Stout in Psychology and Behavioral Studies (SPBS)

DOI: https://doi.org/10.61424/gwmd8w43
Journal Homepage: https://stoutjournals.org/index.php/SPBS

Controllable without Stereotyping: Cross-Context Stability, Utility Preservation, and Uncertainty in Big-Five-Conditioned LLM Dialogue

Toby Ma
Computer Science, University of Utah, Salt Lake City, UT, USA
Corresponding Author: Toby Ma, **E-mail**: t_ma_uofu@outlook.com

ARTICLE INFORMATION

Submitted: February 23, 2026
Accepted: June 13, 2026
Published: August 01, 2026
Volume: 2
Issue: 1

KEYWORDS

Big Five personality; large language models; persona control; response selection; calibration; selective abstention; personality caricature.

ABSTRACT

Personality-controllable dialogue systems must express intended traits without losing contextual fidelity or collapsing into formulaic cues. We evaluated 100,000 BIG5-CHAT responses representing 10,000 scenarios crossed with five Big Five dimensions and high and low levels. After listwise removal of 64 incomplete scenarios, 9,936 scenarios remained. Scenario-grouped training, validation, and test partitions contained 6,955, 1,490, and 1,491 scenarios. Three locally fitted response-selection strategies were compared: a fixed cue-based persona-prompt proxy, a word-TF-IDF weighted-nearest-neighbor persona retriever, and a character-ngram logistic lightweight adapter. Evaluation combined response-level discrimination, paired high-versus-low ordering, calibration, relation-wise stability, selective abstention, content-fidelity proxies, cue and name masking, length controls, and cross-trait entanglement. On the held-out test set, paired accuracy was 0.7797, 0.9896, and 0.9935 for prompt, retrieval, and adapter, respectively; corresponding AUROC values were 0.7993, 0.9927, and 0.9958. The adapter achieved the lowest Brier score (0.0234), while retrieval had the lowest expected calibration error (0.0066). Worst-relation paired accuracy remained 0.7602, 0.9824, and 0.9889. At 90% coverage, selection risk fell from 0.1892 for the prompt proxy to 0.0003 for retrieval and 0.0000 for the adapter. However, length alone ordered 67.3% of pairs, and cross-trait entanglement was higher for retrieval (0.2712) and adaptation (0.2897) than for prompting (0.0933). The findings demonstrate strong, stable corpus-level persona discrimination but also expose a control-entanglement trade-off; they do not establish demographic stereotype safety or generative-model confidence.

1. Introduction

1.1 Motivation

Personalized dialogue promises responses that fit a user's preferred tone, interaction style, and conversational goals. Personality conditioning is especially attractive because the Big Five framework offers a compact vocabulary for recurring behavioral tendencies: openness, conscientiousness, extraversion, agreeableness, and neuroticism. These dimensions have a substantial psychometric tradition, but they are continuous, multifaceted constructs rather than five sets of catchphrases (McCrae & John, 1992). Modern inventories therefore distinguish broad domains from narrower facets and emphasize reliable measurement across many items (Johnson, 2014; Soto & John, 2017). A dialogue system that repeatedly inserts "creative" words for openness or emphatic social language for extraversion may be easy to classify while remaining psychologically shallow.

Longitudinal personalization also depends on how a system represents and governs preference history. Confidence-gated memory writing, time-aware retrieval, and explicit updating or forgetting address persistence across sessions (Sun et al., 2023), while federated topic-preference learning shows how session interests can be modeled without centralizing raw conversations (Kuo et al., 2025). Other work represents customers through self-supervised behavioral histories or retrieves prior behavior before response generation (Xin, 2026b, 2026f). Review-grounded recommendation adds a further requirement: an explanation should remain faithful to the evidence that motivated the recommendation (Chang et al., 2026). These strands make personality

control more than a one-turn style instruction; a useful system must distinguish a current request from durable preference, retain relevant evidence, and permit revision when the user's behavior changes.

Language nevertheless carries personality-relevant information. Function words, affect terms, syntactic choices, and discourse patterns can correlate with individual differences (Pennebaker & King, 1999), and computational models have long used linguistic cues to recognize or generate personality-marked utterances (Mairesse et al., 2007). This creates a central design tension. Strong persona control requires a detectable distinction between high and low trait expressions, yet maximizing detectability can reward lexical shortcuts, exaggerated affect, and entanglement between traits. Cheng et al. (2023) describe this broader problem as caricature in language-model simulations: a representation may appear vivid precisely because it overconcentrates a few recognizable signals.

BIG5-CHAT makes this tension measurable through matched counterfactuals. Its 100,000 records derive from 10,000 social scenarios, each paired with ten responses spanning five traits and two levels (W. Li, 2024; W. Li et al., 2025). The scenarios originate in SODA, a socially contextualized dialogue resource constructed around relations such as intent, reaction, attribute, effect, want, and need (Kim et al., 2023). Within a trait, the high and low records share the same prompt content, so a controller can be tested on whether it orders the intended alternatives correctly while holding the scenario constant. The design also requires scenario-level splitting: placing sibling variants in both training and test data would turn evaluation into context recognition.

The present study asks four questions. First, how strongly can fixed-response persona prompting, retrieval, and lightweight adaptation discriminate high from low Big Five responses? Second, does control remain stable across the six relation types? Third, do the methods preserve the content of the available response candidates and provide calibrated confidence suitable for abstention? Fourth, is performance sustained after removing narrow trait cues, personal names, and length advantages, and how much does a trait-specific evaluator respond to other conditioned traits? These questions connect control, stability, utility, and uncertainty in a single matched design.

Adjacent applications show why these dimensions should be evaluated jointly rather than as isolated accuracy targets. Retrieval and strategy control have been studied in therapist-facing copilots and emotional-support dialogue (Y. Zhang & H. Zhang, 2025a; Y. Zhang & Zhou, 2026), while evidence-linked counseling cards connect a recommendation to supporting evidence for human review (Y. Zhang & H. Zhang, 2025b). Multilingual humanitarian RAG similarly couples answerability with trust-calibrated presentation (Y. Chen & Xu, 2026). In education, intervention rationales connect predictive decisions to teacher review (Xin, 2026c); in commerce, explainable recommendation cards and uplift policies connect personalization to interpretable action selection (Mu et al., 2023; B. Zhang et al., 2025). Although these domains do not establish Big Five validity, they motivate the same separation adopted here between control strength, retained utility, confidence, and the conditions under which a system should defer.

The experiments evaluate response selectors rather than newly generated dialogue. For every scenario-trait pair, a method scores the two BIG5-CHAT responses and selects the candidate corresponding to the requested level. This design isolates whether conditioning information is recoverable from the corpus while avoiding variation from decoding, serving systems, or closed-model updates. Accordingly, "stereotyping" is operationalized narrowly as personality caricature: dependence on fixed trait lexicons, surface length, or signals shared across nominally distinct Big Five dimensions. The dataset contains no demographic labels, so the study makes no claim about protected-group bias. Its contribution is an empirical map of where high persona separability is reliable and where it may reflect entangled or compressed representations.

1.2 Contributions

The analysis contributes a leakage-controlled benchmark protocol, a three-way comparison of transparent control strategies, and a unified diagnostic suite. Control is measured both row-wise and through paired direction accuracy; calibration is measured with Brier score and expected calibration error; context stability is evaluated by SODA relation; utility is assessed as textual fidelity among fixed candidates; and caricature risk is probed through masking, length matching, and cross-trait transfer. Cluster bootstrap intervals and scenario-level permutation tests preserve the dependence structure of the ten responses associated with each scenario. All primary conclusions follow from the untouched test split.

2. Literature Review

2.1 Personality measurement and linguistic expression

The Five-Factor Model organizes personality descriptors into broad dimensions that recur across instruments and samples (McCrae & John, 1992). The model is useful because it offers shared axes, but broad scores can conceal distinct facets. The IPIP-NEO-120 measures 30 facets beneath the five domains (Johnson, 2014), while the BFI-2 was designed to balance domain

breadth, facet fidelity, and practical length (Soto & John, 2017). These measurement traditions caution against treating a single sentence as a diagnostic personality test. In dialogue research, trait conditioning is better understood as control over an intended expressive style than as evidence that a model possesses a human personality.

Text-based personality work provides both motivation and warning. Pennebaker and King (1999) showed that linguistic style behaves as an individual-difference signal, and Mairesse et al. (2007) demonstrated automatic personality recognition and generation from conversational cues. Such results support text-level control, but they also make cue exploitation likely: the easiest features for a classifier need not reflect the construct's full meaning. Recent psychometric studies of language models reinforce this distinction. Shu et al. (2024) found that apparent psychometric performance can be unreliable under modest changes in testing conditions. TRAIT was subsequently designed to assess distinctness and consistency with stronger psychometric controls (Lee et al., 2025), and Serapio-García et al. (2025) proposed a broad framework for evaluating and shaping model personality. Together, this work motivates cross-context testing rather than a single aggregate score.

The same evaluation principle appears in transfer research outside dialogue: performance must be separated from robustness to topology, tenant, subject, corpus, or dataset shift. Cross-cloud transfer and few-shot cold-start forecasting explicitly test workload changes (He et al., 2024; He et al., 2025); parameter-efficient medical pretraining and subject-independent BCI evaluate transfer under data scarcity or participant change (Mi et al., 2025; Wu, Mi, et al., 2025); and shared-feature learning and cross-dataset activity recognition examine which representations survive domain boundaries (J. Zhang, 2025; Zhong, Pan, et al., 2025). Self-supervised log modeling provides a further example of learning transferable structure without direct task labels (Xin, 2026g). These methodological precedents motivate held-out-context evaluation and relation-specific performance floors. The decision to keep all ten responses for an original_index in one split follows separately from BIG5-CHAT's matched data structure and the need to prevent sibling-scenario leakage.

Robust control also depends on whether an evaluation reflects the intended decision. Offline counterfactual and off-policy studies distinguish a candidate-ranking policy from the data-collection policy that produced its observations (Mu et al., 2025; Ye et al., 2026), while spectral rank-aggregation theory studies stable ordering from noisy pairwise comparisons (Zhong & Ling, 2024a). These are methodological analogies: BIG5-CHAT evaluates a matched ordering—given the same scenario and a requested trait pole, which of two available responses better realizes that pole?

2.2 Persona-conditioned dialogue

Persona dialogue research initially emphasized consistency with biographical facts. Persona-Chat established a widely used setting in which speakers condition responses on short profile sentences (Zhang et al., 2018). Scaling work showed that many personalized agents could be trained from distributed persona data (Mazaré et al., 2018), while meta-learning targeted rapid adaptation to limited personal evidence (Madotto et al., 2019). BoB used a layered transformer design to improve persona-based dialogue with limited personalized data (Song et al., 2021). More recent synthetic-data work has generated personality-oriented dialogue at scale; PSYDIAL, for example, conditions synthetic conversations on personality profiles (Han et al., 2024).

BIG5-CHAT differs from free-form persona profiles by grounding its conditions in human trait data and constructing high/low response variants for the same social scenarios (W. Li et al., 2025). Its source scenarios come from SODA's commonsense contextualization pipeline (Kim et al., 2023). The dataset's steering process is related to controlled decoding with expert and anti-expert distributions, as developed in DExperts (Liu et al., 2021). This matched structure supports a stricter question than general conversational plausibility: given two responses to the same input, does a controller reliably rank the one written for the requested trait level?

2.3 Control through prompting, retrieval, and adaptation

Prompting, retrieval, and parameter-efficient adaptation represent increasingly data-driven forms of control. A persona prompt supplies an explicit instruction or descriptor at inference time. Retrieval adds examples or memories selected for relevance, and retrieval-augmented generation has shown how external evidence can condition a model without embedding all information in its parameters (Lewis et al., 2020). Adaptation changes a learned representation. Adapter layers (Houlsby et al., 2019), prefix tuning (Li & Liang, 2021), low-rank adaptation (Hu et al., 2022), and quantized low-rank adaptation (Dettmers et al., 2023) provide efficient ways to specialize large models.

Retrieval itself spans several decisions that are easy to collapse into one label. Multi-source attribution combines heterogeneous operational signals (Liu et al., 2023); prototype retrieval explains an anomaly through similar normal cases (Xin, 2025a); and retrieval-summary integration links findability to an interpretable synthesis (Y. Li, 2024). Budgeted multi-hop search determines when additional evidence is worth its cost (Su et al., 2025), whereas hybrid query rewriting and listwise reranking change the ordering of already retrieved candidates (Meng et al., 2026). Retrieval-quality prediction then treats the success of that pipeline

as a forecastable outcome rather than an assumption (R. Zhang et al., 2025). These distinctions motivate the present retriever's deliberately narrow description: it is a weighted nearest-neighbor response selector, not a generator with an unspecified retrieval component.

Grounding also requires provenance after retrieval. Evidence-grounded DevOps question answering and bilingual incident triage emphasize answers tied to operational documents (B. Zhang et al., 2024; Liu et al., 2024). Evidence-chain methods add word-level attribution or deterministic provenance explanations (C. Li et al., 2024; Jin, 2025b). Retrieval-grounded log work combines anomaly evidence with deterministic failure narratives, while troubleshooting copilots add command-safety evaluation (Sun et al., 2026; B. Zhang et al., 2026). These studies motivate reporting evidence preservation separately from task accuracy; in the present fixed-candidate study, cosine, overlap, and entity measures quantify candidate fidelity only, not provenance or factual correctness.

Evidence must finally be presented in a form that a reviewer can inspect. Accounting due-diligence work focuses on evidence-aware retrieval (Y. Chen et al., 2025), while scientific-search and disclosure-review cards organize retrieved evidence into reviewable units and visual hierarchies (Jin, 2025c; K. Zhang et al., 2025). Structural/visual consistency evaluation and human-judgment-aligned design critique illustrate separate ways to validate the presentation layer after a model decision has been made (Y. Chen & Li, 2025; Y. Li et al., 2025). Numerical guardrails and scientific claim verification add deterministic checks or evidence ranking before a conclusion is shown (Z. Li et al., 2026; Su, Chen, et al., 2026). Incident cards and cross-regulation legal RAG similarly connect evidence to an actionable human decision (J. Li & Zhou, 2026; Nie et al., 2026), while text-grounded rationale interfaces make design decisions explicit in explainable decision cards (Wu, Meng, et al., 2025). These interface studies do not alter the selector metrics used here, but they clarify how calibrated margins and retrieved exemplars could be exposed in a later interactive system.

The current comparison translates these families into transparent offline probes. The prompt condition is a fixed cue score, retrieval is weighted nearest-neighbor classification over response exemplars, and adaptation is regularized logistic regression over character n-grams. These implementations are intentionally smaller than neural generators: they test how strongly BIG5-CHAT encodes its labels before attributing success to a particular model architecture. The comparison also separates explicit lexicon dependence from corpus-level similarity and learned subword patterns.

2.4 Utility, caricature, and bias boundaries

Dialogue quality cannot be reduced to persona identification. Reference-free evaluation such as USR was developed to capture multiple aspects of dialog quality (Mehri & Eskenazi, 2020), BERTScore aligns contextual token representations with references (Zhang et al., 2020), and MAUVE compares text distributions rather than individual strings (Pillutla et al., 2021). Those metrics are designed for generated text. Because the present experiment selects between two existing candidates, its utility measures are correspondingly narrower: similarity to the intended candidate, overlap with scenario content, preservation of entities, and length distortion. They quantify selection fidelity, not human judgments of helpfulness or naturalness.

Bias benchmarks define a second boundary. StereoSet (Nadeem et al., 2021) and CrowS-Pairs (Nangia et al., 2020) measure social stereotypes associated with protected groups, while Social Bias Frames represents offensiveness and power implications in language (Sap et al., 2020). Dialogue-specific work has also investigated gender bias and mitigation in generation (Dinan et al., 2020). BIG5-CHAT provides neither demographic group labels nor counterfactual protected identities. Consequently, importing those benchmark claims would be invalid. The relevant concern here is whether a controlled personality becomes a caricature of itself or leaks into other trait axes, consistent with the simulation-focused analysis of Cheng et al. (2023).

Controllability should likewise not be conflated with safety. Continual red-teaming addresses guardrail drift under changing jailbreak distributions (Zheng et al., 2023), evidence-based self-verification checks an agent response before release (Zheng et al., 2024), and multi-stage refusal explicitly treats utility preservation as a constraint rather than rewarding blanket refusal (Zheng & Li, 2024). Dark-pattern work treats manipulative interface microcopy as a distinct harm from factual inaccuracy (H. Xu et al., 2025). The present cue-masking and cross-trait tests do not measure any of those harms. They instead adopt the same general discipline of decomposing a broad quality claim into observable failure modes.

Personalization also creates privacy and fairness obligations because user history can become both a predictive feature and a source of exposure. Privacy-robust uplift estimation and privacy-safe aggregate marketing models examine decision quality when persistent identifiers are unavailable (Bai et al., 2026; Bai & Wu, 2026). Counterfactual credit explanations illustrate how an actionable rationale can be audited for fairness rather than treated as decorative text (Xin, 2026d), and privacy/data-integrity risk cards make prompt-injection and data-flow hazards visible to a human overseer (Su, Rao, et al., 2026). BIG5-CHAT lacks user

identifiers, protected attributes, and longitudinal logs, so those properties cannot be tested here. They remain important deployment considerations for systems that turn a corpus-level control signal into a persistent user model.

2.5 Calibration and selective prediction

A useful control system should know when its choice is uncertain. Neural probabilities are often miscalibrated, which motivates post-hoc calibration and metrics such as expected calibration error (Guo et al., 2017). Language-model studies have examined whether verbalized confidence can be calibrated (Tian et al., 2023) and whether models can express uncertainty reliably across prompting conditions (Xiong et al., 2024). Here, uncertainty is more limited and more directly testable: it is the calibrated probability assigned by a response selector. Selective classification then orders matched pairs by confidence and withholds the least certain cases, tracing risk as coverage falls (Geifman & El-Yaniv, 2017). This framework reveals whether confidence separates easy from difficult persona distinctions rather than merely accompanying a high average score.

Calibration, rejection, and reliability prediction have been paired in several decision settings. Credit-risk tiering and probability-of-default workflows combine calibrated probabilities with rejection or distribution-free uncertainty (Y. Chen et al., 2023; Jin et al., 2024). Pairwise resume-job matching applies probability calibration and selective refusal to a ranking problem (Jin, 2025a), while evidence-calibrated RAG treats insufficient retrieval coverage as a reason to abstain (Zhong, Chen, et al., 2025). Conformal alert control and trajectory-reliability forecasting make uncertainty actionable in operational workflows (Xin, 2026e; Zhong et al., 2026), while cost-aware routing illustrates a separate policy choice among analysis paths (C. Li et al., 2026). These precedents motivate reporting both calibration error and risk-coverage curves here; a high paired accuracy alone does not show that the remaining errors receive lower confidence.

Operational forecasting offers a second useful distinction between a point estimate and a risk envelope. Capacity forecasting with calibrated uncertainty, conformal demand envelopes, and multi-horizon operational risk curves all expose the cost of acting on an uncertain forecast (S. Chen et al., 2024; He et al., 2023; Zhao et al., 2025). Visual brief cards for capacity dashboards translate those curves into decisions rather than presenting a single predicted value (Liu et al., 2025). The application domain is different, but the reporting principle provides a useful analogy: the persona selector should expose how error changes as coverage is reduced, not only its full-coverage mean.

Uncertainty is nevertheless estimator-specific. Confidence-calibrated late fusion learns how evidence streams should be combined (Xin, 2025b), medical vision-language classification evaluates uncertainty attached to cross-modal predictions (Lu & Zou, 2026), and spectral-estimator analysis derives uncertainty for a structured synchronization problem (Zhong & Ling, 2024b). None of those estimators is imported into the present experiment. Their methodological relevance is that confidence must be tied to the decision rule that produced it. Accordingly, the probabilities reported below come from positive-scale Platt calibration of each selector's own scores and are never described as a language model's internal confidence.

Communicating uncertainty is also distinct from estimating it. Medical-image explanation cards pair uncertainty-aware predictions with visual explanations, while patient-facing safety-card studies organize risk information through rubric-based frameworks or interface benchmarks (Zhong, Wu, et al., 2025; C. Li et al., 2025; B. Zhou et al., 2025). Breast-ultrasound explanation cards and financial-risk dashboards provide further examples of visual hierarchy for uncertainty and evidence communication (Ye et al., 2025; Z. Li et al., 2025). The current paper reports curves and subgroup metrics rather than testing an interface; these studies suggest candidate elements for future evaluation rather than demonstrating that the proposed elements improve user understanding.

3. Methodology

3.1 Dataset, audit, and analysis unit

BIG5-CHAT contains 100,000 English response records and 14 fields. Each of the 10,000 original_index values identifies one base scenario with five trait dimensions crossed by high and low levels (W. Li, 2024; W. Li et al., 2025). The head, tail, literal, and narrative fields describe the scenario; relation identifies one of six SODA relations; train_input contains the conditioned input; and train_output contains the response. Trait cells were balanced at 10,000 records per level.

The file was read as UTF-8 text with empty strings preserved. Audit rules required exactly ten records, all ten trait-level instructions, nonblank input, and nonblank output for every retained scenario. There were 294 blank input cells and 65 blank or whitespace-only output cells. These omissions affected 64 scenarios, which were excluded listwise. The analysis therefore used 9,936 scenarios, 99,360 response records, and 49,680 matched high/low pairs. This complete-case rule guarantees that every method is evaluated on the same paired material.

Table 1. Dataset audit and leakage-controlled scenario split. All ten trait-level records for a scenario remain in one partition.

Item	Value	Note
Source records	100000	Rows in the BIG5-CHAT CSV
Unique scenarios	10000	Defined by original_index
Complete scenarios used	9936	All 10 trait-level variants present
Excluded incomplete scenarios	64	Listwise exclusion for matched-pair analyses
Training scenarios	6955	Grouped 70% split
Validation scenarios	1490	Half tuning, half calibration
Test scenarios	1491	Untouched until final evaluation
Traits × levels	5 × 2	Balanced within every complete scenario
Relation types	6	xIntent, xReact, xAttr, xEffect, xWant, xNeed

Descriptive statistics were computed before model fitting on the retained records. High-level outputs were longer than low-level outputs for openness, conscientiousness, and extraversion, with smaller differences for agreeableness and neuroticism. This pattern motivated an explicit length-only baseline and a length-matched evaluation. Type-token ratios and the share of unique outputs were also summarized to detect trivial duplication.

Table 2. Descriptive statistics for the 99,360 records retained after complete-scenario filtering.

Trait	Level	n	Words M	Words SD	Sentences M	TTR M	Unique (%)
Openness	Low	9936	38.87	7.40	4.97	0.841	99.57
Openness	High	9936	45.16	5.77	4.65	0.855	99.61
Conscientiousness	Low	9936	41.09	6.87	5.53	0.857	99.63
Conscientiousness	High	9936	45.52	5.55	4.36	0.842	99.64
Extraversion	Low	9936	37.93	8.53	4.80	0.858	99.54
Extraversion	High	9936	43.89	6.40	5.03	0.862	99.61
Agreeableness	Low	9936	43.56	6.14	5.20	0.828	99.36
Agreeableness	High	9936	44.05	6.08	4.72	0.850	99.68
Neuroticism	Low	9936	43.20	6.30	4.78	0.857	99.68
Neuroticism	High	9936	43.33	5.89	5.34	0.805	99.73

3.2 Leakage-controlled splitting

Scenarios, not rows, were assigned to partitions. Exact normalized duplicates in head, train_input, or train_output were treated as links; all scenarios within a connected duplicate component were placed in the same split. A seeded allocation (seed 20250729) produced 6,955 training, 1,490 validation, and 1,491 test scenarios, corresponding to 69,550, 14,900, and 14,910 records. A post-split audit found no normalized output value crossing partitions. The validation scenarios were divided evenly: 745 scenarios for hyperparameter selection and 745 for probability calibration. Test data were not used for either choice.

Within every retained trait pair, the high and low records had identical train_input text. Consequently, paired direction accuracy compares two outputs under a controlled input. Cross-trait inputs sometimes differ because instructions name different traits; cross-trait analyses were therefore treated as sensitivity diagnostics rather than matched causal contrasts.

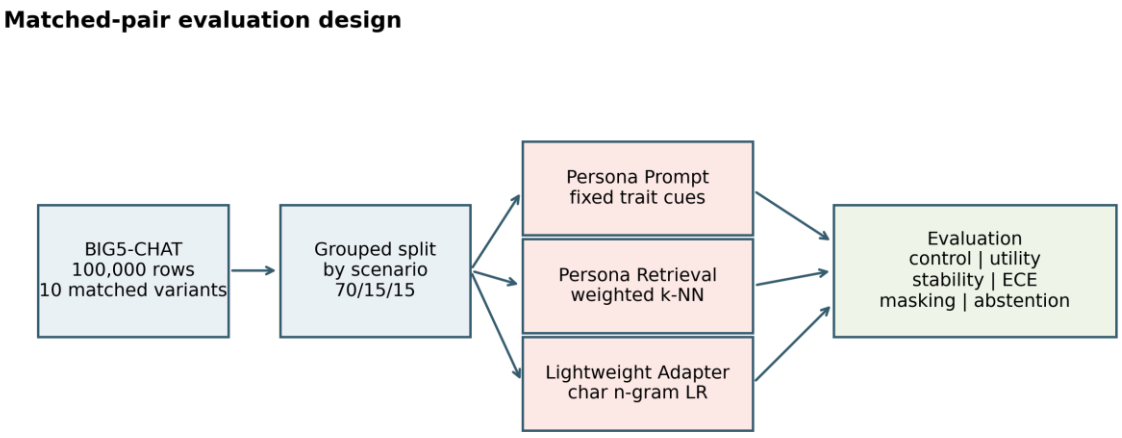

Figure 1. Leakage-controlled matched-pair workflow. The three methods score fixed BIG5-CHAT responses; no dialogue is regenerated.

3.3 Response-selection strategies

Each strategy produced a score for the high level of a specified trait. For a test scenario-trait pair, the response with the larger score was selected for a high request and the response with the smaller score for a low request. Table 3 distinguishes the conditioning signal, fitted component, and selection statistic.

Table 3. Operational comparison of the three fixed-response persona-control strategies.

Method	Conditioning	Fitted component	Selection signal
Persona Prompt	Fixed high/low Big Five descriptor cues	Positive-scale Platt calibration only	Normalized high-minus-low cue count
Persona Retrieval	Target response similarity to stored exemplars	Word TF-IDF; weighted k-NN; k tuned by validation pairs	Similarity-weighted proportion of high-level neighbours
Lightweight Adapter	Learned trait-level response score	Character 3–5 gram TF-IDF logistic model; C tuned	Calibrated high-level decision score

The Persona Prompt proxy used fixed, theory-informed cue sets for the two poles of each trait. Text was lowercased and tokenized; the raw score was the normalized count of high-pole cues minus low-pole cues. The cue lists were fixed before test evaluation. Because this is an offline corpus study, the score represents the lexical component of a conventional persona instruction rather than an API call to a generative model.

The Persona Retrieval strategy represented responses with word TF-IDF features using unigrams and bigrams, Unicode accent stripping, sublinear term frequency, a minimum document frequency of two, and a maximum of 30,000 features. Cosine similarities to training exemplars were converted to a similarity-weighted fraction of high-level neighbors. The number of neighbors was selected independently by trait on the tuning split: 31 for openness, 5 for conscientiousness, 31 for extraversion, 15 for agreeableness, and 15 for neuroticism.

The Lightweight Adapter used character-boundary TF-IDF features with 3- to 5-grams, a minimum document frequency of two, sublinear term frequency, and at most 45,000 features. A separate L2-regularized logistic model was fitted for each trait. The regularization values selected on tuning pairs were 4.0 for openness, 0.25 for conscientiousness, 1.0 for extraversion, 1.0 for agreeableness, and 4.0 for neuroticism. The term "adapter" denotes this compact learned scoring layer; it is not a low-rank update to a generative model.

For all three strategies, raw scores were mapped to high-level probabilities with positive-scale Platt calibration fitted on the calibration half of validation. Constraining the scale to be positive preserved the ordering learned or specified by each method. Thus, calibration changed confidence but not paired direction decisions.

3.4 Outcome measures

Row-level discrimination was evaluated with accuracy, balanced accuracy, macro F1, AUROC, Brier score, and expected calibration error. ECE used ten equal-width probability bins and the absolute difference between mean confidence and observed high-level frequency in each occupied bin, weighted by bin size (Guo et al., 2017). The primary control outcome was paired direction accuracy (PDA), the proportion of scenario-trait pairs in which the high response received a larger high-level score than the low response; ties counted as one half. Mean absolute probability margin summarized pair confidence.

Cross-context stability was measured by computing PDA separately for xIntent, xReact, xAttr, xEffect, xWant, and xNeed. For each method, the worst relation, standard deviation across relations, and maximum-minus-minimum gap were recorded. This formulation uses the context taxonomy present in the data rather than assigning external domains.

Selective performance followed the risk-coverage framework of Geifman and El-Yaniv (2017). Pairs were sorted by absolute probability margin. At coverage c, the least-confident (1 − c) fraction was withheld, and risk was the error proportion among retained pairs. Area under the risk-coverage curve (AURC) and risk at 100%, 90%, 80%, 70%, and 50% coverage were computed. These probabilities describe evaluator confidence in choosing between two candidates; they are not uncertainty estimates produced by a dialogue generator.

3.5 Textual fidelity and caricature diagnostics

Utility preservation was evaluated within each matched pair. For a requested level, each method selected one of the two existing responses; the intended BIG5-CHAT response served as the target candidate. A shared word TF-IDF space measured cosine similarity between the selected response and (a) train_input, (b) the combined scenario description, and (c) the target response. Keyword coverage used non-stopword scenario terms, entity retention checked whether available personx or persony mentions survived, and absolute log length deviation compared selected and target word counts. The oracle row selects the intended candidate and establishes the ceiling for reference similarity and length fidelity. Because both choices come from the same context, these measures quantify candidate fidelity rather than open-ended response quality.

The metric bundle is intentionally rubric-like: it separates the control decision from evidence retention, entity preservation, and length distortion. Rubric-grounded automated scoring provides a related example of making a composite judgment auditable by exposing its criteria and formative rationale (Xin, 2026a). Here, the criteria are computed rather than generated, and no aggregate utility score is formed. Keeping them separate prevents a high similarity on one dimension from concealing a failure on another.

Four diagnostics examined personality caricature. First, all fixed trait-cue tokens were masked and PDA was recomputed. Second, personal names were masked to test whether incidental entity strings affected ordering; a separate name-only character-ngram probe provided a direct leakage check. Third, each method was evaluated only on pairs whose high/low length difference was no larger than the median absolute difference, and a length-only rule selected the longer response. Fourth, a score trained for one trait was applied to the other four traits. Cross-trait entanglement was the mean absolute deviation of AUROC from 0.5 across these off-target applications. A larger value indicates that a nominally trait-specific score responds systematically to other conditioned dimensions, regardless of direction.

3.6 Statistical analysis and reproducibility

Uncertainty respected the scenario-level cluster. Pair-correctness values were first averaged within each original_index. Differences between methods were then evaluated with 5,000 scenario-cluster bootstrap samples for 95% percentile intervals and 10,000 two-sided sign-flip permutations. The three permutation probabilities were adjusted by Holm's sequential procedure (Holm, 1979). Comparisons were planned for adapter minus prompt, adapter minus retrieval, and retrieval minus prompt. All preprocessing, fitting, calibration, tables, and figures were generated by one deterministic Python workflow. Package versions and the dataset checksum were recorded with the run outputs.

4. Results and Discussion

4.1 Overall control and calibration

All three methods exceeded chance, but the size of the advantage varied sharply (Table 4). The Persona Prompt proxy reached 73.23% row accuracy, 0.7993 AUROC, and 77.97% PDA. Persona Retrieval increased these values to 95.83%, 0.9927, and 98.96%. The Lightweight Adapter performed best on the primary control outcomes, with 96.76% row accuracy, 0.9958 AUROC, and 99.35% PDA. The difference between row accuracy and PDA is expected: PDA only requires the two responses within a matched pair to be ordered correctly, whereas row accuracy also requires the calibrated score to cross a fixed 0.5 threshold.

Table 4. Held-out response discrimination, calibration, and paired direction accuracy (PDA). Accuracy, F1, AUROC, and PDA are percentages.

Method	Accuracy	Macro F1	AUROC	Brier	ECE	PDA
Persona Prompt	73.23	72.62	79.93	0.1786	0.0255	77.97
Persona Retrieval	95.83	95.83	99.27	0.0311	0.0066	98.96
Lightweight Adapter	96.76	96.76	99.58	0.0234	0.0069	99.35

Calibration was strong in aggregate for the learned selectors. The adapter had the lowest Brier score (0.0234), indicating the smallest mean squared probability error, while retrieval had the lowest ECE (0.0066 versus 0.0069 for the adapter). The prompt proxy was less accurate and less sharp, with a Brier score of 0.1786 and ECE of 0.0255. Figure 3 shows that validation-only calibration brought all aggregate reliability curves near the diagonal. The aggregate prompt result nevertheless concealed a subgroup weakness: neuroticism ECE was 0.1257. This discrepancy illustrates why one pooled calibration number cannot guarantee reliable confidence for every trait.

Figure 2 and Table 5 show that the learned methods were consistently strong across dimensions. Prompt PDA ranged from 69.79% for neuroticism to 87.49% for agreeableness. Retrieval ranged from 98.26% for extraversion to 99.73% for agreeableness. Adapter PDA ranged from 99.03% for openness to 99.77% for agreeableness. Thus, the prompt proxy's overall score did not describe a uniform control mechanism; its performance depended on how readily each pole was captured by the fixed lexicon. Retrieval and adaptation reduced this variance because they exploited broader distributional features.

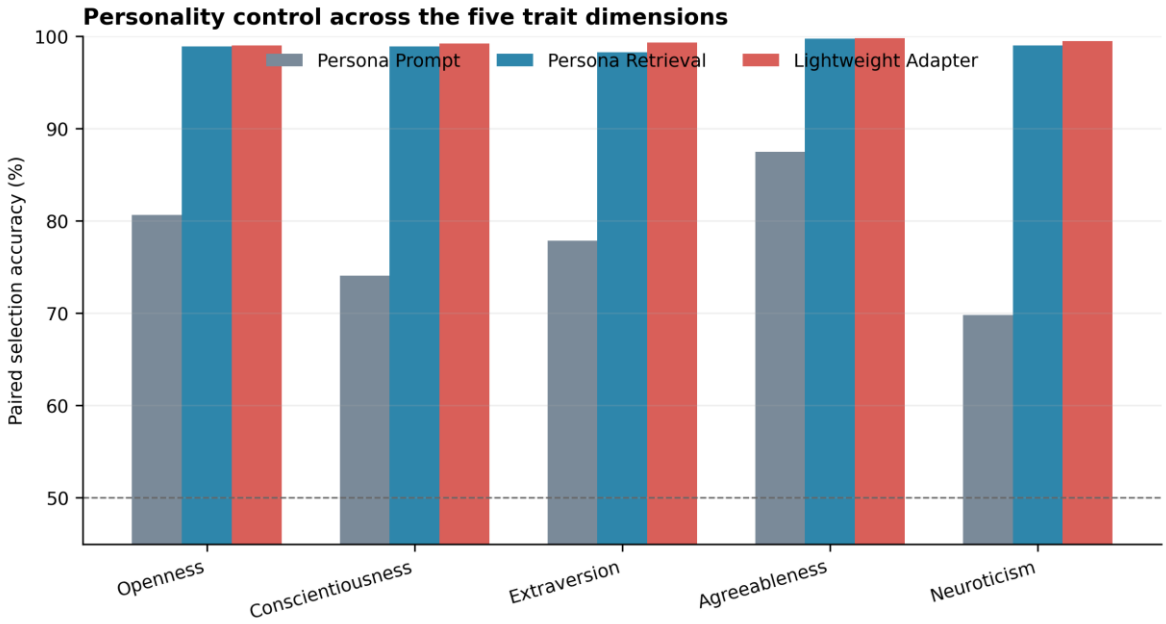

Figure 2. Paired direction accuracy by Big Five dimension. The dashed line marks chance performance for ordering each high/low response pair.

Table 5. Paired direction accuracy (%) by Big Five dimension on the test set.

Trait	Prompt	Retrieval	Adapter
Openness	80.65	98.89	99.03
Conscientiousness	74.08	98.89	99.20
Extraversion	77.87	98.26	99.30
Agreeableness	87.49	99.73	99.77
Neuroticism	69.79	99.03	99.46

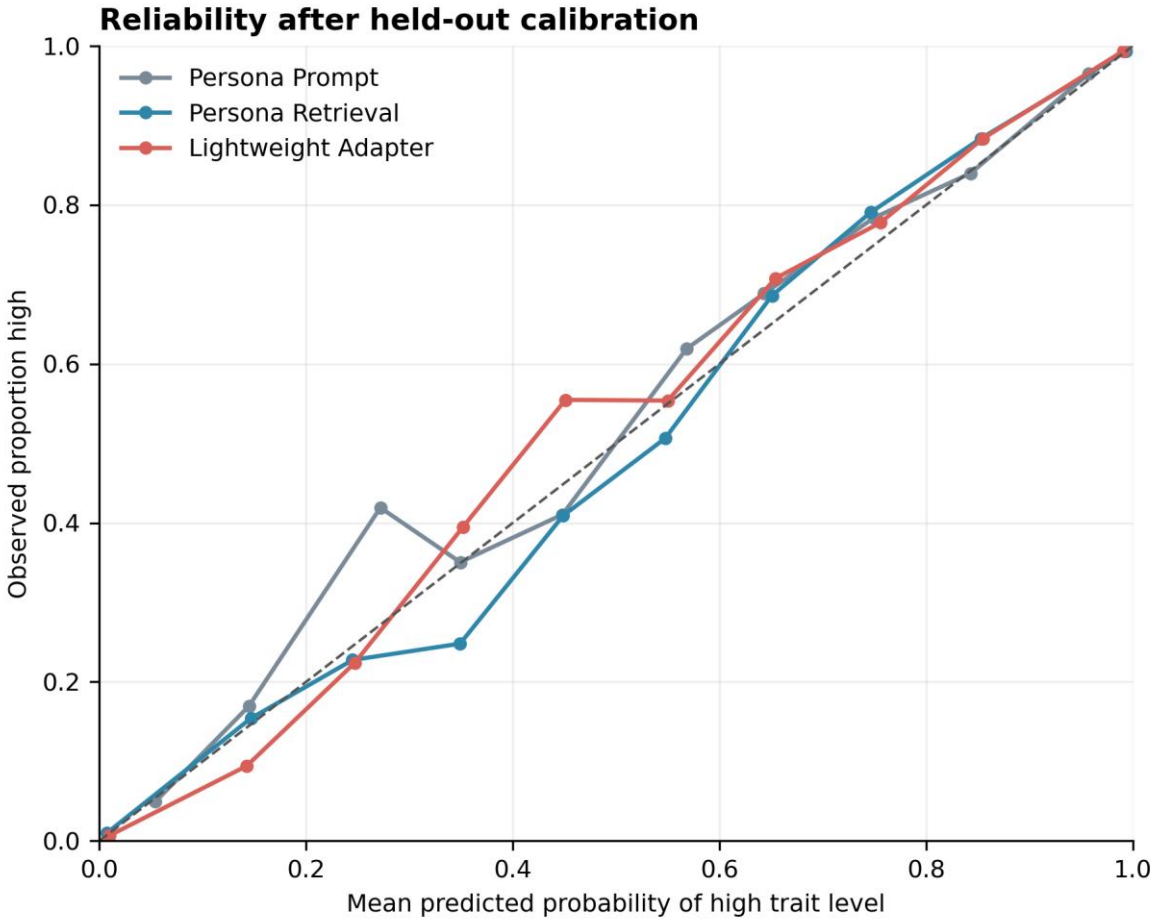

Figure 3. Reliability of held-out high-level probabilities after validation-only positive-scale Platt calibration.

4.2 The diagonal indicates ideal calibration.

The high learned-probe scores should be interpreted as corpus separability, not as evidence that a deployed language model will maintain a personality through an unconstrained conversation. BIG5-CHAT deliberately contrasts high and low responses, and both learned strategies were trained on responses drawn from the same construction process. Near-perfect PDA therefore shows that label-linked patterns generalize to unseen scenarios under a duplicate-safe split. It does not show that free generation would retain factual coherence, avoid repetition, or pass a psychometric inventory across turns.

4.3 Cross-context stability

Relation-wise results support the claim that separability was not confined to one social context (Table 6). Prompt PDA was lowest for xEffect at 76.02% and highest for xNeed at 79.46%, producing a 3.44-point gap and a 1.14-point standard deviation across relations. Retrieval's worst relation was xAttr at 98.24%; its range was 1.25 points and its cross-relation standard deviation was 0.39 points. Adapter PDA was 98.89% on xAttr, 99.26% on xEffect, 99.59% on xIntent, 99.05% on xNeed, 99.67% on xReact, and 99.21% on xWant. Its worst-to-best gap was only 0.79 points.

Table 6. Cross-context paired direction accuracy (%) by SODA relation type. Pair counts aggregate the five traits.

Relation	Pairs	Prompt	Retrieval	Adapter
xIntent	1815	77.88	99.09	99.59
xReact	1840	79.27	99.48	99.67
xAttr	1530	77.61	98.24	98.89
xEffect	1080	76.02	99.07	99.26
xWant	820	77.87	98.84	99.21
xNeed	370	79.46	98.65	99.05

The stability pattern is visible in Figure 5. Relation frequency varied substantially, from 370 test pairs for xNeed to 1,840 for xReact, yet the learned methods showed no collapse in the smaller stratum. This supports cross-context robustness within SODA's relation taxonomy. It remains a bounded conclusion: the six relations are semantic roles within social scenarios, not separate application domains such as counseling, education, or customer support.

4.4 Textual fidelity of selected candidates

The utility analysis asked what content was lost when a selector chose the wrong member of a matched pair. As Table 7 shows, the oracle target response had a mean scenario cosine of 0.1250, keyword coverage of 0.1054, entity mention rate of 0.7919, and zero reference or length error. The adapter nearly reached that ceiling: reference cosine was 0.9968 and absolute log length deviation was 0.0006. Retrieval achieved 0.9937 and 0.0015 on the same measures. Prompt selection was visibly less faithful, with reference cosine 0.8084 and length deviation 0.0349.

Table 7. Textual fidelity of selected fixed candidates. Larger cosine, keyword, and entity values are better; smaller length deviation is better.

Method	Context cosine	Scenario cosine	Reference cosine	Keyword coverage	Entity rate	Length deviation
Persona Prompt	0.0791	0.1232	0.8084	0.1046	0.7767	0.0349
Persona Retrieval	0.0795	0.1249	0.9937	0.1054	0.7915	0.0015
Lightweight Adapter	0.0795	0.1250	0.9968	0.1054	0.7919	0.0006
Oracle target response	0.0795	0.1250	1.0000	0.1054	0.7919	0.0000

Context and scenario cosine values were low in absolute terms for every method, including the oracle. This is a property of the representation and task: the scenario fields often describe an event abstractly, whereas the response realizes a conversational reaction using different vocabulary. The meaningful comparison is therefore with the oracle and across selectors, not against an assumed universal cosine threshold. Adapter matched the oracle's context cosine, scenario cosine, keyword coverage, and entity rate after rounding because its errors were rare. Retrieval was close. Prompt errors more often substituted the opposite-level candidate, which changed reference similarity and length while usually preserving the scenario and named participants.

These measures do not replace human evaluation. USR, BERTScore, and MAUVE address broader aspects of generated-text quality (Mehri & Eskenazi, 2020; Pillutla et al., 2021; Zhang et al., 2020), whereas the present candidates were already written for the same input. The result is best described as utility preservation under response selection: when the learned selector expressed the requested level, it also retained almost all text associated with that intended candidate.

4.5 Cue reliance, length, and cross-trait entanglement

Table 8 separates three kinds of shortcut. Removing the fixed cue lexicons reduced prompt PDA from 77.97% to chance, a 27.97-point loss, because those counts are the prompt proxy's entire signal. The same masking changed retrieval PDA by only 0.10 points and adapter PDA by 0.07 points. The learned methods therefore did not depend on the small theory-derived vocabulary, although they could still use correlated words or constructions outside it.

Table 8. Cue reliance, length and name controls, and cross-trait entanglement. Values are percentages or percentage-point changes.

Method	PDA	Cue drop	Length-match PDA	Length-only PDA	Name drop	Entanglement
Persona Prompt	77.97	27.97	76.47	67.28	0.00	9.33
Persona Retrieval	98.96	0.10	98.91	67.28	0.00	27.12
Lightweight Adapter	99.35	0.07	98.97	67.28	-0.01	28.97

Names contributed no useful label information. Name masking left prompt and retrieval unchanged and improved adapter PDA by 0.013 points, a negligible fluctuation. The separate name-only probe produced 0.5000 AUROC, accuracy, and PDA for every trait. This exact chance result follows from the matched design: high and low variants share their scenario names, and scenario grouping prevents the same pair from crossing the split.

Length was a genuine corpus cue. A rule that always ranked the longer response as high achieved 67.28% PDA. Restricting evaluation to length-matched pairs reduced prompt PDA by 1.51 points, retrieval by 0.05 points, and adapter by 0.38 points. Thus, length explained part of the dataset distinction but not the learned methods' near-perfect ordering. The high conditions for openness, conscientiousness, and extraversion were several words longer on average, so future generators trained on these records should be checked for verbosity changes that masquerade as personality.

The most important caution came from cross-trait entanglement. Mean absolute off-target AUROC deviation from chance was 9.33% for prompt, 27.12% for retrieval, and 28.97% for adapter. A classifier optimized for one Big Five dimension therefore responded strongly to outputs conditioned on other dimensions, especially in the learned representations. This does not mean that the model predicted the same label in every off-target case; the absolute statistic also counts systematic inverse associations. It does show that the dataset's trait signals are not cleanly orthogonal.

The pattern clarifies the title's qualification. Retrieval and adaptation controlled requested labels without relying on the fixed cue list, but their greater separability coincided with stronger cross-trait coupling. In psychometric terms, a detectable factor can still have weak discriminant validity. The result is compatible with concerns about model-personality distinctness and testing instability (Lee et al., 2025; Serapio-García et al., 2025; Shu et al., 2024). It also parallels Cheng et al.'s (2023) warning that simulated personas can become recognizable through compressed caricatures. Figure 6 summarizes this trade-off: learned selectors occupy the high-control, low-fixed-cue-dependence region, yet Table 8 shows that their remaining signal is more entangled across traits.

4.6 Selective abstention

Confidence was useful for deciding when not to select a persona response. At full coverage, risk equaled 22.03% for prompt, 1.04% for retrieval, and 0.65% for adapter (Table 9). Retaining the most confident 90% of pairs reduced prompt risk to 18.92%, retrieval risk to 0.03%, and adapter risk to zero. Retrieval also reached zero observed error at 80% coverage, while the adapter remained at zero through 50%. AURC was 0.0737 for prompt, 0.00024 for retrieval, and 0.00009 for adapter.

Table 9. Selective persona-control risk at retained coverage levels. Risk equals one minus paired direction accuracy; smaller values are better.

Method	AURC	Risk 100%	Risk 90%	Risk 80%	Risk 70%	Risk 50%
Persona Prompt	0.0737	22.03	18.92	15.03	10.14	5.39
Persona Retrieval	0.0002	1.04	0.03	0.00	0.00	0.00
Lightweight Adapter	0.0001	0.65	0.00	0.00	0.00	0.00

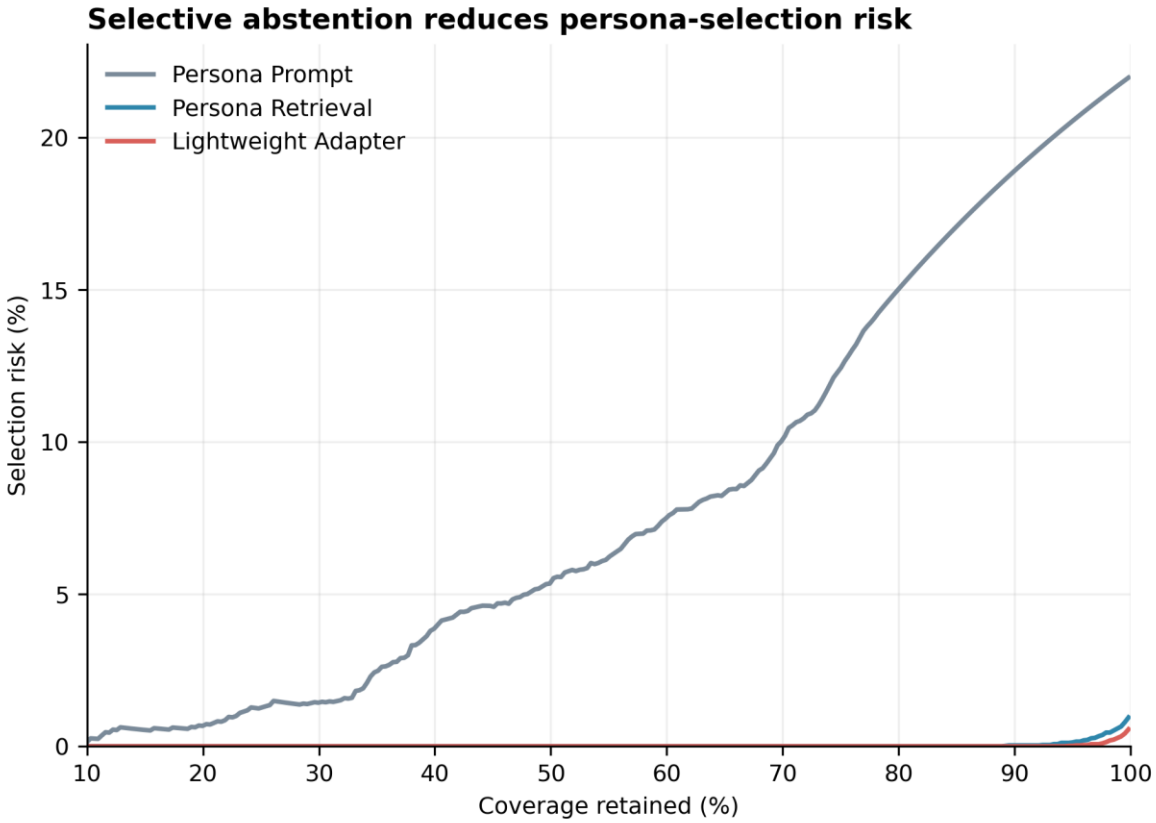

Figure 4. Persona-selection risk as low-confidence matched pairs are withheld. Coverage is the retained fraction of the 7,455 test pairs.

Figure 4 makes the operational difference clear. Prompt risk declined gradually as coverage fell, indicating that its confidence ranking separated some difficult pairs but retained many errors. The learned selectors concentrated nearly all errors at the lowest margins. In a candidate-ranking system, an abstention policy could therefore route uncertain pairs to a neutral style or a second-stage check. These results concern uncertainty in the evaluator's binary choice; they should not be generalized to factual confidence, conversational safety, or a generator's verbalized certainty (Tian et al., 2023; Xiong et al., 2024).

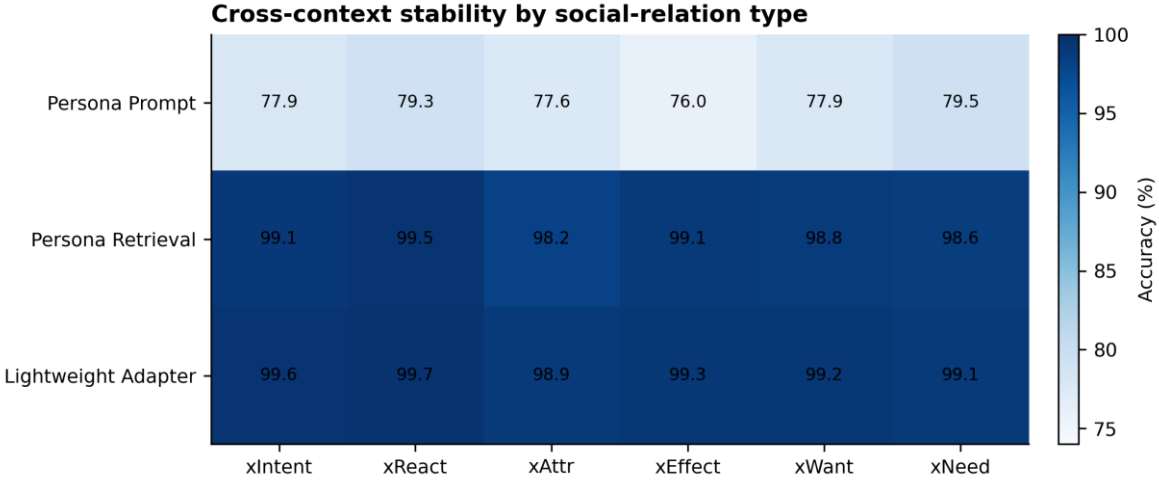

Figure 5. Cross-context paired direction accuracy across the six social-relation types represented in BIG5-CHAT.

4.7 Statistical comparisons and integrated interpretation

Scenario-cluster inference confirmed the large differences (Table 10). Adapter exceeded prompt by 21.37 PDA points, with a 95% cluster-bootstrap interval of [20.54, 22.17] and Holm-adjusted $p < .001$. Retrieval exceeded prompt by 20.99 points, interval

[20.15, 21.84], also p < .001. Adapter exceeded retrieval by 0.39 points, interval [0.17, 0.60], p < .001. The last comparison is statistically detectable because the test contains 7,455 paired decisions, but its practical size is modest relative to the roughly 21-point gains over the cue proxy.

Table 10. Cluster-bootstrap paired-accuracy differences and Holm-adjusted scenario-level sign-flip tests.

Comparison	Difference (pp)	95% CI (pp)	Holm p
Lightweight Adapter minus Persona Prompt	21.37	[20.54, 22.17]	< .001
Lightweight Adapter minus Persona Retrieval	0.39	[0.17, 0.60]	< .001
Persona Retrieval minus Persona Prompt	20.99	[20.15, 21.84]	< .001

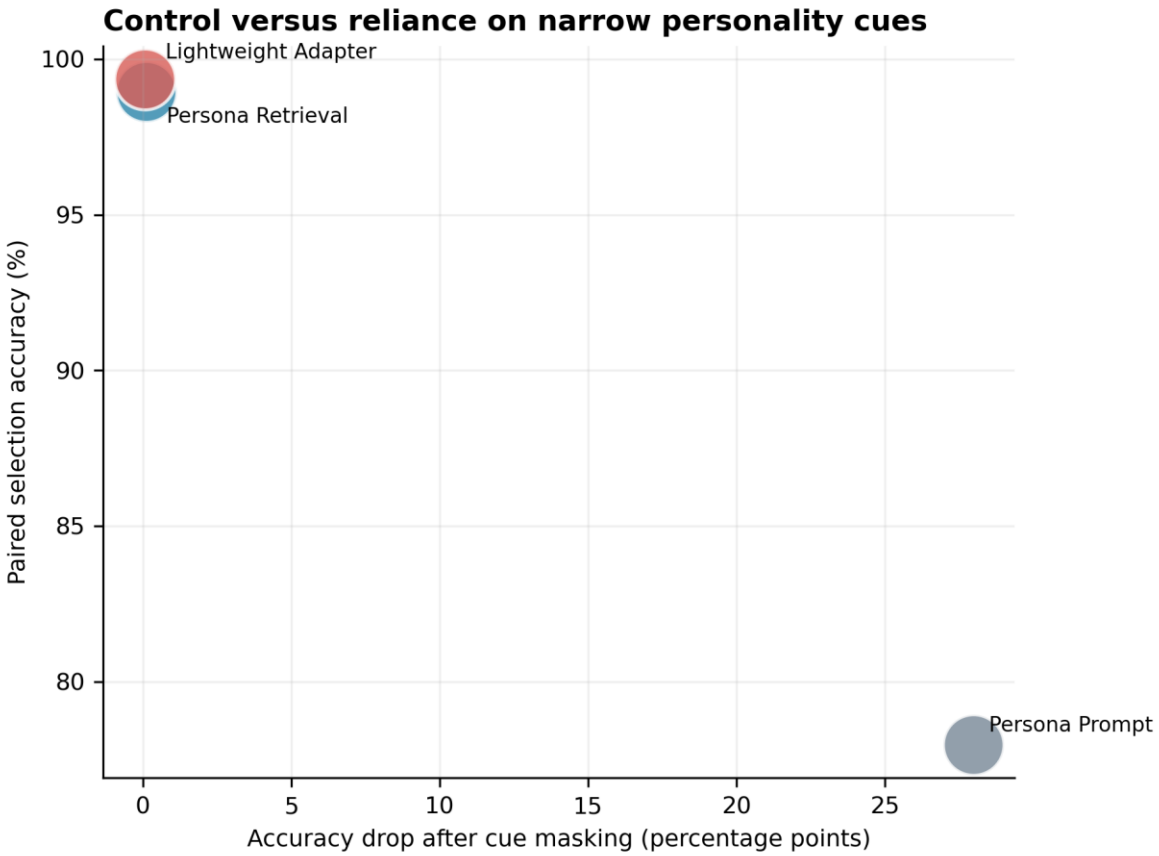

Figure 6. Control versus dependence on the fixed narrow trait-cue lexicons. Bubble area is proportional to mean scenario cosine, a textual fidelity proxy.

Taken together, the results favor retrieval when near-maximum control, slightly lower aggregate ECE, and a simple exemplar mechanism are priorities. The adapter provides the best PDA, AUROC, Brier score, relation floor, and AURC, but only narrowly improves on retrieval and has the highest cross-trait entanglement. The prompt proxy is transparent and less entangled, yet its control is weaker, more trait-dependent, and entirely removed by masking its lexicon. A practical design would therefore combine a learned selector with trait-specific audits, length controls, calibrated abstention, and evaluation on genuinely generated multi-turn conversations before deployment.

5. Conclusions

This study evaluated Big Five persona control on the complete matched portion of BIG5-CHAT. A duplicate-safe scenario split retained 9,936 scenarios and 49,680 high/low pairs. On 1,491 unseen test scenarios, PDA was 77.97% for a fixed cue-based prompt proxy, 98.96% for weighted retrieval, and 99.35% for a character-ngram lightweight adapter. The adapter-prompt and retrieval-prompt advantages were approximately 21 points under scenario-cluster inference, while the adapter's 0.39-point gain over retrieval was statistically reliable but practically small.

Strong aggregate control was accompanied by stable performance across all six SODA relations. The adapter's worst relation remained at 98.89% PDA, and retrieval's at 98.24%. Calibration and abstention added an operational benefit: at 90% coverage, retrieval made two errors per approximately 6,700 retained pairs and the adapter made none in this test sample. Selected-response fidelity was correspondingly high because both learned methods almost always chose the intended fixed candidate. These outcomes establish a robust corpus-level control signal across unseen scenarios.

The diagnostic results place important limits on that success. Length alone ordered 67.28% of pairs, so response verbosity carried label information. Fixed cue masking destroyed the prompt proxy but barely changed retrieval or adaptation, showing that their control was distributed beyond the narrow lexicons. Personal names were uninformative. More consequentially, cross-trait entanglement was 27.12% for retrieval and 28.97% for the adapter, compared with 9.33% for prompt. Near-perfect target discrimination therefore did not imply independent Big Five axes.

The safest conclusion is that BIG5-CHAT supports highly reliable selection between matched high- and low-level responses, especially with retrieval or lightweight adaptation, while also encoding shared stylistic structure across traits. The study does not measure demographic stereotypes, open-ended generation, multi-turn persistence, factual helpfulness, or model-internal uncertainty. Future work should generate new responses under the three conditioning families, obtain blinded human judgments of trait strength and conversational quality, apply established social-bias benchmarks where demographic constructs are actually represented, and test stability across longitudinal interactions. Until then, calibrated abstention, relation-wise reporting, length checks, and cross-trait diagnostics provide a concrete standard for judging whether stronger persona control remains useful rather than merely more recognizable.

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., 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

[2] Bai, J., & Wu, Q. (2026). Privacy-safe marketing mix modeling and budget optimization under identifier loss: A controlled simulation study. *International Journal of Electronic Communication Systems, 6*(1). https://doi.org/10.24042/ijecs.v6i1.30533

[3] Chang, X., Lu, Y., & Zhong, Z. S. (2026). Review-grounded explainable recommendation with faithfulness evaluation on Amazon reviews. *Journal of Electrical Engineering and Computer Science, 11*(1), 9–22. https://doi.org/10.54732/jeecs.v11i1.2

[4] 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

[5] 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

[6] 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

[7] 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

[8] 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

[9] Cheng, M., Piccardi, T., & Yang, D. (2023). CoMPosT: Characterizing and evaluating caricature in LLM simulations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 10853–10875). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.669

[10] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems* (Vol. 36, pp. 10088–10115). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract-Conference.html

[11] Dinan, E., Fan, A., Williams, A., Urbanek, J., Kiela, D., & Weston, J. (2020). Queens are powerful too: Mitigating gender bias in dialogue generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 8173–8188). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.656

[12] Geifman, Y., & El-Yaniv, R. (2017). Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 4885–4894). Curran Associates, Inc. https://papers.nips.cc/paper/7073-selective-classification-for-deep-neural-networks

[13] 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

[14] Han, J.-E., Koh, J.-S., Seo, H.-T., Chang, D.-S., & Sohn, K.-A. (2024). PSYDIAL: Personality-based synthetic dialogue generation using large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 13321–13331). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.1166/

[15] 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

[16] 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

[17] 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

[18] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70. https://www.jstor.org/stable/4615733

[19] Houlsby, N., Giurgiu, A., Jastrzębski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 2790–2799). PMLR. https://proceedings.mlr.press/v97/houlsby19a.html

[20] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[21] 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

[22] 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

[23] 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

[24] Jin, J., Huang, T., & Lu, S. (2024). 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

[25] Johnson, J. A. (2014). Measuring thirty facets of the Five Factor Model with a 120-item public domain inventory: Development of the IPIP-NEO-120. *Journal of Research in Personality, 51*, 78–89. https://doi.org/10.1016/j.jrp.2014.05.003

[26] Kim, H., Hessel, J., Jiang, L., West, P., Lu, X., Yu, Y., Zhou, P., Bras, R. L., Alikhani, M., Kim, G., Sap, M., & Choi, Y. (2023). SODA: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 12930–12949). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.799

[27] 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

[28] Lee, S., Lim, S., Han, S., Oh, G., Chae, H., Chung, J., Kim, M., Kwak, B.-W., Lee, Y., Lee, D., Yeo, J., & Yu, Y. (2025). Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics. In *Findings of the Association for Computational Linguistics: NAACL 2025* (pp. 8412–8452). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.findings-naacl.469

[29] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[30] 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

[31] 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

[32] 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

[33] 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

[34] Li, W. (2024). *BIG5-CHAT* [Data set]. Hugging Face. https://huggingface.co/datasets/wenkai-li/big5_chat

[35] Li, W., Liu, J., Liu, A., Zhou, X., Diab, M. T., & Sap, M. (2025). BIG5-CHAT: Shaping LLM personalities through training on human-grounded data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 20434–20471). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.999

[36] Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 4582–4597). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.353

[37] 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

[38] 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

[39] 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

[40] 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

[41] Liu, A., Sap, M., Lu, X., Swayamdipta, S., Bhagavatula, C., Smith, N. A., & Choi, Y. (2021). DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 6691–6706). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.522

[42] 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

[43] 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

[44] 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

[45] 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

[46] Madotto, A., Lin, Z., Wu, C.-S., & Fung, P. (2019). Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5459). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1542

[47] Mairesse, F., Walker, M. A., Mehl, M. R., & Moore, R. K. (2007). Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research, 30*, 457–500. https://doi.org/10.1613/jair.2349

[48] Mazaré, P.-E., Humeau, S., Raison, M., & Bordes, A. (2018). Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 2775–2779). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1298

[49] McCrae, R. R., & John, O. P. (1992). An introduction to the Five-Factor Model and its applications. *Journal of Personality, 60*(2), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

[50] Mehri, S., & Eskenazi, M. (2020). USR: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 681–707). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.64

[51] 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

[52] Mi, G., Ye, T., & Wood, D. (2025). A lightweight medical foundation model for cross-modal multi-task pretraining and parameter-efficient few-shot transfer on MedMNIST. *Journal of Technology Informatics and Engineering, 4*(3), 572–589. https://doi.org/10.51903/jtie.v4i3.492

[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] Nadeem, M., Bethke, A., & Reddy, S. (2021). StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 5356–5371). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.416

[56] Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020). CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1953–1967). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.154

[57] 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

[58] Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology, 77*(6), 1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296

[59] Pillutla, K., Swayamdipta, S., Zellers, R., Thickstun, J., Welleck, S., Choi, Y., & Harchaoui, Z. (2021). MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems* (Vol. 34, pp. 4816–4828). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2021/hash/260c2432a0eecc28ce03c10dadc078a4-Abstract.html

[60] Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N. A., & Choi, Y. (2020). Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5477–5490). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.486

[61] Serapio-García, G., Safdari, M., Crepy, C., Sun, L., Fitz, S., Romero, P., Abdulhai, M., Faust, A., & Matarić, M. J. (2025). A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence, 7*(12), 1954–1968. https://doi.org/10.1038/s42256-025-01115-6

[62] Shu, B., Zhang, L., Choi, M., Dunagan, L., Logeswaran, L., Lee, M., Card, D., & Jurgens, D. (2024). You don't need a personality test to know these models are unreliable: Assessing the reliability of large language models on psychometric instruments. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 5263–5281). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.295

[63] Song, H., Wang, Y., Zhang, K., Zhang, W.-N., & Liu, T. (2021). BoB: BERT over BERT for training persona-based dialogue models from limited personalized data. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 167–177). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.14

[64] Soto, C. J., & John, O. P. (2017). The next Big Five Inventory (BFI-2): Developing and assessing a hierarchical model with 15 facets to enhance bandwidth, fidelity, and predictive power. *Journal of Personality and Social Psychology, 113*(1), 117–143. https://doi.org/10.1037/pspp0000096

[65] 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

[66] 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

[67] 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

[68] 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

[69] 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

[70] Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., & Manning, C. D. (2023). Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 5433–5442). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.330

[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] Wu, Q., Mi, G., & Wood, D. (2025). Calibration-light subject-independent motor imagery BCI via self-supervised pretraining and conformer. *Journal of Technology Informatics and Engineering, 4*(1), 239–262. https://doi.org/10.51903/jtie.v5i1.493

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

[74] 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

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

[76] Xin, Q. (2026b). Behavior retrieval plus response generation for interpretable conversational personalized recommendation. *International Journal of Electrical, Energy and Power System Engineering, 9*(2), 120–136. https://doi.org/10.31258/ijeepse.9.2.120-136

[77] Xin, Q. (2026c). 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

[78] Xin, Q. (2026d). 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

[79] Xin, Q. (2026e). 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

[80] Xin, Q. (2026f). Self-supervised customer representation learning for segmentation and next-purchase prediction on UCI online retail. *Journal of Information and Technology, 14*(1). https://doi.org/10.32664/j-intech.v14i01.2229

[81] Xin, Q. (2026g). Self-supervised log anomaly detection with LogBERT-style transformers: Full empirical evaluation on a reproducible SynHDFS benchmark. *Journal of Electrical Engineering and Computer Science, 11*(1), 23–35. https://doi.org/10.54732/jeecs.v11i1.3

[82] Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2024). Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=gjeQKFxFpZ

[83] 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

[84] 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

[85] 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

[86] 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

[87] 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

[88] Zhang, B., Sun, X., Liu, G., & Zhou, B. (2026). LLM-style DevOps copilot for cloud-native troubleshooting: Retrieval-augmented runbook generation and command-safety evaluation. *Journal of Technology Informatics and Engineering, 5*(2), 104–118. https://doi.org/10.51903/jtie.v5i2.534

[89] Zhang, J. (2025). From general human activity recognition to volleyball-oriented wearable transfer learning: Cross-dataset evidence from UCI HAR and WISDM for domain adaptation and edge deployment. *Journal of Technology Informatics and Engineering, 4*(1), 263–283. https://doi.org/10.51903/jtie.v4i1.524

[90] 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

[91] 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

[92] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., & Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2204–2213). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1205

[93] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr

[94] 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

[95] 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

[96] 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

[97] 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

[98] 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

[99] 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

[100] 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

[101] 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

[102] 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

[103] 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

[104] 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

[105] Zhong, Z. S., Pan, X., & Lei, Q. (2025). Bridging domains with approximately shared features. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics* (Vol. 258, pp. 559–567). PMLR.

[106] 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

[107] 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