024) found benefits from a probabilistic latent preference model with architecture and scaling designed for multimodality; the simpler linear profile tested here did not reproduce those benefits.

The result is therefore better read as a representation failure than as evidence that individual preferences do not exist. Controlled memory, federated profiles, review-grounded recommendation, and strategy-aware dialogue control make retrieval or update rules explicit (Chang et al., 2026; Kuo et al., 2025; X. Sun et al., 2023; Y. Zhang & Z. Zhou, 2026). In the present stack, by contrast, one normalized average compressed distinct topics and exposure paths. Its negative test result shows that access to personal history is insufficient unless the representation preserves the aspects of preference that transfer to a new conversation.

4.2 Calibration and preference uncertainty

Temperature scaling improved probability quality without changing AUC. For the global model, scaling reduced NLL from 0.6251 to 0.6247 and ECE from 0.0196 to 0.0144. For the demographic model, it reduced NLL from 0.6260 to 0.6254 and ECE from

0.0183 to 0.0133. Figure 4 shows that all three reliability curves remained close to the diagonal through the dense central probability range, although the personalized curve produced the largest aggregate deviation.

Table 6. Calibration treatment and held-out probability quality.

Model	Probability treatment	Temperature	AUC	NLL	ECE
Global	Raw	1.000	0.693	0.625	0.020
Global	Temperature-scaled	1.058	0.693	0.625	0.014
Demographic-conditioned	Raw	1.000	0.692	0.626	0.018
Demographic-conditioned	Temperature-scaled	1.075	0.692	0.625	0.013
Personalized	Stacked logistic output	1.000	0.682	0.634	0.018

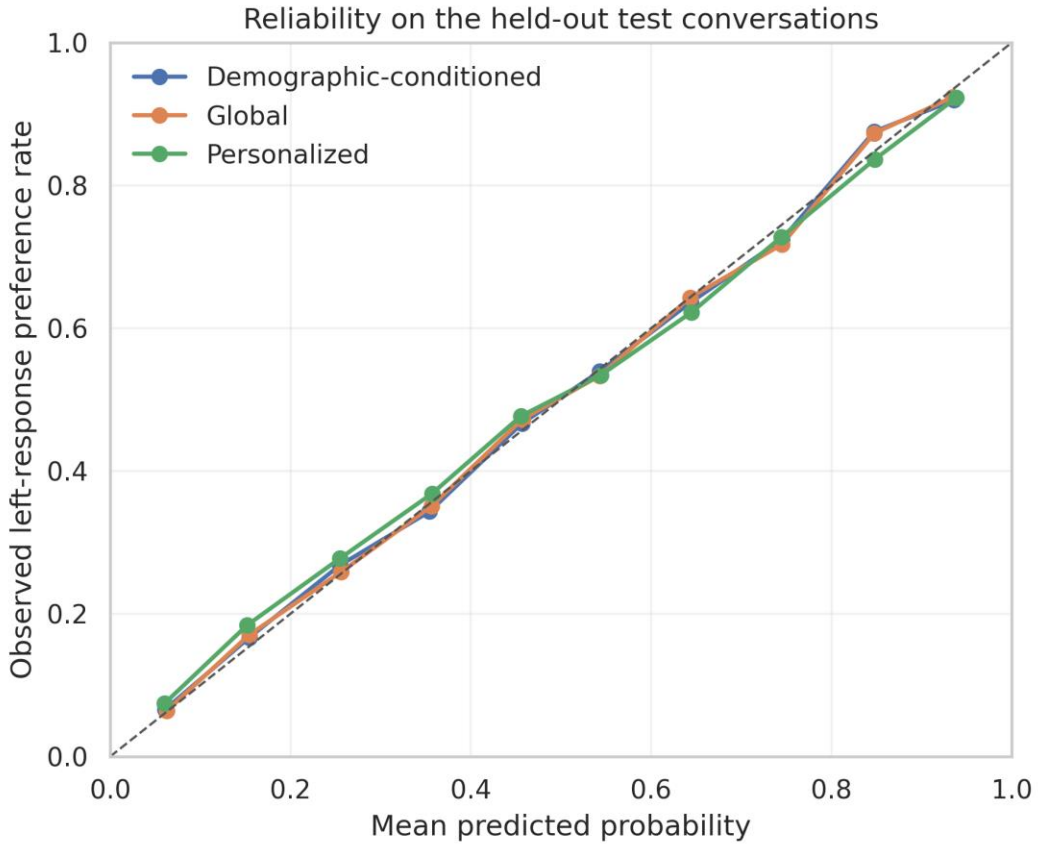

Figure 4. Reliability on held-out conversations. Temperature scaling produced low aggregate ECE for the global and demographic-conditioned models.

Human rating margin exposed a much larger uncertainty gradient than model identity alone. On 2,321 pairs separated by only 1-9 score points, global AUC was 0.5474, NLL was 0.7090, and accuracy was 0.5252. These comparisons were nearly indistinguishable from chance. AUC rose to 0.6304 for 10-24-point gaps, 0.7229 for 25-49-point gaps, and 0.8545 for 50-99-point gaps. The demographic model followed the same pattern. The personalized model did not resolve close judgments and reached AUC 0.5396 in the smallest-margin bin.

Table 7. Preference performance by absolute difference in human response scores.

Score gap	Pairs	Model	AUC	NLL	ECE	Accuracy
1-9	2,321	Global	0.547	0.709	0.061	0.525
1-9	2,321	Demographic-conditioned	0.546	0.710	0.068	0.535
1-9	2,321	Personalized	0.540	0.717	0.078	0.525
10-24	2,750	Global	0.630	0.669	0.029	0.590
10-24	2,750	Demographic-conditioned	0.628	0.671	0.032	0.589
10-24	2,750	Personalized	0.608	0.683	0.040	0.571
25-49	2,169	Global	0.723	0.613	0.032	0.655
25-49	2,169	Demographic-conditioned	0.724	0.613	0.028	0.657
25-49	2,169	Personalized	0.717	0.616	0.026	0.651
50-99	2,046	Global	0.855	0.483	0.059	0.760
50-99	2,046	Demographic-conditioned	0.855	0.481	0.063	0.762
50-99	2,046	Personalized	0.848	0.492	0.052	0.758

Selective prediction translated confidence into operational risk. When the global model retained only its most confident 10% of test pairs, classification error was 0.100; full coverage produced error 0.374. Corresponding demographic errors were 0.098 and 0.370. Personalized errors were 0.117 and 0.380. Thus, calibrated abstention can materially improve the reliability of an automated preference gate even when a richer personal model cannot.

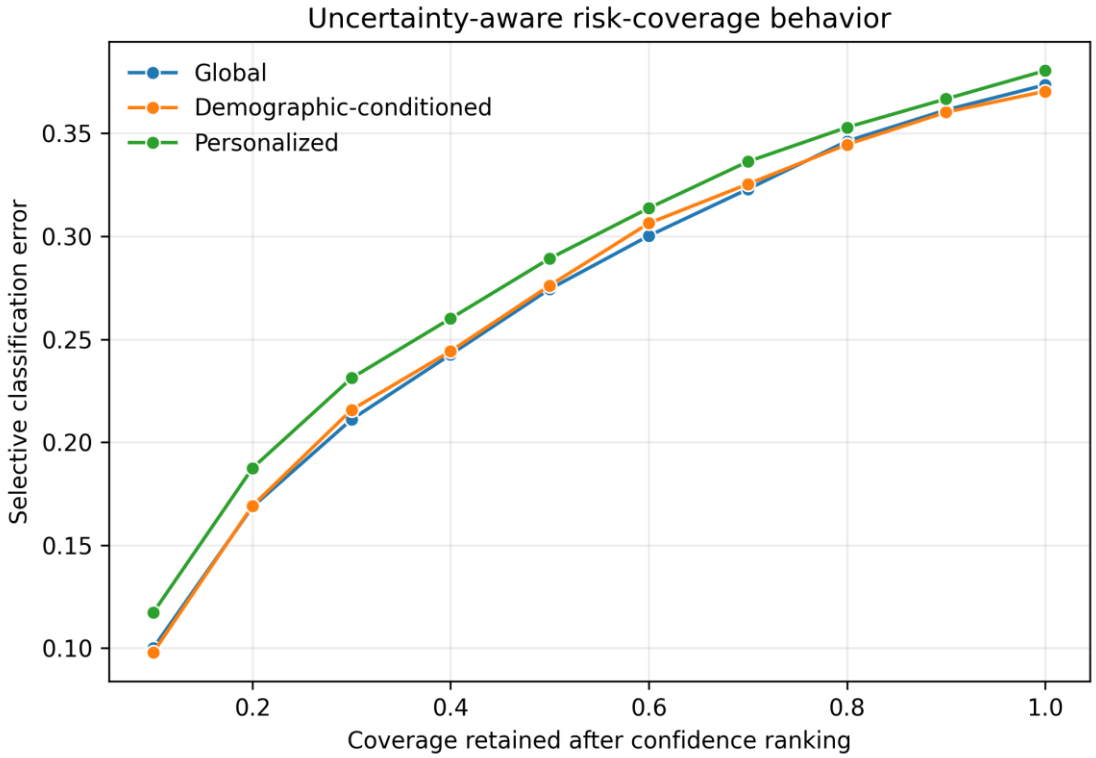

Figure 5. Uncertainty-aware risk-coverage behavior. Retaining the most confident comparisons reduced error for every model.

The risk-coverage result is consistent with selective systems in which uncertain cases are deferred rather than forced into a binary decision. Calibrated credit tiering, recruiter screening, unanswerable-question answering, and conformal alert control apply this logic to different decision surfaces (Y. Chen et al., 2023; Jin, 2025a; Xin, 2026d; Zhong, Chen, et al., 2025). The present experiment does not import their thresholds; it demonstrates the same principle empirically on PRISM by reporting error as coverage changes.

These findings support an interpretation of disagreement as uncertainty rather than automatic evidence for a demographic or personal rule. Mostafazadeh Davani et al. (2022) and Fleisig et al. (2023) established that disagreement can be structured, but close cardinal ratings in PRISM remained intrinsically difficult for every conditioning level. Reporting only binary accuracy would hide this distinction between decisive and marginal human judgments.

4.3 Subgroup performance and worst-group behavior

Subgroup results varied more across residence region than gender. For the personalized model, age-group AUC ranged from 0.6672 among participants aged 35-44 to 0.7033 among those aged 45-54. Gender AUCs were 0.6838 for female participants, 0.6807 for male participants, and 0.6828 for the combined gender-diverse or withheld category. The latter category contained only 185 test pairs and had ECE 0.1337; its apparent AUC parity does not imply stable probability calibration.

Residence-region AUC ranged from 0.6508 in Africa to 0.7256 in Asia for the personalized model. Africa also had NLL 0.6651 and ECE 0.0836. Asia contained 481 test pairs and Africa 516, so both estimates are less stable than those for Europe or the Americas. The variation can reflect prompts, model availability, English usage, and sample composition. It must not be interpreted as an inherent regional preference hierarchy.

Table 8. Personalized-model performance across supported age, gender, and residence-region groups.

Attribute	Group	Pairs	AUC	NLL	ECE
Age	18-24	1,931	0.684	0.634	0.028
Age	25-34	2,829	0.680	0.633	0.019
Age	35-44	1,471	0.667	0.641	0.032
Age	45-54	1,232	0.703	0.619	0.034
Age	55+	1,823	0.680	0.638	0.026
Gender	Female	4,448	0.684	0.630	0.023
Gender	Gender diverse/prefer not	185	0.683	0.616	0.134
Gender	Male	4,653	0.681	0.637	0.025
Residence region	Africa	516	0.651	0.665	0.084
Residence region	Americas	3,418	0.675	0.638	0.021
Residence region	Asia	481	0.726	0.605	0.075
Residence region	Europe	3,904	0.680	0.635	0.025
Residence region	Oceania	957	0.711	0.611	0.028

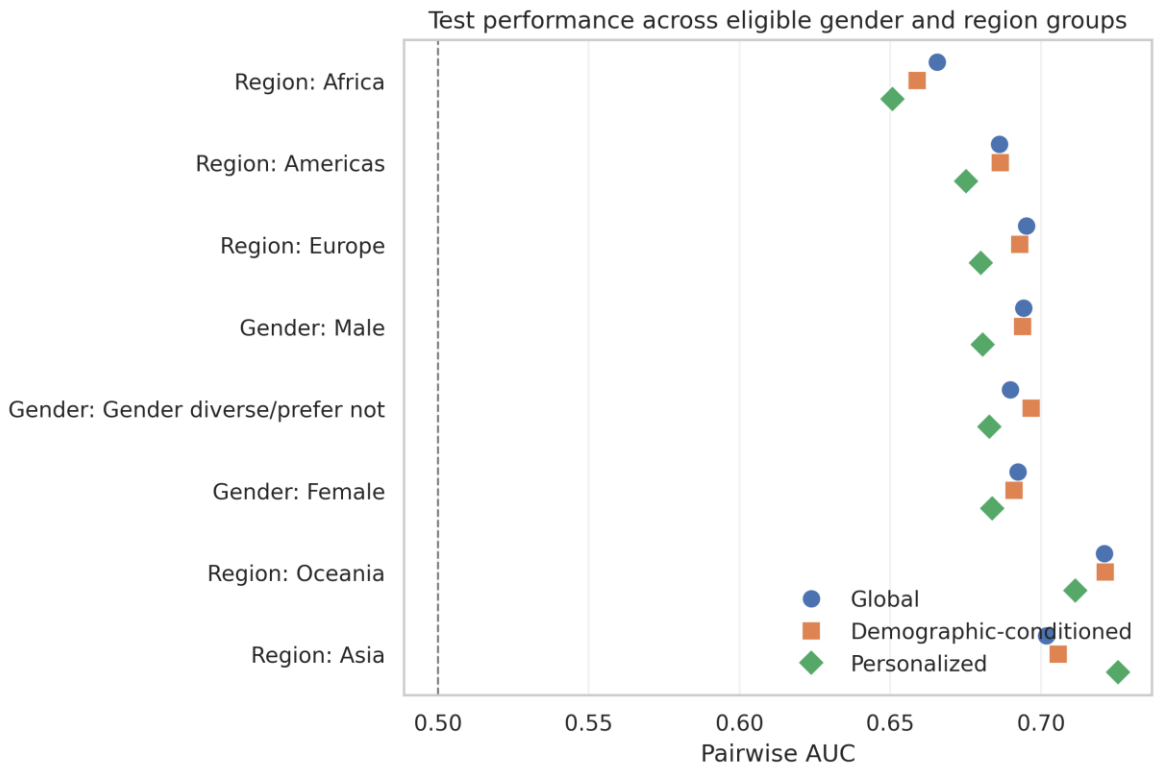

Figure 6. Predictive performance by gender and residence region. The chart reports associations within this sample and does not identify causal demographic effects.

Across all 30 supported single and gender-by-region categories, the global model's minimum AUC was 0.6402 and its AUC gap was 0.0949. Demographic conditioning raised the pooled minimum to 0.6448 and narrowed the gap to 0.0906. Personalization produced a minimum of 0.6466 and gap 0.0852, but its worst-group NLL increased to 0.6947 and average utility declined. The apparent worst-AUC improvement is therefore not a general fairness win. A model can compress rank differences while degrading probability quality for many users.

Table 9. Worst-group predictive-performance summary for categories with at least 150 test pairs.

Model	Scope	Worst AUC	AUC gap	Worst NLL	Worst ECE
Demographic-conditioned	Gender	0.691	0.006	0.627	0.127
Demographic-conditioned	Residence region	0.659	0.062	0.646	0.065
Global	Gender	0.690	0.004	0.626	0.131
Global	Residence region	0.666	0.055	0.645	0.070
Personalized	Gender	0.681	0.003	0.637	0.134
Personalized	Residence region	0.651	0.075	0.665	0.084
Demographic-conditioned	All eligible categories	0.645	0.091	0.667	0.127
Global	All eligible categories	0.640	0.095	0.672	0.131
Personalized	All eligible categories	0.647	0.085	0.695	0.134

For residence region alone, demographic conditioning reduced the worst AUC from 0.6655 to 0.6590 and increased the regional gap from 0.0555 to 0.0623. For gender, all three models had a small AUC gap, but the small combined category produced ECE above 0.12. These mixed directions demonstrate why fairness evaluation requires multiple metrics and declared support thresholds. Hashimoto et al. (2018) motivated protection against hidden worst-case groups; the present results add that a demographic model is not automatically group robust merely because it consumes demographic labels.

The subgroup findings also caution against treating a group label as a causal explanation. Approximately shared-feature learning separates common structure from domain-specific residuals, while multilingual humanitarian retrieval and fair credit scoring audit transfer or error across populations without assuming that membership itself generates the outcome (Y. Chen & Xu, 2026; Xin, 2026c; Zhong, Pan, et al., 2025). PRISM's regional differences should therefore remain descriptive performance audits, exactly as the sampling and exposure limitations imply.

4.4 History size and the limits of personalization

Personalization behaved differently across history sizes. Among the 2,246 test pairs for participants with 1-9 training interactions, personalized AUC was 0.6803 compared with 0.6765 for the global model, and personalized ECE was lower at 0.0321 versus 0.0413. This was the only substantial history band in which the personalized model improved discrimination. Among 6,464 pairs with 10-19 interactions, personalized AUC was 0.6852 versus global AUC 0.7003. For 20-39 interactions, personalized AUC was 0.6618 versus 0.6853. The 40-or-more band contained only 16 pairs and was not inferentially useful.

Table 10. Model performance by the number of participant training-history interactions.

Training interactions	Pairs	Model	AUC	NLL	ECE
1-9	2,246	Global	0.676	0.639	0.041
1-9	2,246	Demographic-conditioned	0.674	0.639	0.043
1-9	2,246	Personalized	0.680	0.639	0.032
10-19	6,464	Global	0.700	0.619	0.016
10-19	6,464	Demographic-conditioned	0.700	0.620	0.008
10-19	6,464	Personalized	0.685	0.630	0.019
20-39	560	Global	0.685	0.634	0.036
20-39	560	Demographic-conditioned	0.684	0.634	0.029
20-39	560	Personalized	0.662	0.655	0.032
40+	16	Global	0.267	0.724	0.168
40+	16	Demographic-conditioned	0.300	0.716	0.163
40+	16	Personalized	0.317	0.737	0.223

More recorded history did not yield monotonic benefit. The profile averaged directions from distinct prompts and model sets, so additional interactions could dilute rather than clarify a stable preference. The result also reflects adaptive exposure: the selected model in an early turn shaped later responses, making a participant's history partly a record of the system path. Li et al. (2024) and Poddar et al. (2024) used representations designed to separate users or preference modes. The current linear summary was intentionally transparent, but its limitations show that personalization needs a validated representation of transferable preference, not merely an identifier or a longer log.

4.5 Privacy-utility frontier

Randomized response produced a shallow privacy-utility curve. The unperturbed personalized model had AUC 0.6821 and NLL 0.6337. At epsilon 2, where each historical direction was retained with probability 0.8808, mean AUC was 0.6820 and NLL was 0.6338. At epsilon 1, retention probability was 0.7311, AUC was 0.6817, and NLL was 0.6340. At epsilon 0.5, retention probability

was 0.6225, AUC was 0.6815 with standard deviation 0.0002, and NLL was 0.6340 with standard deviation 0.0001. Worst-group AUC changed from 0.6466 without perturbation to a mean of 0.6459 at epsilon 0.5.

Table 11. Event-level local privacy and utility under randomized response of preference history.

Epsilon	Retention	AUC mean (SD)	NLL mean (SD)	ECE	Worst-group AUC
Non-private	100.0%	0.6821	0.6337	0.0183	0.6466
8.0	100.0%	0.6821 (0.0000)	0.6337 (0.0000)	0.0182	0.6466
4.0	98.2%	0.6821 (0.0001)	0.6337 (0.0001)	0.0181	0.6466
2.0	88.1%	0.6820 (0.0002)	0.6338 (0.0001)	0.0178	0.6465
1.0	73.1%	0.6817 (0.0002)	0.6340 (0.0001)	0.0179	0.6458
0.5	62.2%	0.6815 (0.0002)	0.6340 (0.0001)	0.0168	0.6459

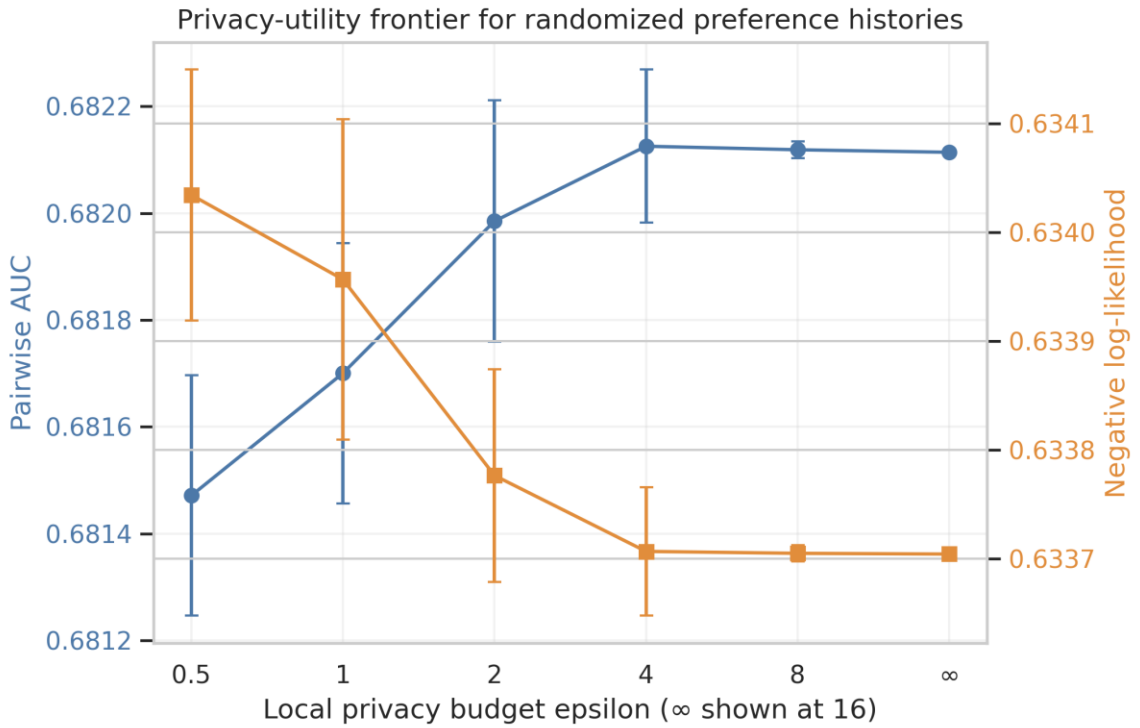

Figure 7. Privacy-utility frontier. Randomized response altered the history component only; survey attributes and the shared model remained unchanged.

The small degradation does not demonstrate cost-free private personalization. It shows that the observed-history similarity contributed little to this stack. The demographic-conditioned logit and unperturbed stated-preference interactions remained available at every budget, while the full personalized model already underperformed the global baseline. A privacy mechanism preserved most of the utility of a weak component. Duchi et al. (2013) established that stronger local privacy generally reduces information; the empirical effect here was bounded by the component's limited predictive influence.

The scope also matters. Each comparison received event-level protection, but a participant contributed multiple comparisons. Basic composition makes the privacy loss for an entire history larger than the per-event epsilon. A deployment seeking user-level protection must allocate a total budget across events or use user-level clipping and accounting. Survey responses would require a separate policy. The experiment nevertheless demonstrates a concrete frontier with a stated mechanism rather than treating generic noise as a privacy guarantee.

A shallow frontier is informative only in relation to the protected component. Federated preference learning and privacy-robust incrementality preserve different information flows, while prompt-injection risk cards address disclosure and approval at the interface (Bai, Wang, et al., 2026; Kuo et al., 2025; Su, Rao, et al., 2026). The present curve cannot be generalized to those settings: it shows that perturbing one weak history feature had little additional cost, not that privacy is free or that the full profile was protected.

4.6 Feature ablation and validity

Feature ablation showed that model identity carried substantial predictive information. A 35-dimensional model-and-style specification achieved AUC 0.6846 and NLL 0.6323. Removing model identity while retaining text, style, context, prompt similarity, and moderation features reduced AUC to 0.6750 and increased NLL to 0.6387. The full 147-dimensional model reached AUC 0.6931 and NLL 0.6247.

Table 12. Global-model feature ablation on the held-out conversations.

Feature set	Dimensions	AUC	NLL	ECE	Accuracy
Model/style only	35	0.685	0.632	0.012	0.625
Text/style, no model identity	126	0.675	0.639	0.014	0.617
Full global feature set	147	0.693	0.625	0.014	0.626

Model identity is both legitimate signal and a benchmark shortcut. Opening turns compare several distinct models, whose average quality differs. Later turns compare samples from the model selected earlier, making identity differences zero while the selection path remains endogenous. The full result therefore predicts preferences in the PRISM interaction process; it is not a causal ranking of model families under identical prompts and exposure. The no-identity ablation provides a more conservative estimate of text-conditioned preference discrimination.

Several additional limitations bound interpretation. First, the within-participant split supports warm-start personalization but not unseen-user generalization. Second, expanding opening interactions creates correlated pairs; conversation-cluster bootstrapping addresses interval estimation but does not change the micro-averaged training objective. Third, participants chose their own prompts, so demographic associations combine preferences with topic and language distributions. Fourth, automated English and PII flags introduce measurement error and remove some interactions unevenly. Fifth, the 64-component semantic representation retained a limited fraction of sparse TF-IDF variance. Sixth, the local privacy experiment covered history labels only. Finally, PRISM is an English-dominant, finite sample; broad geographic participation does not make every subgroup estimate representative.

These constraints are consistent with Gordon et al. (2022), who treated the constitution of an annotator jury as an explicit design choice, and with Chakraborty et al. (2024) and Ramesh et al. (2024), who optimized group objectives directly rather than assuming that conditioning alone provides robustness. Demographic data can support auditing even when it adds little predictive power. Personal data should be collected only when a demonstrated benefit justifies its privacy and governance cost.

5. Conclusions

This study evaluated global, demographic-conditioned, personalized, calibrated, and privacy-aware preference prediction on PRISM Alignment 2024. The global model provided the best average held-out discrimination and likelihood. Strongly regularized demographic residuals produced nearly identical AUC, slightly better aggregate ECE and accuracy, and mixed worst-group effects. A warm-start personal stack based on earlier choices and stated preferences reduced average AUC and worsened NLL. Its modest advantage among short-history users did not persist with longer histories.

Three conclusions follow. First, pluralistic alignment requires evidence that a conditioning variable captures transferable preference rather than prompt choice, exposure, or sample noise. Demographics remain valuable for disaggregated auditing, but their presence does not make a model fair or pluralistic. Second, calibration and rating-margin analysis are essential. The models ranked decisive 50-99-point comparisons well but approached chance on 1-9-point differences; confidence-based coverage reduced error sharply. Third, privacy evaluation must name its protected unit. Event-level randomized response preserved the utility of the history component across epsilon 0.5-8, but that component was weak and the mechanism did not protect the full user profile.

Deployment should therefore separate three questions: whether conditioning improves held-out prediction, whether confidence is calibrated well enough to support abstention, and whether the resulting rationale can be inspected and contested. Evidence and risk-card research offers practical patterns for communicating sources, uncertainty, and safer alternatives to human reviewers (Jin, 2025c; G. Liu et al., 2025; Nie et al., 2026; Su, Rao, et al., 2026). Those interface patterns complement rather than replace the empirical gate established here.

The answer to "Who defines a good answer?" is therefore neither a generic average user nor a demographic label by itself. A defensible system must state whose feedback it models, how it aggregates disagreement, how well its probabilities transfer to new conversations, which groups bear the largest errors, and what personal information it protects. PRISM enables that evaluation precisely because it links context, ratings, and participant characteristics. The empirical lesson is disciplined restraint: personalization and group conditioning should be deployed after independent utility, calibration, fairness, and privacy tests show that they improve the objective they claim to serve.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers
Artificial Intelligence (AI) Use Disclosure: The authors declare that no artificial intelligence tools were used in the preparation of this manuscript.

References

[1] Bai, J., Chen, S., Zheng, D., & Kuo, M.-J. (2026). Interpretable attack-chain stage detection from AWS CloudTrail event sequences via linear models and HMM smoothing. Information Electrical and Electronics Engineering, 6(1), 28–43. https://doi.org/10.33474/infotron.v6i1.24923
[2] Bai, J., Wang, H., Wu, Q., & Zhang, B. (2026). Privacy-robust incrementality estimation in cookieless settings via uplift modeling: Reproducible evidence from the Hillstrom e-mail experiment. *Journal of Technology Informatics and Engineering,* 5(1), 17–38. https://doi.org/10.51903/jtie.v5i1.468
[3] Bai, J., & Wu, Q. (2026). Privacy-safe marketing mix modeling and budget optimization under identifier loss: A controlled simulation study. *International Journal of Electronics and Communications Systems, 6*(1), 83–95. https://doi.org/10.24042/ijecs.v6i1.30533
[4] Bakker, M., Chadwick, M., Sheahan, H., Tessler, M., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M., & Summerfield, C. (2022). Fine-tuning language models to find agreement among humans with diverse preferences. Advances in Neural Information Processing Systems, 35, 38176-38189. https://proceedings.neurips.cc/paper/2022/hash/f978c8f3b5f399cae464e85f72e28503-Abstract-Conference.html
[5] Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. Biometrika, 39(3-4), 324-345. https://doi.org/10.1093/biomet/39.3-4.324
[6] Chakraborty, S., Qiu, J., Yuan, H., Koppel, A., Manocha, D., Huang, F., Bedi, A., & Wang, M. (2024). MaxMin-RLHF: Alignment with diverse human preferences. In Proceedings of the 41st International Conference on Machine Learning (Vol. 235, pp. 6116-6135). PMLR. https://proceedings.mlr.press/v235/chakraborty24b.html
[7] Chang, X., Lu, Y., & Zhong, Z. S. (2026). Review-grounded explainable recommendation with faithfulness evaluation on Amazon reviews. *Journal of Electrical Engineering and Computer Sciences*, 11(1), 9–22. https://doi.org/10.54732/jeecs.v11i1.2
[8] Chen, S., He, S., & Sun, E. (2024). Risk-bounded GPU resource oversubscription via conformal demand envelopes in production AI clusters. *Journal of Advanced Computing Systems, 4(*5), 119–134. https://doi.org/10.69987/JACS.2024.40509
[9] Chen, Y., & Li, M. (2025). From hand-drawn sketches to interactive web prototypes: A reproducible vision-language approach with structural and visual consistency evaluation. *Journal of Technology Informatics and Engineering, 4*(2), 364–384. https://doi.org/10.51903/jtie.v4i2.490
[10] Chen, Y., & Xu, H. (2026). Trust-calibrated multilingual RAG for humanitarian information platforms: Empirical evaluation on OMoS-QA for migration information access. *International Journal of Graphic Design, 4*(1), 141–164. https://doi.org/10.51903/ijgd.v4i1.3552
[11] Chen, Y., Zhang, Y., Chau, D., & Sherman, M. (2023). Credit card default risk tiering with probability calibration and uncertainty-driven rejection: A reproducible study on the UCI credit card clients dataset. *Journal of Advanced Computing Systems, 3(*4), 31–47. https://doi.org/10.69987/JACS.2023.30403
[12] Chen, Y., Zhang, Y., & Sherman, M. (2024). Going concern and bankruptcy prediction under extreme class imbalance: Cost-sensitive learning, resampling, and focal loss with explainable financial-ratio portraits. *Journal of Advanced Computing Systems*, 4(4), 80–96. https://doi.org/10.69987/JACS.2024.40407
[13] Chen, Y., Zhou, S., & Lin, E. (2025). Accounting-aware evidence retrieval for institutional due diligence of tokenized trade receivable RWA. *Journal of Technology Informatics and Engineering*, 4(3), 649–663. https://doi.org/10.51903/jtie.v4i3.542
[14] Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in Neural Information Processing Systems, 30. https://proceedings.neurips.cc/paper/7017-deep-reinforcement-learning-from-human-preferences
[15] Conitzer, V., Freedman, R., Heitzig, J., Holliday, W. H., Jacobs, B. M., Lambert, N., Mosse, M., Pacuit, E., Russell, S., Schoelkopf, H., Tewolde, E., & Zwicker, W. S. (2024). Position: Social choice should guide AI alignment in dealing with diverse human feedback. In Proceedings of the 41st International Conference on Machine Learning (Vol. 235, pp. 9346-9360). PMLR. https://proceedings.mlr.press/v235/conitzer24a.html
[16] Duchi, J. C., Jordan, M. I., & Wainwright, M. J. (2013). Local privacy and statistical minimax rates. In 2013 IEEE 54th Annual Symposium on Foundations of Computer Science (pp. 429-438). IEEE. https://doi.org/10.1109/FOCS.2013.53

[17] Durmus, E., Nguyen, K., Liao, T. I., Schiefer, N., Askell, A., Bakhtin, A., Chen, C., Hatfield-Dodds, Z., Hernandez, D., Joseph, N., Lovitt, L., McCandlish, S., Sikder, O., Tamkin, A., Thamkul, J., Kaplan, J., Clark, J., & Ganguli, D. (2024). Towards measuring the representation of subjective global opinions in language models. In Proceedings of the First Conference on Language Modeling. https://openreview.net/forum?id=zI16jLb91v

[18] Feng, S., Sorensen, T., Liu, Y., Fisher, J., Park, C. Y., Choi, Y., & Tsvetkov, Y. (2024). Modular pluralism: Pluralistic alignment via multi-LLM collaboration. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (pp. 4151-4171). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.240

[19] Fleisig, E., Abebe, R., & Klein, D. (2023). When the majority is wrong: Modeling annotator disagreement for subjective tasks. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (pp. 6715-6726). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.415

[20] Gordon, M. L., Lam, M. S., Park, J. S., Patel, K., Hancock, J. T., Hashimoto, T., & Bernstein, M. S. (2022). Jury learning: Integrating dissenting voices into machine learning models. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (Article 115, pp. 1-19). Association for Computing Machinery. https://doi.org/10.1145/3491102.3502004

[21] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning (Vol. 70, pp. 1321-1330). PMLR. https://proceedings.mlr.press/v70/guo17a.html

[22] Hashimoto, T., Srivastava, M., Namkoong, H., & Liang, P. (2018). Fairness without demographics in repeated loss minimization. In Proceedings of the 35th International Conference on Machine Learning (Vol. 80, pp. 1929-1938). PMLR. https://proceedings.mlr.press/v80/hashimoto18a.html

[23] He, S., Chang, X., & Sun, E. (2024). Cross-cloud transfer learning for AI training capacity forecasting under workload and topology distribution shift. *Journal of Advanced Computing Systems, 4*(1), 100–120. https://doi.org/10.69987/JACS.2024.40108

[24] He, S., Li, C., & Rao, H. (2025). Few-shot cold-start workload forecasting for new AI inference tenants with time-series foundation models. Journal of Technology Informatics and Engineering, 4(1), 306–324. https://doi.org/10.51903/jtie.v4i1.546

[25] He, S., Nie, J., & Li, C. (2026). Power-aware inventory planning for AI infrastructure using job-level forecasting and LLM workload explanations. Journal of Technology Informatics and Engineering, 5(1), 341–359. https://doi.org/10.51903/jtie.v5i1.548

[26] He, S., Tu, H., & Liu, I. (2023). Safe PD capacity forecasting with time-series foundation models and calibrated uncertainty for heterogeneous GPU clusters. *Journal of Advanced Computing Systems, 3*(4), 48–66. https://doi.org/10.69987/JACS.2023.30404

[27] Jin, J. (2025a). Calibrated resume-job matching for trustworthy LLM-assisted recruiter screening: Pairwise matching, probability calibration, and selective refusal on two public recruitment datasets. *Journal of Technology Informatics and Engineering*, 4(3), 625–648. https://doi.org/10.51903/jtie.v4i3.529

[28] Jin, J. (2025b). Evidence-chain reliable RAG: Hallucination detection, source attribution, and deterministic provenance explanations. Journal of Technology Informatics and Engineering, 4(2), 520–533. https://doi.org/10.51903/jtie.v4i2.535

[29] Jin, J. (2025c). LLM-style evidence cards for scientific search interfaces: A UI/UX design framework for retrieval transparency, ranking trust, and visual evidence hierarchy. *International Journal of Graphic Design, 3*(2), 397–414. https://doi.org/10.51903/ijgd.v3i2.3698

[30] Jin, J., Huang, T., & Lu, S. (2024a). A model-risk-friendly probability of default workflow: Calibration, distribution-free uncertainty quantification, and SHAP explanations on the UCI credit card default dataset. Journal of Advanced Computing Systems, 4(6), 74–85. https://doi.org/10.69987/JACS.2024.40606

[31] Jin, J., Huang, T., & Lu, S. (2024b). Cost-sensitive learning, simulated PU learning, and one-class autoencoding for extreme-imbalance credit card fraud detection. *Journal of Advanced Computing Systems, 4*(6), 64–73. https://doi.org/10.69987/JACS.2024.40605

[32] Kirk, H. R., Vidgen, B., Röttger, P., & Hale, S. A. (2024). The benefits, risks and bounds of personalizing the alignment of large language models to individuals. Nature Machine Intelligence, 6, 383-392. https://doi.org/10.1038/s42256-024-00820-y

[33] Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., & Hale, S. A. (2024a). The PRISM alignment dataset [Data set]. Hugging Face. https://doi.org/10.57967/hf/2113

[34] Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., & Hale, S. A. (2024b). The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. Advances in Neural Information Processing Systems, 37, 105236-105344. https://doi.org/10.52202/079017-3342

[35] Kuo, M.-J., Zheng, D., & Hires, J. (2025). Federated topic-preference learning for knowledge-grounded chat with differential privacy. *Journal of Technology Informatics and Engineering*, 4(2), 385–401. https://doi.org/10.51903/jtie.v4i2.502

[36] Li, C., Bai, J., & Wang, S. (2024). Evidence-chain reliable RAG: Word-level hallucination detection, source attribution, and provenance explanation for LLM applications. *Journal of Advanced Computing Systems, 4(*2), 76–92. https://doi.org/10.69987/JACS.2024.40207

[37] Li, C., Liu, G., & Zhao, Z. (2026). Cost-aware LLM-style routing for AIOps log analysis: Log parsing, anomaly detection, fault diagnosis, and incident summarization on LogEval task files. *Journal of Technology Informatics and Engineering, 5(*2), 91–103. https://doi.org/10.51903/jtie.v5i2.538

[38] Li, C., Zhou, B., & Gao, K. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX benchmark for explainable medical AI response interfaces. *International Journal of Graphic Design*, 3(2), 381–394. https://doi.org/10.51903/ijgd.v3i2.3709

[39] Li, J., & Zhou, A. (2026). Multi-regulation RAG for AI product counsel: A legal governance framework for cross-border digital commerces. *Rule of Law Studies Journal,* 2(2), 105–123. https://doi.org/10.64780/rolsj.v2i2.225

[40] Li, X., Zhou, R., Lipton, Z. C., & Leqi, L. (2024). Personalized language modeling from personalized human feedback. arXiv. https://doi.org/10.48550/arXiv.2402.05133

[41] Li, Y. (2024). Findable then explainable: Retrieval–summary integration for code intelligence on a lightweight CodeSearchNet subset. *Journal of Advanced Computing Systems, 4*(7), 65–82. https://doi.org/10.69987/JACS.2024.40706

[42] Li, Y., Lu, S., & Zhao, L. (2025). LLM-as-design-critic: Aligning AI-generated UI feedback with human graphic design judgment. *International Journal of Graphic Design, 3*(1), 196–215. https://doi.org/10.51903/ijgd.v3i1.3661

[43] Li, Z., Zhang, K., & Wong, A. (2026). Numerical-reasoning guardrails for a quant research assistant: A compact reproducible benchmark using SEC and FRED data. *Journal of Technology Informatics and Engineering, 5*(2), 75–90. https://doi.org/10.51903/jtie.v5i2.541

[44] Li, Z., Zhou, S., & Zhou, Z. (2025). Financial risk dashboard design for institutional RWA investors: Visual hierarchy, chart comprehension, and explainability in FinChart-Bench. *International Journal of Graphic Design*, 3(1), 196–210. https://doi.org/10.51903/ijgd.v3i1.3715

[45] Liu, G., He, S., & Liu, I. (2023). LLM-Augmented multi-source root cause attribution for CPU and network faults in microservices. *Journal of Advanced Computing Systems, 3*(6), 39–57. https://doi.org/10.69987/JACS.2023.30604

[46] Liu, G., He, S., & Wong, H. (2025). LLM-compatible visual brief cards for AI infrastructure capacity dashboards: A UI/UX framework for turning forecast risk into graphic design decisions. *International Journal of Graphic Design, 3*(1), 196–213. https://doi.org/10.51903/ijgd.v3i1.3723

[47] Liu, G., Li, C., & Zhang, E. (2024). OpsLLM for cloud incident triage: Bilingual RAG-based root cause analysis and alert summarization for AI infrastructure operations. Journal of Advanced Computing Systems, 4(4), 97–111. https://doi.org/10.69987/JACS.2024.40408

[48] Lu, S., & Zou, T. (2026). Uncertainty-aware medical vision–language classification on a lightweight MedMNIST-compatible biomedical patch benchmark. *Journal of Technology Informatics and Engineering, 5*(2), 1–19. https://doi.org/10.51903/jtie.v5i2.530

[49] Lu, Y., Zhou, H., & Zhang, Y. (2025). A constrained, data-driven budgeting framework integrating macro demand forecasting and marketing response modeling. *Journal of Technology Informatics and Engineering, 4*(3), 493–520. https://doi.org/10.51903/jtie.v4i3.466

[50] Meng, S., Chen, J., & Zheng, I. (2026). LLM-inspired offline reranking for financial search: Query rewriting, hybrid retrieval, and listwise relevance ranking on FiQA. *Journal of Technology Informatics and Engineering, 5*(1), 361–378. https://doi.org/10.51903/jtie.v5i1.537

[51] Mostafazadeh Davani, A., Díaz, M., & Prabhakaran, V. (2022). Dealing with disagreements: Looking beyond the majority vote in subjective annotations. Transactions of the Association for Computational Linguistics, 10, 92-110. https://doi.org/10.1162/tacl_a_00449

[52] Mu, J., Lu, Y., & Hwang, E. (2026). Structured visual brief interfaces for advertising design: A UI/UX framework for turning creative intentions into designer-editable graphic design cards. *International Journal of Graphic Design, 4*(1), 192–208. https://doi.org/10.51903/ijgd.v4i1.3702

[53] Mu, J., Lu, Y., & Smith, M. (2023). LLM-assisted incrementality (uplift) modeling for email advertising: From feature interactions to interpretable audience–creative–channel policies. *Journal of Advanced Computing Systems, 3*(1), 31–48. https://doi.org/10.69987/JACS.2023.30103

[54] Mu, J., Ye, T., & Patel, P. (2025). Offline counterfactual evaluation for advertising and recommendation slot policies: A reproducible study on the Open Bandit Dataset (small). *Journal of Technology Informatics and Engineering,* 4(3), 521–543. https://doi.org/10.51903/jtie.v4i3.500

[55] Nie, J., Liu, G., Li, C., & Zou, T. (2026). 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. *International Journal of Graphic Design, 4*(1), 179–185. https://doi.org/10.51903/ijgd.v4i1.3703

[56] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744. https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

[57] Poddar, S., Wan, Y., Ivison, H., Gupta, A., & Jaques, N. (2024). Personalizing reinforcement learning from human feedback with variational preference learning. Advances in Neural Information Processing Systems, 37, 52516-52544. https://doi.org/10.52202/079017-1664

[58] Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., & Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 36, 53728-53741. https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html

[59] Ramesh, S. S., Hu, Y., Chaimalas, I., Mehta, V., Sessa, P. G., Bou Ammar, H., & Bogunovic, I. (2024). Group robust preference optimization in reward-free RLHF. Advances in Neural Information Processing Systems, 37, 37100-37137. https://doi.org/10.52202/079017-1171

[60] Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., & Hashimoto, T. (2023). Whose opinions do language models reflect? In Proceedings of the 40th International Conference on Machine Learning (Vol. 202, pp. 29971-30004). PMLR. https://proceedings.mlr.press/v202/santurkar23a.html

[61] Sorensen, T., Jiang, L., Hwang, J. D., Levine, S., Pyatkin, V., West, P., Dziri, N., Lu, X., Rao, K., Bhagavatula, C., Sap, M., Tasioulas, J., & Choi, Y. (2024). Value kaleidoscope: Engaging AI with pluralistic human values, rights, and duties. Proceedings of the AAAI Conference on Artificial Intelligence, 38(18), 19937-19947. https://doi.org/10.1609/aaai.v38i18.29970

[62] Sorensen, T., Moore, J., Fisher, J., Gordon, M. L., Mireshghallah, N., Rytting, C. M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., & Choi, Y. (2024). Position: A roadmap to pluralistic alignment. In Proceedings of the 41st International Conference on Machine Learning (Vol. 235, pp. 46280-46302). PMLR. https://proceedings.mlr.press/v235/sorensen24a.html

[63] Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., & Christiano, P. F. (2020). Learning to summarize with human feedback. Advances in Neural Information Processing Systems, 33, 3008-3021. https://proceedings.neurips.cc/paper/2020/hash/1f89885d556929e98d3ef9b86448f951-Abstract.html

[64] Su, W., Chen, S., & Qian, E. (2026). Narrative-aware scientific claim verification agent with evidence ranking for ClimateCheck. Journal of Technology Informatics and Engineering, 5(1), 327–340. https://doi.org/10.51903/jtie.v5i1.549

[65] Su, W., Chen, S., & Zhao, C. (2025). Budgeted multi-hop retrieval agent for compositional question answering: A retrieval-policy evaluation on the official MultiHop-RAG benchmark. *Journal of Technology Informatics and Engineering,* 4(3), 649–662. https://doi.org/10.51903/jtie.v4i3.543

[66] Su, W., Rao, H., & Ma, E. (2026). Privacy and data-integrity risk cards for LLM agents: A UI/UX design framework for secure human oversight under prompt-injection attacks. *International Journal of Graphic Design, 4*(1), 186–191. https://doi.org/10.51903/ijgd.v4i1.3699

[67] Sun, X., Lu, Y., & Chen, J. (2023). Controllable long-term user memory for multi-session dialogue: Confidence-gated writing, time-aware retrieval-augmented generation, and update/forgetting. Journal of Advanced Computing Systems, 3(8), 9–24. https://doi.org/10.69987/JACS.2023.30802

[68] Sun, X., Zhong, Z. S., & Wu, Q. (2026). Retrieval-grounded HDFS log anomaly detection and deterministic failure narrative generation. *Journal of Computer Systems and Applications,* 3(1), 15–30. https://doi.org/10.64229/j6d7fr94

[69] Tu, H., Zhao, S., & Zhou, A. (2025). Visual brief cards for advertising design: A structured UI/UX framework for turning creative intentions into graphic design decisions. *International Journal of Graphic Design, 3*(1), 210–226. https://doi.org/10.51903/ijgd.v3i1.3714

[70] Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association, 6*0(309), 63-69. https://doi.org/10.1080/01621459.1965.10480775

[71] Wu, Q., Meng, S., & Zhao, J. (2025). Text-grounded LLM-assisted design rationale interfaces: Turning advertising layout metadata into explainable UI/UX decision cards. *International Journal of Graphic Design, 3*(1), 216–240. https://doi.org/10.51903/ijgd.v3i1.3713

[72] Xin, Q. (2025a). Explaining OpenStack failure-injection log anomalies with retrieved normal prototypes. Emerging Information Science and Technology, 6(2), 125–146. https://doi.org/10.18196/eist.v6i2.31232

[73] Xin, Q. (2025b). Uncertainty-aware late fusion for 3D perception (confidence calibration + fusion rule learning*). Journal of Technology Informatics and Engineering, 4(*1), 215–238. https://doi.org/10.51903/jtie.v4i1.485

[74] Xin, Q. (2026a). Auditable automated essay scoring and formative feedback: A rubric-grounded pipeline for secondary and higher education. *Journal of Applied Artificial Intelligence in Education, 2*(1), 1–19. https://doi.org/10.66053/jaaie.v2i1.348

[75] Xin, Q. (2026b). Early-warning analytics with LLM intervention rationales for student retention decisions: Classroom interaction modeling with xAPI-edu-data and dropout/success prediction. *Interdisciplinary Journal of Pedagogical Research and Media Technology, 2(*1). https://doi.org/10.64268/inspire.v2i1.117

[76] Xin, Q. (2026c). Explainable and fair credit risk scoring with counterfactual explanations: A reproducible evaluation on the German credit dataset (HELOC-Motivated). *Journal of Information and Technology, 14*(2). https://doi.org/10.32664/j-intech.v14i02.2228

[77] Xin, Q. (2026d). Log anomaly detection with conformal alert control and evidence-grounded incident ticket generation. AVITEC, 8(2), 247. https://doi.org/10.28989/avitec.v8i2.3974

[78] Xin, Q. (2026e). Self-supervised customer representation learning for segmentation and next-purchase prediction on UCI online retail. J-INTECH, 14(1), 20–37. https://doi.org/10.32664/j-intech.v14i01.2229

[79] Xu, H., Chen, Y., & Med, A. (2025). 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. *Journal of Technology Informatics and Engineering, 4*(3), 590–612. https://doi.org/10.51903/jtie.v4i3.491

[80] Xu, K., Zhou, H., Zheng, H., Zhu, M., & Xin, Q. (2024). Intelligent classification and personalized recommendation of e-commerce products based on machine learning. Applied and Computational Engineering, 64(1), 143–149. https://doi.org/10.54254/2755-2721/64/20241365

[81] Ye, T., Chang, X., & Zhong, E. (2025). Uncertainty-aware breast ultrasound explanation cards: A visual communication framework for image-based AI diagnostic support using BreastMNIST_224. *International Journal of Graphic Design, 3*(2), 365–380. https://doi.org/10.51903/ijgd.v3i2.3701

[82] Ye, T., Mu, J., & Hunter, J. (2026). Off-policy evaluation and conservative policy selection for slot-level dynamic bidding and ranking on the Open Bandit Dataset (small). *Journal of Technology Informatics and Engineering*, 5(1), 178–199. https://doi.org/10.51903/jtie.v5i1.503

[83] Zhang, B., Rao, H., & Zhao, D. (2024). Evidence-grounded RAG for cloud-native DevOps: Hallucination-resistant AIOps question answering over private operations documents. *Journal of Advanced Computing Systems, 4*(3), 109–125. https://doi.org/10.69987/JACS.2024.40308

[84] Zhang, B., Ren, Y., & Zou, J. (2025). LLM-style explainable e-commerce recommendation cards: A UI/UX design framework for trust-calibrated product recommendation. *International Journal of Graphic Design, 3*(2), 381–396. https://doi.org/10.51903/ijgd.v3i2.3697

[85] Zhang, J. (2025). Adaptive user interface design for volleyball learning apps: Empirical evidence from Google Play reviews and mobile screen analysis. *International Journal of Graphic Design, 3*(1), 175–195. https://doi.org/10.51903/ijgd.v3i1.3618

[86] Zhang, K., Chen, Y., & Qian, A. (2025). Evidence-grounded accounting disclosure review cards: A visual communication framework for LLM-style explanations over SEC financial statements and notes. *International Journal of Graphic Design, 3(*2), 395. https://doi.org/10.51903/ijgd.v3i2.3710

[87] Zhang, R., Wen, Z., Wang, C., Tang, C., Xu, P., & Jiang, Y. (2025). Quality analysis and evaluation prediction of RAG retrieval based on machine learning algorithms [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2511.19481

[88] Zhang, Y., & Zhang, H. (2025a). A therapist-facing session copilot for live counseling support: Reasoning-guided retrieval and ranking from multi-turn counseling dialogues. *Journal of Technology Informatics and Engineering*, 4(2), 464–486. https://doi.org/10.51903/jtie.v4i2.547

[89] Zhang, Y., & Zhang, H. (2025b). Visualizing the right counseling support: Evidence-linked recommendation cards for explainable mental health intake interfaces. *International Journal of Graphic Design,* 3(1), 214–229. https://doi.org/10.51903/ijgd.v3i1.3722

[90] Zhang, Y., & Zhou, Z. (2026). Strategy-aware therapist imitation for emotional support dialogues: A reproducible ESConv study for LLM response control. Advances in Educational Technology and Psychology, 10(2), 92–97. https://doi.org/10.23977/aetp.2026.100213

[91] Zhao, S., Bai, J., & Roberson, D. (2025). Multi-horizon GPU demand forecasting with workload semantics and operational risk curves: An empirical study on Alibaba clusterdata GPU trace. *Journal of Technology Informatics and Engineering,* 4(3), 544–571. https://doi.org/10.51903/jtie.v4i3.498

[92] Zheng, D., & Li, C. (2024). Behavior-level jailbreak resistance via multi-stage refusal + utility preservation. *Journal of Advanced Computing Systems, 4(*1), 83–99. https://doi.org/10.69987/JACS.2024.40107

[93] Zheng, D., Li, C., & Davidson, H. (2023). Continual red-teaming for in-the-wild jailbreaks via online guardrail updates and guardrail distillation. *Journal of Advanced Computing Systems, 3(*2), 35–49. https://doi.org/10.69987/JACS.2023.30203

[94] Zheng, D., Zhang, B., & Geibel, J. (2024). VerifySafe: Toxicity-safe agent responses under adversarial prompts with evidence-based self-verification. *Journal of Advanced Computing Systems, 4*(1), 67–82. https://doi.org/10.69987/JACS.2024.40106

[95] Zhong, Z. S., Chen, J., Zhong, E., & Sun, X. (2025). Evidence-calibrated RAG for unanswerable question answering: Retrieval coverage, abstention calibration, and hallucination-proxy analysis on SQuAD 2.0. *Journal of Technology Informatics and Engineering, 4*(2), 502–520. https://doi.org/10.51903/jtie.v4i2.536

[96] Zhong, Z. S., Li, C., & Rao, H. (2026). Trajectory reliability prediction for generalist AI agents: Tool-use failure analysis and success forecasting on ZClawBench. *Journal of Technology Informatics and Engineering*, 5(1), 341–360. https://doi.org/10.51903/jtie.v5i1.539

[97] Zhong, Z. S., & Ling, S. (2024a). Improved theoretical guarantee for rank aggregation via spectral method. Information and Inference: *A Journal of the IMA*, 13(3), iaae020. https://doi.org/10.1093/imaiai/iaae020

[98] Zhong, Z. S., & Ling, S. (2024b). Uncertainty quantification of spectral estimator and MLE for orthogonal group synchronization. [Preprint]. arXiv. https://arxiv.org/abs/2408.05944

[99] Zhong, Z. S., Pan, X., & Lei, Q. (2025). Bridging domains with approximately shared features. In Y. Li, S. Mandt, S. Agrawal, & E. Khan (Eds.), Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (Vol. 258, pp. 559–567). PMLR. https://proceedings.mlr.press/v258/zhong25a.html

[100] Zhong, Z. S., Wu, Q., & Mi, G. (2025). Uncertainty-aware medical image explanation cards: LLM-generated visual explanations for AI-assisted radiology interfaces. *International Journal of Graphic Design*, 3(2), 415–436. https://doi.org/10.51903/ijgd.v3i2.3616

[101] Zhou, B., Li, C., & Liu, L. (2025). Risk-calibrated patient-facing AI safety cards: A UI/UX design framework for rubric-based medical risk communication. *International Journal of Graphic Design, 3*(2), 365–380. https://doi.org/10.51903/ijgd.v3i2.3696

[102] Zhou, H., & Zhang, K. (2025). News-based uncertainty and macro-market fusion for VIX direction forecasting: Evidence from 2015-2024 FRED panel. *Journal of Technology Informatics and Engineering,* 4(2), 487–501. https://doi.org/10.51903/jtie.v4i2.540

[103] Zhou, S., Chen, Y., & Lee, K. (2026). Accounting-aware evidence-constrained agents for disclosure, settlement, and secondary-market risk monitoring in tokenized RWA infrastructure. *Journal of Technology Informatics and Engineering*, 5(2), 60–74. https://doi.org/10.51903/jtie.v5i2.544