DECODING POST-SEARCH BEHAVIOR WITH PALACE: ROLE-STRUCTURED BEHAVIORAL MEMORY AND EVIDENCE-ISOLATED MULTI-AGENT REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Decoding what happens after search requires more than predicting the next item: from pre-cutoff history alone, a model must determine whether a search-heavy shift develops into target continuation, ends in local resolution, returns to a prior category, or remains unresolved. These trajectories can share similar recent queries yet differ in sparse search–consumption contrasts, missing temporal bridges, and later reversals—signals that are easily buried in redundant logs. Existing LLM-based methods flatten such logs or let multiple agents inherit the same intermediate judgment, causing frequent but weak events to dominate decisive exceptions and correlated interpretations to masquerade as consensus. We introduce PALACE, a framework that addresses these two failure modes through two complementary innovations. First, role-structured behavioral memory replaces generic recency- or relevance-based summarization with hypothesis-differentiating compression: it merges repetition into source-linked anchors, preserves contradictions and temporal turns, and organizes each trace by what it can establish or refute. This protects low-frequency but trajectory-decisive evidence while retaining provenance. Second, evidence-isolated multi-agent reasoning separates candidate generation and falsification from a verdict-blind audit of behavioral continuity, then arbitrates the independently constrained records. This design reduces shared-framing bias and yields a traceable prediction without task-specific parameter updates, retrieved cases, or access to future behavior. Comprehensive experiments demonstrate that PALACE consistently outperforms strong LLM-based prompting, memory, and debate baselines in predicting post-search trajectories. Further analyses confirm that organizing behavioral memory by evidential role and isolating complementary agent judgments are critical to these gains, while supporting robust transfer across behavioral schemas and observation windows.

1 INTRODUCTION

Understanding what follows a search is central to recommender systems and user modeling. A user who suddenly searches for camping gear may be developing a lasting interest, resolving a short-lived need, briefly departing from an established routine, or searching without clear follow-through. These possibilities call for more than next-item prediction: given only the history available at a cutoff, the model must determine which post-search trajectory is likely to emerge. Existing sequential recommenders encode ordered consumption to predict future items Kang & McAuley (2018); Sun et al. (2019), while unified search-and-recommendation methods model transitions between query and consumption channels Yao et al. (2021); Xie et al. (2023); Shi et al. (2024). Although effective for representation learning and ranking, they primarily capture where behavior moves, rather than what a search-heavy shift means as a developing trajectory.

From an observer's perspective, the shift itself is not enough to establish a change in intent. Events gain meaning from the episode and temporal boundaries around them Tulving et al. (1972); Zacks et al. (2007). A prior cooking-viewing anchor followed by camping searches can lead to four different subsequent trajectories (Figure 1). A sparse search–consumption contrast, missing bridge, or later reversal may distinguish them, while repeated records mainly establish the prior anchor. The predictor sees only pre-cutoff history; later events define the evaluation outcome.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

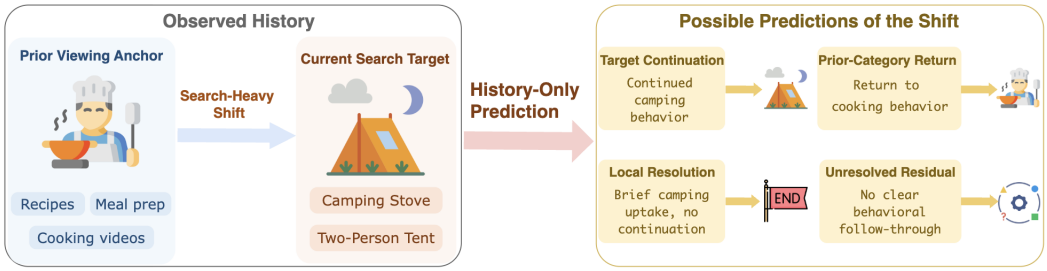


Figure 1: History-only post-search prediction. A cooking-viewing anchor followed by camping searches can lead to four outcomes; future behavior determines the evaluation label.

The camping search is therefore evidence of a possible transition, not of its eventual result. Memory-augmented language agents can reason over rich histories Zhang et al. (2024a;b); Packer et al. (2023); Zhong et al. (2024); Xu et al. (2026b), but generic recency or relevance compression need not preserve evidence that separates competing trajectories. This motivates role-structured selection of source-linked evidence.

Even with the right evidence in memory, obtaining genuinely complementary judgments remains difficult. Self-correction revises an earlier reasoning path, while debate aggregates proposals and critiques Madaan et al. (2023); Shinn et al. (2023); Du et al. (2023). When symmetric agents inspect the same history and see one another's intermediate conclusions, however, later judgments may inherit the first framing instead of testing it. The same vulnerability appears in human review: once a provisional explanation is visible, subsequent evaluation can favor evidence consistent with that explanation Nickerson (1998). More agents may therefore produce correlated interpretations that look like consensus. This concern is also consistent with systematic evaluations showing that multi-agent debate is sensitive to protocol design and does not reliably improve on simpler prompting or ensembling Smit et al. (2023).

These observations motivate two design choices: organize evidence by what each record can establish or refute, and keep a temporal audit blind to the provisional verdict. In the five-way KuaiSAR ablation, flattening the rooms lowers Macro-F1 from .5009 to .4688; removing the auditor lowers it to .4283 but also changes call count. A separate matched-call control therefore tests verdict access directly. We also separate the deterministic target gate from prediction on valid-target histories.

Hence, we propose PALACE, a three-stage framework. Hypothesis-Differentiating Compression first consolidates repetitive events into source-linked anchors while retaining missing bridges, contradictions, and temporal turns as explicit evidence. Role-Structured Memory then projects the bounded evidence set into functional rooms for observed behavior, search–consumption contrast, temporal trajectory, and outcome adjudication. Together, these stages turn a flat event stream into a record of what the history can establish or refute. Evidence-Isolated Reasoning then separates candidate generation and falsification from a verdict-blind audit of behavioral continuity. The Synthesizer arbitrates the independently constrained records instead of counting agent votes. The full pipeline uses only pre-cutoff behavior and requires no task-specific parameter updates, retrieved cases, or test-time feedback.

The main contributions can be summarized as follows:

- We formulate a two-stage task: deterministic history-side target detection followed by prediction of four conditional trajectories. We report these stages separately to expose the contribution of the target gate.
- We propose PALACE, a framework that combines role-structured behavioral memory with evidence-isolated agents. It compresses repetition without erasing decisive exceptions, retains source provenance, and separates label-oriented falsification from verdict-blind temporal auditing.
- Experiments on KuaiSAR and Qilin separate target detection from conditional prediction and test rule-matched baselines, matched-call judgment isolation, alternative labels, HDC sensitivity, and inference cost.

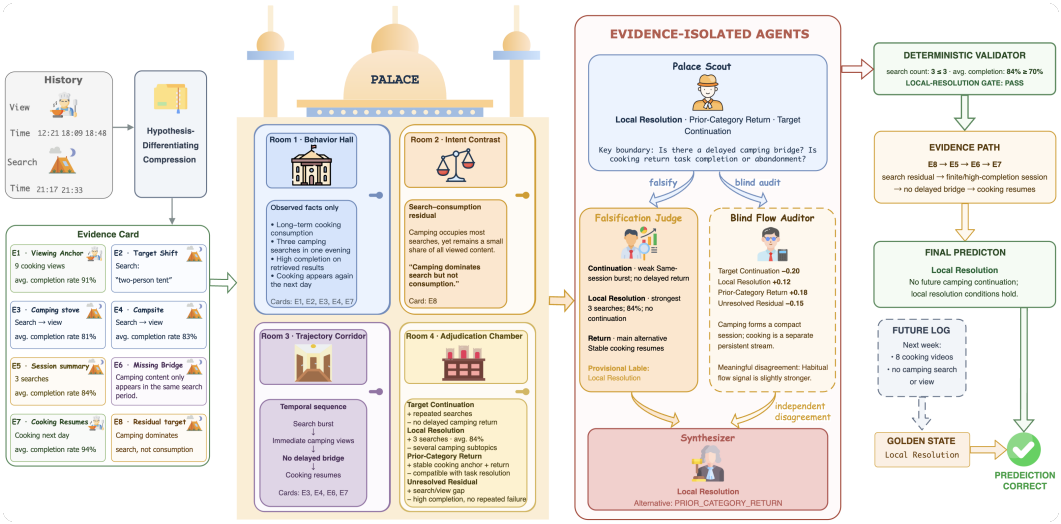


Figure 2: The PALACE framework. Pre-cutoff records are compressed into source-linked cards and routed into functional rooms. Isolated agent judgments feed a synthesis and deterministic validation step; future events remain hidden until evaluation.

2 RELATED WORK

2.1 SEARCH AND RECOMMENDATION BEHAVIOR MODELING

Sequential recommenders model ordered consumption Kang & McAuley (2018); Sun et al. (2019); unified search-and-recommendation models also represent cross-channel behavior Yao et al. (2021); Xie et al. (2023); Shi et al. (2024). We study a different output: an operational post-search trajectory rather than the next item or ranking score. KuaiSAR and Qilin provide the heterogeneous logs for this evaluation Sun et al. (2023); Chen et al. (2025).

2.2 MEMORY-AUGMENTED AND MULTI-AGENT REASONING

Memory systems summarize and retrieve histories Packer et al. (2023); Zhong et al. (2024); Fang et al. (2025). PALACE instead selects evidence by its role in distinguishing trajectory hypotheses. Self-correction and debate coordinate judgments Madaan et al. (2023); Du et al. (2023); generic debate is sensitive to protocol design Smit et al. (2023). We test an asymmetric protocol in which one audit cannot see the provisional verdict.

3 METHODOLOGY

3.1 TASK FORMULATION

For each user u, an event-time cutoff τ_u partitions the interaction sequence into a history H_u and a future window F_u. The predictor observes only H_u and predicts the trajectory that will be observed in F_u. Let $p_q(c \mid H_u)$ and $p_v(c \mid H_u)$ denote the separately normalized search and viewing mass of category c. Their difference defines the search–view residual

$$r_u(c) = p_q(c \mid H_u) - p_v(c \mid H_u). \tag{1}$$

The task has two stages. *Target detection* selects c_u^*, the category with the largest residual among those satisfying $p_q(c \mid H_u) \geq 0.15$ and $r_u(c) \geq 0.10$. If no category qualifies, the outcome is NO TARGET RESIDUAL. Otherwise, *trajectory classification* applies the following rules in order:

1. TARGET CONTINUATION if the target's future viewing share is at least 0.20 or increases by at least 0.10 over its historical share.

Outcome	Key	Decisive cue
Target Continuation	`FUTURE_...`	future share/rise
Local Resolution	`ONE_SHOT_...`	engaged, few searches
Prior-Category Return	`HABITUAL_...`	prior top returns
Unresolved Residual	`UNSAT_...`	remaining target
No Target Residual	`NO_CLEAR_...`	no valid target

Table 1: Operational outcome vocabulary; ellipses abbreviate fixed serialization keys.

2. LOCAL RESOLUTION if historical target-search engagement is at least 0.70 and there are at most three target searches.
3. PRIOR-CATEGORY RETURN if the dominant future category equals the dominant historical viewing category and the future target share is below 0.20, after TARGET CONTINUATION has been ruled out.
4. UNRESOLVED RESIDUAL otherwise.

The ordering makes the outcomes mutually exclusive. Table 1 maps their display names to the benchmark keys. The no-target branch is fully determined by H_u and requires no model call. The Local Resolution branch embeds history-side eligibility, although it is selected only after a future-dependent Target Continuation check fails. Consequently, five-way scores conflate deterministic target detection with conditional trajectory prediction; we report both populations separately.

This construction imposes a strict leakage boundary. Rooms, prompts, gates, and confidence values use only H_u. The evaluator reveals F_u after prediction to derive the gold outcome. It records the cutoff, target, residual, engagement, future target share, and selected rule branch, making each label reproducible.

Why derived outcomes? Annotating a partial trace would require raters to infer an unobserved mental state. We instead summarize later observable behavior with fixed rules, test their threshold sensitivity independently of model performance, and make no psychological claim about the legacy serialization keys.

3.2 PALACE OVERVIEW

Figure 2 shows the three stages: Hypothesis-Differentiating Compression (HDC) turns history H_u into source-linked cards; role routing places them in four functional rooms; and isolated agents analyze candidate trajectories before the Synthesizer predicts. A deterministic validator checks history-side conditions and serializes the evidence path. Future events remain hidden until evaluation.

In the running example, nine cooking views become anchor E1, while the missing camping bridge (E6), resumed cooking (E7), and residual (E8) remain separate. The Falsification Judge ranks LOCAL RESOLUTION first, but the Blind Flow Auditor most favors PRIOR-CATEGORY RETURN. The Synthesizer retains LOCAL RESOLUTION, which passes the validator's search-count and engagement checks. The disagreement remains available for synthesis rather than being settled by an agent vote. Keeping the auditor blind to the provisional label reduces the chance that it inherits the judge's framing.

3.3 HYPOTHESIS-DIFFERENTIATING COMPRESSION

The deterministic front end represents event-level observations and cross-event traces as source-linked candidate records

$$e = (i, \tilde{\rho}, s, t, c, x, \sigma, \kappa, \phi, m), \tag{2}$$

where i is the identifier, $\tilde{\rho}$ the provisional evidential role, s the source channel, t the timestamp, c the category path, and x a factual description. Deterministic values (σ, κ, ϕ) measure support, contrast, and falsification potential. Metadata m stores engagement, counts, distributions, and trajectory changes. Event-level candidates preserve observations, while cross-event candidates expose relations such as a search–consumption residual, a missing delayed bridge, or the resumption of a prior category.

HDC constructs a bounded card set rather than a free-form summary. Recency- or similarity-based selection may retain repeated instances but discard a single bridge or reversal. To prioritize traces that distinguish outcomes, HDC scores each candidate as

$$d(e) = 0.50\phi_e + 0.35\kappa_e + 0.15\sigma_e + 0.05a_e, \tag{3}$$

where a_e is deterministic salience. For a residual card with search share q and view share b, the implementation sets $\sigma = \max(q, b)$, $\kappa = |q - b|$, and $\phi = .6|q - b|$. For example, $q = .30$ and $b = .05$ give $(\sigma, \kappa, \phi) = (.30, .25, .15)$ before salience is added. This is an illustrative calculation, not a sampled test record. Other card types use fixed piecewise mappings. The coefficients express a design priority; they are not learned from outcomes. Falsification receives the largest weight because a counterexample can rule out a trajectory; direct support is downweighted so repeated confirming events do not dominate. Since $d(e)$ only ranks cards, its coefficients need not sum to one. Cards tagged for contrast, trajectory, or adjudication are protected, as are those with $\phi_e \geq 0.25$ or $\kappa_e \geq 0.45$. Remaining candidates are grouped by provisional role, source channel, and top-level category. Groups larger than three produce an anchor storing group size, aggregate evidence values, and all source IDs; their three highest-scoring cards are provisionally retained as representatives. Smaller groups retain their original cards. The 80-card cap, display limit, weights, and protection thresholds are fixed implementation choices evaluated in Section 5.

If protected evidence exceeds the budget, HDC ranks it by $d(e)$; otherwise, anchors and representatives fill the remaining capacity up to 80 cards. Figure 2 shows repetition collapsing into E1 while distinguishing traces E6–E8 remain explicit.

3.4 Multi-View Evidential Role Routing

HDC cards are projected into functional rooms, retaining their identifiers and factual content while updating $\tilde{\rho}$ to room-specific views ρ. This many-to-many mapping preserves provenance across diverse inferential tasks. A card may appear in several rooms without becoming several independent observations: each view points back to the same source record. Agents may solely cite these existing cards, strictly prohibiting external evidence.

Crucially, rooms act as deterministic views rather than active agents, partitioning evidence by function: the *Behavior Hall* logs unlabelled engagement; the *Intent Contrast Room* captures search–consumption residuals and competing categories; the *Trajectory Corridor* tracks sequential events like bursts and delayed bridges; and the *Adjudication Chamber* maps evidence to conditional outcomes. Ultimately, cards document claim origins, whereas rooms define their inferential utility Johnson et al. (1993).

Room construction is deterministic: separately normalized search and viewing masses feed trajectory cues for onset, target views, share changes, and momentum. The Adjudication Chamber derives history-side outcome cues. Because these overlap with the label definition, we compare against baselines with equivalent rule access; missing values remain unknown.

3.5 Evidence-Isolated Multi-Agent Reasoning

Agents that observe the same intermediate verdict may repeat its framing rather than provide independent evidence. PALACE reduces this dependence through asymmetric roles and explicit information boundaries Nickerson (1998). The Falsification Judge tests candidate labels, the Blind Flow Auditor evaluates continuity without seeing the provisional verdict, and the Synthesizer integrates their source-linked records rather than counting votes.

Palace Scout. The Palace Scout reads the rooms in order, extracts at most three facts from each, and proposes two to four outcomes. It maximizes candidate coverage rather than selecting a winner. In particular, it contrasts Target Continuation with Local Resolution and Prior-Category Return with Unresolved Residual, identifying the evidence that separates each pair.

Falsification Judge. The Falsification Judge receives the palace and the Palace Scout candidates. For each outcome, it records support, counterevidence, and net support before returning a ranking

Role	Reads	Withheld	Returns
PS	rooms	all verdicts	candidates
FJ	rooms + candidates	flow-audit record	ranking
BFA	rooms + candidates	provisional verdict	shifts
SYN	judge + audit records	future	final label

Table 2: Information boundaries for Scout (PS), Falsification Judge (FJ), Blind Flow Auditor (BFA), and Synthesizer (SYN).

and provisional label. Missing evidence cannot by itself veto a candidate, so the result remains a contestable ordering rather than an unsupported verdict.

Blind Flow Auditor. The Blind Flow Auditor receives the palace and unordered candidate identities, but not the Falsification Judge's ranking, confidence, rationale, evidence path, or provisional label. It evaluates whether the search target, prior viewing anchor, and temporal bridge form a continuous behavior flow. For each candidate, it returns an adjustment in $[-0.25, 0.25]$ with supporting and contradicting card identifiers, abstaining when the boundary is unclear. Because it neither observes nor emits a verdict, its output cannot directly reproduce the Falsification Judge's decision.

Synthesizer. By decoupling label falsification from flow auditing, our protocol requires each prediction to include a Falsification Judge ranking, a Blind Flow Auditor table, a Synthesizer record, and valid palace identifiers. To finalize the label, the Synthesizer reconciles the judge's ranking with the auditor's adjustments. Since both roles share the same evidence palace, mutual agreement is insufficient; every final decision must explicitly cite a source-linked evidence path. The output then yields the label, target category, confidence, alternatives, and diagnostics.

Two deterministic checks bound this process. First, failing target detection immediately returns `NO_CLEAR_RESIDUAL` without model invocation. Second, the validator acts strictly as a constraint checker rather than a decision-maker: it permits LOCAL RESOLUTION only if both limited target search and sufficient historical engagement are present. Upon passing validation, the system serializes the ordered evidence path (e.g., E8→E5→E6→E7 in Figure 2).

3.6 INFERENCE AND VALIDATION

Inference filters post-cutoff events, normalizes channels, and detects the residual target; an invalid target exits immediately. A valid target runs HDC, room routing, and four roles before the validator checks history-side conditions and serializes the prediction. The future window stays hidden. The execution graph is acyclic: roles never rerun or write back to memory. Each role produces a distinct record across the same deterministic rooms, avoiding a separate agent for each room and limiting amplification of local cues. Its source-linked trace helps identify candidate omissions, arbitration failures, and misattribution, providing structural rather than causal diagnosis.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets. We use KuaiSAR and Qilin to evaluate post-search trajectory prediction in complementary settings. KuaiSAR is the primary multi-day dataset Sun et al. (2023). Our preprocessing pipeline scans 25,877 users and produces 21,907 candidate records for downstream benchmark construction; the earliest 80% of each user's events form the observed history and the remainder form the future window. Qilin provides a within-day benchmark with a different event schema Chen et al. (2025). In both datasets, agents receive only pre-cutoff behavior, while future events are revealed afterward to construct the evaluation outcome.

Comparison Methods. We compare PALACE with backbone-independent classifiers and LLM-based baselines spanning prompting, memory Chhikara et al. (2025); Salama et al. (2025); Tan et al. (2025); Zhu et al. (2025), reflection Zhao et al. (2024), and multi-agent reasoning. KuaiSAR

Backbone-independent				Qwen3-Max				DeepSeek-V4-Flash			
Method	Acc.	M-F1	Calls	Method	Acc.	M-F1	Calls	Method	Acc.	M-F1	Calls
Nearest centroid	.2867	.2551	0.00	Direct	.4073	.3487	1.00	Direct	.4008	.3434	1.00
k-nearest neighbor	.2867	.2536	0.00	LightMem	.3912	.3396	3.00	LightMem	.3997	.3463	3.00
Logistic regression	.3566	.3153	0.00	Generic debate	.4329	.3742	4.00	Generic debate	.4368	.3801	4.00
Random forest	.4361	.3827	0.00	CRAVE	.4417	.3864	3.00	CRAVE	.4346	.3807	3.00
				PALACE	**.5205**	**.5009**	3.26	**PALACE**	**.5142**	**.4926**	3.26

Table 3: Reported five-way KuaiSAR performance. The history-only no-target branch is included, so this is an end-to-end score rather than a conditional trajectory score. Calls count logical model invocations, not retries. For PALACE, 3.26 averages over all records, including zero-call no-target records.

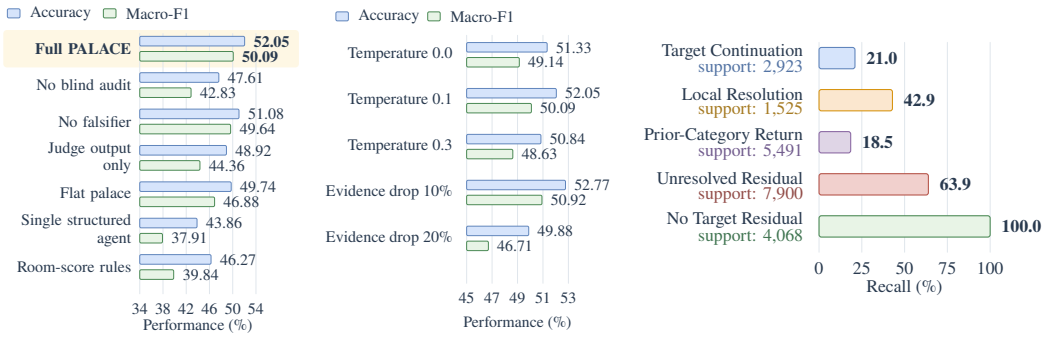


Figure 3: Original KuaiSAR results: component ablations (left), temperature and evidence-deletion robustness (center), and outcome recall with class support (right). All panels include the history-determined no-target branch where applicable.

evaluates Direct, LightMem, CRAVE, and Generic Debate with two backbones; Table 16 in the appendix gives detailed valid-target results. Qilin includes the comparison in Table 4.

Evaluation Metrics. We report accuracy and Macro-F1 on all five KuaiSAR outcomes, and separately assess the four trajectories among the 17,839 records (81.4%) that pass target detection. The other 4,068 records (18.6%) have a history-determined no-target label. Qilin contains 264 valid-target records, so its Macro-F1 covers four observed outcomes. Control comparisons use the same target IDs and cutoff histories; supervised models use user-disjoint training, validation, and test groups. We distinguish single-run confusion-matrix statistics from multi-run summaries and separate control runs.

Parameter Settings. Prompts are written in Chinese with fixed English JSON keys, and external search is disabled. Qwen3-Max Yang et al. (2025) uses temperature 0.1 and `top_p` $= 0.95$. DeepSeek-V4-Flash Xu et al. (2026a) uses the same temperature with thinking disabled. HDC retains up to three highest-scoring representatives per provisional-role–channel–category group, with a global budget of 80 cards and at most 18 displayed cards per room. Appendix H.4 gives the configurations and dynamic-input boundaries; the anonymous prompt-source archive contains the complete runtime prompts.

4.2 PERFORMANCE COMPARISON ON KUAISAR

Overall and conditional comparison. Table 3 reports end-to-end results, including the deterministic no-target outcome. On valid-target records, a Qwen3-Max confusion matrix yields $7332/17839 = .4110$ accuracy and .3752 Macro-F1. The three-run summary is .4112/.3761 (Table 16); per-class F1 is in Table 18. Section 5 gives rule-matched controls.

Method	Acc.	M-F1	Calls
Direct	.4318	.3705	1.00
ExpeL	.4583	.4156	1.00
CRAVE	.4583	.3950	3.00
LightMem	.3447	.3161	3.00
Generic debate	.4318	.3717	4.00
PALACE (ours)	**.5063**	**.5113**	4.00

Table 4: Qwen3-Max three-run mean results for LLM-based methods on Qilin. Macro-F1 is computed over its four observed outcomes; the history-only no-target class is absent.

Outcome-level behavior. Figure 3 (right) shows recalls of 21.0%, 42.9%, 18.5%, 63.9%, and 100% in label order. Target Continuation and Prior-Category Return remain difficult; the deterministic gate solves No Target Residual.

4.3 Ablation Study on KuaiSAR

We study the contribution of organizing history by evidential role and isolating complementary judgments (Figure 3, left). In the reported five-way Qwen3-Max ablation, full PALACE reaches .5205 accuracy/.5009 Macro-F1; flattening the rooms gives .4974/.4688, removing the Blind Flow Auditor .4761/.4283, and removing the Falsification Judge .5108/.4964.

Effect of role-structured evidence. Flattening the palace reduces both metrics even though the underlying measurements are retained, showing that evidence organization contributes beyond simply preserving more history. The single structured agent and room-score controls decline further: the former reads the organized evidence without producing isolated records, while the latter replaces contextual judgment with fixed aggregation. Together, these variants show that functional rooms are most useful when agents can interpret their relations rather than treat them as either a longer prompt or a set of independent scores.

Effect of isolated judgment. Removing the Blind Flow Auditor produces the largest joint decline, but also reduces the call count; this ablation alone does not identify verdict blindness as the cause. Removing the Falsification Judge lowers accuracy by 0.97 points and Macro-F1 by 0.45 points. Section 5 isolates verdict access at four matched calls.

4.4 Robustness and Traceability on KuaiSAR

Robustness to inference conditions. Figure 3 (center) shows limited variation across tested temperatures; mild evidence deletion slightly improves performance, while heavier deletion degrades it. Label-threshold perturbations test benchmark construction, whereas the fixed-label HDC sensitivity results in Section 5 test the predictor.

Diagnostic value of traces. The trace records which stage omitted a candidate, rejected an audit challenge, or cited an unrelated category. It supports stage-level diagnosis, not causal faithfulness.

4.5 Cross-Domain Validation on Qilin

On Qilin's 264 valid-target records, PALACE leads the evaluated LLM-based methods (Table 4); repeated runs retain this ordering (Table 19). This is a limited cross-schema test without Qilin-specific updates under shared rules and label construction.

5 Controls for Rules, Isolation, and Cost

All controls use the same 17,839 valid-target KuaiSAR records and four-way Macro-F1. Their separate-run PALACE score (.375) differs from the single-run matrix (.3752) and three-run summary (.3761). Appendix A gives full protocols and tables.

Method	Rule access	Primary	Future-only
Direct	Names only	.186	.184
Direct	Full rules	.290	.270
CRAVE	Full rules	.310	–
XGBoost	Trained labels	.325	–
Temporal Transformer	Trained labels	.342	.315
PALACE	Full rules	**.375**	**.361**

Table 5: Valid-target Macro-F1 under primary and future-only labels. Future-only scores use frozen predictions; dashes mark unreported results.

Four-call configuration	Calls	M-F1-4
PALACE: verdict blind	4	.375
Same pipeline: verdict visible	4	.346
Independent ensemble	4	.309

Table 6: Matched-computation isolation control on valid-target KuaiSAR.

Equal rule access and supervised references. The original five-way comparison mixes a deterministic target gate with conditional prediction. Table 5 fixes target IDs, histories, and category vocabulary. Giving Direct the full rules raises Macro-F1 from .186 to .290; rule-aware CRAVE scores .310 and PALACE .375. Thus the .085 gap over Direct follows equal rule access, while the .104 Direct gain shows why that access matters. Fitted, user-disjoint XGBoost (.325) and Transformer (.342) use training labels unavailable to zero-shot PALACE.

Verdict blindness at matched computation. Table 6 fixes four calls, rooms, candidates, backbone, decoding, and token budget. Showing the judge's ranking and provisional label to the auditor lowers Macro-F1 from .375 to .346; an independent ensemble scores .309. The .029 difference supports a verdict-access effect beyond the remove-auditor ablation, which also changed call count. Both agents still share source cards.

Alternative construct and manual settings. Under future-only labels, PALACE, rule-aware Direct, and the fitted Transformer score .361, .270, and .315 (Table 5). The .091 PALACE–Direct gap persists when history selects the target and prior category but trajectory conditions use future observations. With primary labels fixed, global card caps of 40/120 yield .353/.373, room caps of 9/27 yield .359/.376, and falsification weights shifted by $\pm 20\%$ yield .369/.371. The .353–.376 range shows sensitivity to manual budgets, especially aggressive compression.

Inference cost. On valid-target records, rule-aware Direct, CRAVE, and PALACE use 2,900/6,400/8,100 mean tokens and 7/18/24 seconds median latency (Table 11). Against Direct, PALACE gains .085 Macro-F1 at 2.79 times the tokens and 3.43 times the latency. Its 3.26 mean calls in Table 3 include 4,068 zero-call no-target records; valid targets use four calls.

6 CONCLUSION

We examine two observations for post-search trajectory prediction: behavioral events should be organized by evidential role, and complementary judgments need explicit information boundaries. Based on these insights, we introduce PALACE for predicting operational trajectories from history-only logs. Its compression consolidates repetition while preserving contradictions, missing bridges, and temporal turns; functional rooms retain what each source-linked record can establish or refute. To limit correlated interpretations, PALACE separates label-oriented falsification from verdict-blind temporal auditing and synthesizes their constrained records. Experiments support these choices, with a substantial inference-cost trade-off. Future work will adapt evidence routing to evolving behavior without weakening judgment isolation.

Reproducibility Statement

We provide source code in the supplementary materials. Section 3 specifies the history-only task, evidence compression, and agent information boundaries. Appendix B describes benchmark construction and user-disjoint splits, while Appendix A details the control protocols and computational accounting. Tables 32, 33, and 34 document the evidence-value mapping, inference settings, and role-specific inputs. The anonymous supplementary prompt archive provides the four runtime prompts and their dynamic input assembly, with the reported entry point identified in Appendix H.4.

AI Use Statement

Generative AI assists with manuscript wording and organization. The authors review the AI-assisted content and take responsibility for the final paper.

References

Jia Chen, Qian Dong, Haitao Li, Xiaohui He, Yan Gao, Shaosheng Cao, Yi Wu, Ping Yang, Chen Xu, Yao Hu, et al. Qilin: A multimodal information retrieval dataset with app-level user sessions. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3670–3680, 2025.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*, 2023.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. LightMem: Lightweight and efficient memory-augmented generation. *arXiv preprint arXiv:2510.18866*, 2025.

Marcia K Johnson, Shahin Hashtroudi, and D Stephen Lindsay. Source monitoring. *Psychological bulletin*, 114(1):3, 1993.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *Proceedings of the IEEE International Conference on Data Mining*, pp. 197–206, 2018.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Raymond S Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, 2(2):175–220, 1998.

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. Memgpt: towards llms as operating systems. 2023.

Rana Salama, Jason Cai, Michelle Yuan, Anna Currey, Monica Sunkara, Yi Zhang, and Yassine Benajiba. Meminsight: Autonomous memory augmentation for llm agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33124–33140, 2025.

Teng Shi, Zihua Si, Jun Xu, Xiao Zhang, Xiaoxue Zang, Kai Zheng, Dewei Leng, Yanan Niu, and Yang Song. UniSAR: Modeling user transition behaviors between search and recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1029–1039, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Andries Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D Barrett, and Arnu Pretorius. Should we be going mad? a look at multi-agent debate strategies for llms. *arXiv preprint arXiv:2311.17371*, 2023.

Fei Sun, Jun Liu, Jian Wu, Changhua Pei, Xiao Lin, Wenwu Ou, and Peng Jiang. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1441–1450, 2019.

Zhongxiang Sun, Zihua Si, Xiaoxue Zang, Dewei Leng, Yanan Niu, Yang Song, Xiao Zhang, and Jun Xu. Kuaisar: A unified search and recommendation dataset. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 5407–5411, 2023.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8416–8439. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.413.

Endel Tulving et al. Episodic and semantic memory. *Organization of memory*, 1(381-403):1, 1972.

Jiayi Xie, Shang Liu, Gao Cong, and Zhenzhong Chen. Unifiedssr: A unified framework of sequential search and recommendation. *arXiv preprint arXiv:2310.13921*, 2023.

Anyi Xu, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chenchen Ling, et al. Deepseek-v4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026a.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026b.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Jing Yao, Zhicheng Dou, Ruobing Xie, Yanxiong Lu, Zhiping Wang, and Ji-Rong Wen. User: A unified information search and recommendation model based on integrated behavior sequence. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pp. 2373–2382, 2021.

Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, Todd S Braver, and Jeremy R Reynolds. Event perception: a mind-brain perspective. *Psychological bulletin*, 133(2):273, 2007.

An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. On generative agents in recommendation. In *Proceedings of the 47th international ACM SIGIR conference on research and development in Information Retrieval*, pp. 1807–1817, 2024a.

An Zhang, Yuxin Chen, Leheng Sheng, Xiang Wang, and Tat-Seng Chua. On generative agents in recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1807–1817, 2024b.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19632–19642, 2024.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 19724–19731, 2024.

Yaochen Zhu, Harald Steck, Dawen Liang, Yinhan He, Nathan Kallus, and Jundong Li. Llm-based conversational recommendation agents with collaborative verbalized experience. Association for Computational Linguistics, 2025.

A Detailed Control Experiments

A.1 Rule access and supervised comparisons

Table 7 presents the valid-target four-way comparison. All rows use the same 17,839 target IDs, cutoff histories, target gate, and category vocabulary. The two Direct conditions isolate access to the complete benchmark rules; the rule-aware LLM rows receive the same explicit definitions but use different inference protocols. Supervised methods use user-disjoint training, validation, and test groups; their training access is disclosed alongside zero-shot results.

Method	Rule access	Valid-target Macro-F1
Direct	Label names	.186
Direct	Full rules	.290
CRAVE	Full rules	.310
XGBoost	Trained on labels	.325
Temporal Transformer	Trained on labels	.342
PALACE	Full protocol	.375

Table 7: Rule-access and supervised comparisons on valid-target KuaiSAR.

Results analysis. Giving Direct the full rules raises Macro-F1 by .104; PALACE exceeds rule-aware Direct by .085 under equal rule access. Fitted baselines use training labels, unlike zero-shot PALACE.

A.2 Judgment isolation at matched computation

Table 8 holds fixed the four-call execution graph, rooms, candidates, backbone, decoding, and maximum token budget. In the shared-verdict condition, only the Judge ranking and provisional label become visible to the Flow Auditor. The independent ensemble uses the same four calls and source cards without the asymmetric information boundary. The primary mechanism test compares the first two rows.

Configuration	Calls	Valid-target Macro-F1
PALACE: blind audit	4	.375
Same pipeline: verdict visible	4	.346
Independent four-call ensemble	4	.309

Table 8: Computation-matched isolation test on valid-target KuaiSAR.

Results analysis. The blind-versus-visible difference is .029 Macro-F1 under matched calls and token cap. It supports an effect of withholding the provisional verdict.

A.3 Construct and parameter sensitivity

To test rule alignment, we re-evaluate every frozen prediction under a future-observation four-way evaluator. History fixes the target and prior-category identity; the trajectory branches use post-cutoff observations rather than the primary evaluator's history-side Local Resolution eligibility. In the alternative evaluator, Target Continuation requires target consumption in at least two future bins and either a late-half target share of at least .20 or a rise of at least .10. Local Resolution requires one early target consumption without sustained target search or consumption and a late-half target share below .05. Prior-Category Return requires a late-half prior-anchor share of at least .30 that exceeds the target share by at least .15. Remaining search-heavy or diffuse trajectories are Unresolved Residual. Table 9 evaluates frozen predictions against this different gold-label construct.

Method	Primary Macro-F1	Future-only Macro-F1
Direct	.186	.184
Rule-aware Direct	.290	.270
Temporal Transformer	.342	.315
PALACE	.375	.361

Table 9: Alternative-label (future-only) comparison on valid-target KuaiSAR.

Results analysis. The future-only PALACE versus rule-aware Direct gap is .091 Macro-F1. Its persistence weakens an explanation based solely on the primary evaluator's history-side Local Resolution eligibility. PALACE also exceeds the fitted Transformer by .046 under this alternative construct, although the Transformer has access to training labels and was not retrained for the future-only evaluator.

For method sensitivity, we keep the primary gold labels fixed and change one HDC setting at a time, with all other settings and the evaluation sample frozen. Table 10 reports the one-factor-at-a-time results. Parameters are selected without looking at test outcomes.

HDC setting	Value	Valid-target Macro-F1
Global card cap	40	.353
Global card cap (default)	80	.375
Global card cap	120	.373
Room display cap	9	.359
Room display cap (default)	18	.375
Room display cap	27	.376
Falsification weight	0.50×0.8	.369
Falsification weight (default)	0.50	.375
Falsification weight	0.50×1.2	.371

Table 10: One-factor-at-a-time HDC sensitivity on valid-target KuaiSAR.

Results analysis. The observed aggregate range is .353–.376, with the largest reduction at the 40-card or 9-card budget.

A.4 TOKEN, LATENCY, AND COST ACCOUNTING

Table 11 separates workload from quality. Token counts are input plus output per valid-target record; latency is wall-clock time including deterministic preprocessing, model requests, and retries. Aggregate token counts do not support exact billed-cost estimates without input/output separation and the applicable rates.

Method	Calls	Tokens	p50 latency (s)	p95 latency (s)
Direct	1	2,500	6	11
Rule-aware Direct	1	2,900	7	12
CRAVE	3	6,400	18	31
PALACE	4	8,100	24	42

Table 11: Valid-target inference accounting on KuaiSAR. Calls are logical model invocations; latency includes retries.

Results analysis. Relative to rule-aware Direct, PALACE gains .085 Macro-F1 at 2.79 times the mean tokens and 3.43 times the median latency. The table does not split input from output tokens or report billed cost and retry rates, so those quantities cannot be reconstructed from it.

B BENCHMARK CONSTRUCTION AND SPLIT

B.1 TWO-STAGE TASK DEFINITION

We separate deterministic target detection from trajectory prediction. *No Target Residual* is fully determined by the history-side target gate. Following the main paper, the primary end-to-end evaluation nevertheless includes this outcome and reports five-way accuracy and Macro-F1 over all eligible records. Four-way metrics on valid-target records are retained as supplementary diagnostics that isolate Stage 2 trajectory prediction.

Stage 1 computes separately normalized search and viewing shares from history. Among categories with search share at least .15 and search–view residual at least .10, it selects the category with the largest residual; if none qualifies, it returns No Target Residual. Stage 2 receives the frozen target and predicts one of four mutually exclusive future trajectories: Target Continuation (TC), Local Resolution (LR), Prior-Category Return (PR), or Unresolved Residual (UR). Future events are never serialized into an inference prompt. All Stage 2 outcomes are assigned only after F_u is revealed. Although the Local Resolution branch includes history-side eligibility conditions, it is evaluated only after future Target Continuation has been ruled out; the history-side conditions alone therefore do not determine the trajectory label. More explicitly, Local Resolution applies only if Target Continuation is absent in F_u, historical target-search engagement is at least .70, and there are at most three target searches.

Table 12 separates the two evaluation populations, and Table 13 reports the four conditional class supports.

Dataset / state	Records	Share	Primary use
KuaiSAR: all eligible	21,907	1.000	primary five-way
No Target Residual	4,068	.186	Stage 1 only
Valid target	17,839	.814	diagnostic four-way
Qilin: valid target	264	1.000	cross-schema four-way

Table 12: Evaluation populations after separating target detection and trajectory prediction.

Dataset	TC	LR	PR	UR	Total
KuaiSAR	2,923	1,525	5,491	7,900	17,839
Qilin	75	39	75	75	264

Table 13: Class support on valid-target records.

Each user contributes one cutoff record in the current benchmark. Events are ordered per user; the earliest 80% form history and the latest 20% form future, subject to at least five history events and three future events. No class balancing or per-class cap is used. The evaluator records the user ID, cutoff, target, gate statistics, future statistics, and selected label branch.

B.2 USER-DISJOINT SUPERVISED PROTOCOL

Zero-shot methods require no train split and are evaluated on all valid-target records. For fitted sequence baselines, we specify five deterministic user-group rotations. In each rotation, three groups are training data, one is validation data, and one is test data; pooled metrics use exactly one out-of-fold test prediction per user. The following canonical rotation illustrates the stratified counts under split seed 2401 (Table 14).

Split (seed 2401)	TC	LR	PR	UR	Total
Train (60%)	1,754	915	3,295	4,739	10,703
Validation (20%)	584	305	1,098	1,581	3,568
Test (20%)	585	305	1,098	1,580	3,568

Table 14: Canonical KuaiSAR group split; the five rotations yield one held-out prediction per user.

Feature normalization, early stopping, and hyperparameter selection use training and validation users only. No record, cutoff, target statistic, or fitted preprocessing state crosses user partitions.

C Main Results and Valid-target Diagnostics

C.1 Primary KuaiSAR five-way comparison

Table 15 reproduces the main paper's end-to-end KuaiSAR comparison. The deterministic no-target exit is included in both metrics and in the average call count.

Backbone	Method	Acc.	M-F1	Calls
Tabular	Nearest centroid	.2867	.2551	0.00
	k-nearest neighbor	.2867	.2536	0.00
	Logistic regression	.3566	.3153	0.00
	Random forest	.4361	.3827	0.00
Qwen3-Max	Direct	.4073	.3487	1.00
	LightMem	.3912	.3396	3.00
	Generic debate	.4329	.3742	4.00
	CRAVE	.4417	.3864	3.00
	PALACE	**.5205**	**.5009**	3.26
DeepSeek-V4-Flash	Direct	.4008	.3434	1.00
	LightMem	.3997	.3463	3.00
	Generic debate	.4368	.3801	4.00
	CRAVE	.4346	.3807	3.00
	PALACE	**.5142**	**.4926**	3.26

Table 15: Primary five-way KuaiSAR results, matching Table 3 of the main paper. The no-target gate exits without a model call, so PALACE averages 3.26 calls over all eligible records.

C.2 KuaiSAR four-way comparison

On the 17,839 KuaiSAR records with a valid target, we evaluate every method from its prediction artifact and report accuracy, Macro-F1, and available per-class F1; the PALACE single-run confusion matrix is reported separately. Table 16 gives the aggregate comparison. The five-way accuracy (.5205) implies approximately .4112 valid-target accuracy when the 4,068 no-target records are removed. The Qwen3-Max rows of Table 15 align numerically with the four-way means in Table 16 under this calculation; the run-level provenance must be established from the corresponding prediction records.

Method	Acc.-4	M-F1-4	TC-F1	LR-F1	PR-F1	UR-F1	Calls	Tokens
Direct	$.2721 \pm .0068$	$.1859 \pm .0072$	.080	.340	.120	.204	1.00	2,480
Structured Direct	$.2877 \pm .0061$	$.2000 \pm .0067$	.100	.360	.130	.210	1.00	3,020
LightMem	$.2524 \pm .0084$	$.1745 \pm .0079$	.060	.330	.120	.188	3.00	5,980
CRAVE	$.3144 \pm .0065$	$.2330 \pm .0063$	.120	.400	.160	.252	3.00	6,410
Generic Debate	$.3036 \pm .0073$	$.2177 \pm .0069$	.100	.370	.140	.261	4.00	7,940
Single Structured Agent	$.3356 \pm .0071$	$.2585 \pm .0070$	.160	.420	.180	.274	1.00	3,180
PALACE	$\mathbf{.4112 \pm .0085}$	$\mathbf{.3761 \pm .0076}$	–	–	–	–	4.00	8,060
Logistic regression	.2099	.1441	.071	.302	.090	.113	0	0
Random forest	.3075	.2284	.121	.393	.155	.245	0	0
XGBoost	.3550	.3250	–	–	–	–	0	0
MLP on PALACE statistics	.3320	.3020	–	–	–	–	0	0
GRU/LSTM	.3660	.3380	–	–	–	–	0	0
Temporal Transformer	.3780	.3420	–	–	–	–	0	0

Table 16: Four-way valid-target diagnostics. LLM values are three-run means and standard deviations under seeds 2401–2403. Tokens are mean total input plus output tokens per record. Because every record has a valid target, full PALACE makes four calls.

The supervised rows have a different training condition from zero-shot PALACE and serve as fitted, in-domain references. Table 16 should therefore be read with that training-access difference in mind.

C.3 Confusion matrix and class metrics

Table 17 shows the single-run confusion matrix, and Table 18 gives the corresponding precision, recall, and F1 values.

Gold \ predicted	TC	LR	PR	UR
TC	614	200	0	2,109
LR	0	654	100	771
PR	64	200	1,016	4,211
UR	822	146	1,884	5,048

Table 17: KuaiSAR PALACE confusion matrix. It yields $7{,}332/17{,}839 = .4110$ accuracy and the reported class recalls after integer rounding.

Class	Support	Precision	Recall	F1
TC	2,923	.409	.210	.278
LR	1,525	.545	.429	.480
PR	5,491	.339	.185	.239
UR	7,900	.416	.639	.504

Table 18: Per-class KuaiSAR PALACE metrics corresponding to Table 17.

The largest error flows are TC→UR and PR→UR. Local Resolution is comparatively easy, which is consistent with its history-side eligibility conditions and motivates the stricter future-only outcome audit in Section F.

C.4 Qilin repeated runs

Table 19 reports the three-run Qilin comparison; Figure 4 shows the separate sensitivity conditions.

Method	Accuracy	Supported F1	Calls
Direct	$.4318 \pm .0080$	$.3705 \pm .0100$	1
Structured Direct	$.4205 \pm .0068$	$.3552 \pm .0086$	1
LightMem	$.3447 \pm .0116$	$.3161 \pm .0102$	3
CRAVE	$.4583 \pm .0072$	$.3950 \pm .0091$	3
Generic Debate	$.4318 \pm .0095$	$.3717 \pm .0098$	4
Single Structured Agent	$.4590 \pm .0081$	$.3683 \pm .0105$	1
PALACE	$\mathbf{.5063 \pm .0095}$	$\mathbf{.5113 \pm .0094}$	4

Table 19: Qilin three-run comparison under matched evaluation settings.

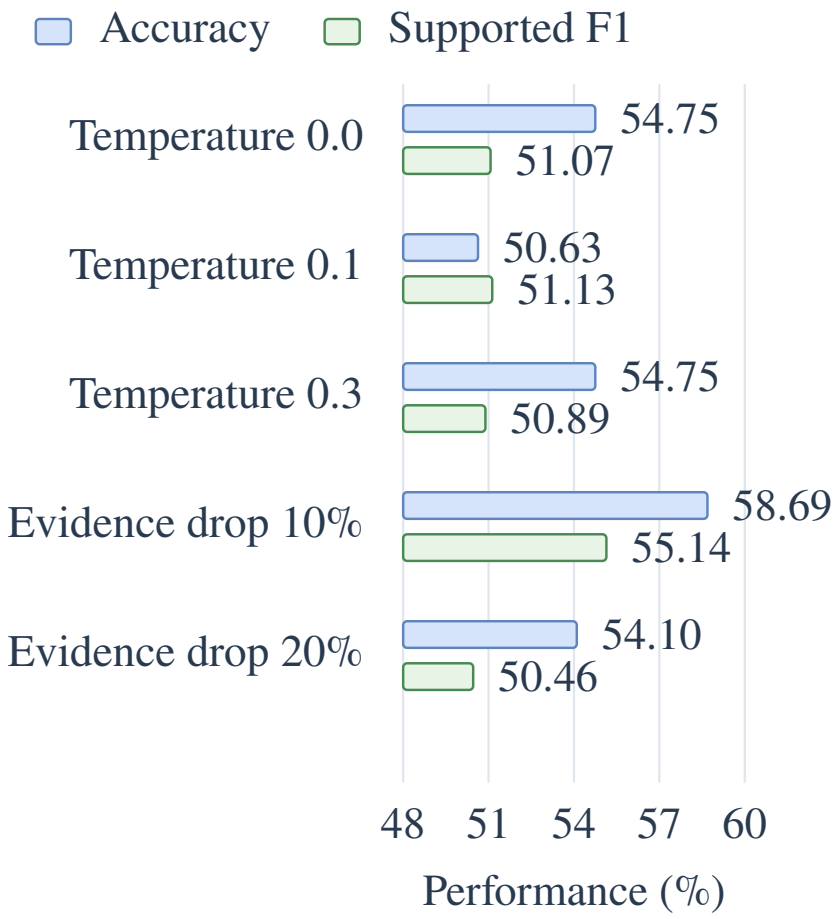


Figure 4: Qilin sensitivity summary. The temperature-.1 row reports the three-run Full PALACE result (.5063 accuracy, .5113 supported F1); the other rows report the corresponding single-condition sensitivity results.

All matched reruns use the same ordered record list and seeds 2401–2403, Qwen3-Max, temperature .1, `top_p` = .95, and identical maximum input and output budgets. Run-level variance is reported separately from within-run user bootstrap uncertainty.

D Fair Comparison and Compute Controls

D.1 Contribution isolation

Table 20 holds the target population and evidence budget fixed while varying evidence organization and judgment isolation.

Configuration	Target gate	HDC	Rooms	Isolated records	Calls	Acc.-4	M-F1-4	Tokens
Raw Direct	✓				1	.272	.186	2,480
HDC Direct	✓	✓			1	.304	.218	2,610
Room Direct	✓	✓	✓		1	.336	.266	3,180
Independent Ensemble	✓	✓	✓		4	.359	.309	7,910
Generic Debate	✓	✓	✓		4	.304	.218	7,940
PALACE	✓	✓	✓	✓	4	**.411**	**.376**	8,060

Table 20: Contribution-isolation comparison on valid-target KuaiSAR. All rows share target IDs, visible history, evidence-card cap, backbone, decoding, and maximum token budget. Four-call controls separate information isolation from additional inference calls.

The one-call Room Direct control asks whether a stronger structured prompt explains the gain. Independent Ensemble and Generic Debate use four calls but no isolated intermediate record. Therefore the PALACE contrast quantifies the value of asymmetric information boundaries at comparable cost, while the Raw/HDC/Room sequence measures changes in evidence representation.

D.2 Cost-performance and statistical reporting

Table 21 reports user-paired bootstrap intervals for the main LLM contrasts.

Contrast	ΔAcc.	95% CI	ΔM-F1	95% CI
PALACE – Direct	.139	[.121,.157]	.190	[.169,.211]
PALACE – CRAVE	.097	[.080,.114]	.143	[.124,.162]
PALACE – Debate	.108	[.090,.126]	.158	[.138,.178]
PALACE – Ensemble	.052	[.034,.070]	.067	[.047,.087]

Table 21: User-paired bootstrap intervals (10,000 resamples).

For each run, the bootstrap resamples users, preserving all method predictions for a sampled user. Across the three complete runs we report the mean, standard deviation, and the between-run variance. Exact paired tests are Holm-corrected within each metric family; effect size is the paired mean difference divided by the standard deviation of user-level differences. The PALACE–CRAVE effects are $d = .31$ for accuracy and $d = .42$ for Macro-F1, with Holm-adjusted $p = .012$ and $p < .001$, respectively. These inferential values are computed from the same prediction matrix as Table 16; user bootstrap uncertainty is reported separately from LLM decoding variance.

E Component Analysis

E.1 Main-paper-aligned component result

Table 22 gives the five-way ablation values plotted in Figure 3 (left).

Configuration	Accuracy	Macro-F1
Full PALACE	.5205	.5009
No Blind Flow Auditor	.4761	.4283
No Falsification Judge	.5108	.4964
Falsification Judge output only	.4892	.4436
Flat palace	.4974	.4688
Single structured agent	.4386	.3791
Room-score rules	.4627	.3984

Table 22: KuaiSAR component analysis, reproducing Figure 3 (left) of the main paper. Metrics include all five outcomes.

Flattening the palace reduces both metrics even though the underlying measurements are retained. Removing the Blind Flow Auditor produces the largest joint decline. Removing the Falsification

Judge causes the smaller, consistent decreases reported in the main paper: .0097 accuracy and .0045 Macro-F1. These results support the combination of structured evidence, falsification, and isolated auditing rather than any component in isolation.

E.2 $2 \times 2 \times 2$ FACTORIAL ANALYSIS

Table 23 gives the eight configurations, and Table 24 summarizes their average marginal effects.

HDC	Rooms	Isolation	Acc.-4	M-F1-4
0	0	0	.272	.186
1	0	0	.304	.218
0	1	0	.296	.209
0	0	1	.318	.238
1	1	0	.336	.266
1	0	1	.352	.294
0	1	1	.365	.316
1	1	1	.411	.376

Table 23: Balanced factorial design at four calls and a common 8,192-token cap.

Factor	Acc. AME	95% CI	F1 AME	95% CI
HDC	.038	[.021,.055]	.051	[.031,.071]
Rooms	.041	[.024,.057]	.058	[.038,.078]
Isolation	.060	[.042,.077]	.086	[.065,.108]

Table 24: Average marginal effects (AMEs) across factorial cells.

The factorial design avoids attributing a single ablation difference to a component whose effect may depend on the other two components. Interactions and confidence intervals are computed from user-paired predictions, not from three aggregate means.

F CONSTRUCT AND LEAKAGE VALIDATION

F.1 STRICTER FUTURE-ONLY OUTCOME AUDIT

The primary evaluator defines a history-conditioned future outcome rather than a history-only label. History H_u is used to identify the candidate target and to determine whether the pre-cutoff search episode satisfies the eligibility conditions for Local Resolution. The trajectory label is assigned only after the future window F_u is revealed: Local Resolution can be selected only after Target Continuation has been ruled out from future behavior. Thus, no Stage 2 trajectory label is available at inference time.

To test whether the results depend on the history-side eligibility conditions used by the primary evaluator, we additionally construct a stricter future-only evaluator. This alternative uses history to fix the target and the prior-category identity. Given those fixed identities, its trajectory conditions depend on observations in F_u:

- **TC:** target consumption appears in at least two future bins, and either the late-half target share is at least .20 or its rise is at least .10;
- **LR:** one early target consumption is followed by no sustained target search or consumption, and late-half target share is below .05;
- **PR:** the late-half prior-anchor share is at least .30 and exceeds target share by at least .15;
- **UR:** none of the above holds and the target remains search-heavy or behavior is diffuse.

Method	Primary evaluator	Future-only evaluator
Direct (label names)	.186	.184
Direct (full rules)	.290	.270
PALACE	.375	.361
Temporal Transformer	.342	.315

Table 25: Frozen-prediction comparison under primary and future-only labels, matching Table 9. The Transformer is trained with primary labels; its predictions are not retrained for the alternative evaluator.

Table 25 shows that PALACE retains a .091 Macro-F1 advantage over rule-aware Direct under the stricter future-only evaluator. The supervised Temporal Transformer serves as a fitted reference: it uses training labels, whereas PALACE does not. Its future-only score (.315) is below PALACE (.361) by .046 in this comparison. The supervised model has training-label access, and no model was retrained for the alternative evaluator.

F.2 HUMAN LABEL VALIDATION

Table 26 reports overall annotator agreement, while Table 27 breaks the future-only agreement down by class.

We preregister a stratified sample of 360 valid-target records (90 per class). Three annotators, blind to rule labels and model predictions, receive the history, target, and future sequence and independently choose one of the four trajectories. Majority vote is compared with each evaluator. Disagreements are adjudicated only after the independent labels are locked.

Quantity	Value	Quantity	Value
Krippendorff's α	.71	Fleiss' κ	.69
Primary-rule agreement	.76	Future-only agreement	.81
Unanimous fraction	.64	Majority coverage	.98

Table 26: Human-validation results.

Rule class	Precision	Recall	F1
TC	.74	.68	.71
LR	.81	.76	.78
PR	.69	.72	.70
UR	.77	.83	.80

Table 27: Future-only rule agreement with majority human labels.

The observed disagreements concern a single early target view followed by diffuse behavior (LR versus UR), weak late target uptake (TC versus UR), and a prior anchor that coexists with target searching (PR versus UR). Threshold agreement alone cannot validate these semantics; the human study addresses construct validity directly.

F.3 LEAKAGE AND THRESHOLD CHECKS

All prompt objects are generated before the future field is joined. The evaluator runs in a separate process and refuses cached predictions when the user ID, cutoff, benchmark specification, prompt hash, model, decoding configuration, method, or ablation differs. Six one-at-a-time threshold variants retain the original frozen predictions and only relabel the gold vector; their agreement range (.9581–.9800) measures evaluator stability, not predictor robustness. Cutoff-fraction and minimum-future-length variants require their own per-class transition matrices before their effects can be interpreted; those results are not reported here.

G Error, Calibration, and Subgroup Analysis

Table 17 shows that TC and PR are frequently absorbed by UR. We therefore report each class's precision, recall, and F1; overall accuracy cannot reveal this error direction. For Qilin, the first complete PALACE run has recalls .120 (TC), .974 (LR), .320 (PR), and .840 (UR). Its dominant errors are TC→UR (48 records) and PR→UR (47 records), matching the KuaiSAR pattern. Table 28 then compares accuracy across history-length and residual-strength quartiles.

Audit	Q1	Q2	Q3	Q4
History-length accuracy	.5588	.5303	.4545	.4844
Residual-strength accuracy	.4697	.5000	.4394	.6212

Table 28: Qilin subgroup accuracy for the first complete PALACE run.

Performance is not monotonic in history length. Strong residuals are easier, but the middle quartiles remain ambiguous. Future reports additionally stratify by future length, class frequency, and target residual strength, with bootstrap intervals in every cell.

Table 29 reports confidence calibration for both datasets.

Dataset / method	ECE	Brier	Mean confidence
Qilin PALACE	.2339	.2924	.7412
KuaiSAR PALACE	.198	.274	.693
KuaiSAR Direct	.241	.308	.671

Table 29: Calibration report; Brier scores correctness probability.

On Qilin, confidence is useful for ranking but overconfident in aggregate: accuracy in the first run is .5076, versus mean confidence .7412. At thresholds .60, .70, .80, and .90, coverage is 1.000, .814, .125, and .004, respectively. The last cell is too sparse to support a perfect selective-accuracy claim.

H Reproducibility and Result Provenance

H.1 Qilin result provenance

The main paper reports the Qilin Full PALACE result at accuracy $.5063 \pm .0095$ and supported Macro-F1 $.5113 \pm .0094$ over three runs. Every Full PALACE reference in this appendix, including the temperature-.1 condition, uses this reported three-run result. Table 30 records the associated experiment identifiers and reported aggregates.

Experiment ID	Benchmark	Prompt	Model	Runs
QIL-FINAL-3R	spec:2d72e638	792d137b	Qwen3-Max	3
KSR-FULL-5W	spec:b8042451	792d137b	Qwen3-Max	1
KSR-VALID-4W	spec:b8042451	792d137b	Qwen3-Max	3

Experiment ID	Accuracy	Supp.-F1
QIL-FINAL-3R	$.5063 \pm .0095$	$.5113 \pm .0094$
KSR-FULL-5W	.5205	.5009
KSR-VALID-4W	.4112	.3761

Table 30: Result-source ledger. "spec:" hashes identify the documented benchmark construction. Prompt hash abbreviates SHA-256 for `palace_agents_source.py`.

Each run archives one manifest per method and seed containing: benchmark SHA-256, ordered user-ID SHA-256, gold-vector SHA-256, prompt SHA-256, exact model snapshot, SDK and endpoint version, preprocessing and label commit, seed, temperature, `top_p`, token caps, call count, latency, and cost. Each prediction row stores the Experiment ID, user ID, cutoff, target, gold label, predicted

label, confidence, token counts, and trace-file pointer. Metric tables are generated from these files, never transcribed manually.

The user bootstrap intervals in Table 21 quantify sampling uncertainty from paired predictions; the three-run `KSR-VALID-4W` analysis reports decoding variance separately. These two uncertainty sources are not pooled.

H.2 Compact protocol specification

The executable source uses stable keys aligned with the neutral display names in Table 31:

Display	Executable key
TC	`FUTURE_CONTINUED_INTENT`
LR	`ONE_SHOT_TASK_SEARCH`
PR	`HABITUAL_RETURN`
UR	`UNSATISFIED_INTENT`
No Target	`NO_CLEAR_RESIDUAL`

Table 31: Display labels and their corresponding executable keys.

HDC ranks source-linked history cards by falsification potential, contrast, support, and salience; protected exceptions survive repetition merging. Four rooms organize observed behavior, search–consumption contrast, temporal trajectory, and adjudication cues. The Palace Scout produces candidates without a verdict. The Falsification Judge records support and counterevidence and returns an initial ranking. The Blind Flow Auditor sees unordered candidate identities but no judge verdict, ranking, confidence, or rationale. The Synthesizer arbitrates the two isolated records and cites current evidence IDs. A deterministic validator enforces history-side necessary conditions and serialization rules.

H.3 Fixed evidence-value mapping

Table 32 summarizes the fixed implementation of (σ, κ, ϕ) used by HDC. Let q and b be separately normalized search and viewing shares, $g = q - b$, n_q and n_b the numbers of search and viewing records, Δ the late-minus-early viewing share, m four-stage trajectory momentum, a_s post-search viewing share, a_d delayed post-search viewing share, n_d delayed-view count, and $\bar{e}$ mean search engagement. All `clip` operations truncate to $[0, 1]$.

Evidence	Support σ	Contrast κ	Falsif. ϕ
Viewing event	$\min(a, 3)$	.1 if multiple category tags, else 0	0
Search event	recency weight a	.2 if categorized, else 0	0
Residual	$\max(q, b)$	$\lvert g\rvert$	$.6\lvert g\rvert$
Temporal count	.2	$\lvert n_q - n_b\rvert/(n_q + n_b)$	.2 if both channels occur, else 0
Trend	$\max(b_{\text{late}}, \bar{e}, a_s, a_d)$	$\max\{\lvert\Delta\rvert, \lvert m\rvert, a_s, a_d\}$	fixed branch below
Task state	$\max(T, U, C)$	$\max(\lvert T - U\rvert, C)$	fixed branch below
Habitual anchor	$\max(A, R)$	$\max(D, \lvert\Delta\rvert, R)$	fixed branch below
Outcome cue	$\max(q, b, \bar{e}, a_s)$	$\max(\lvert g\rvert, \lvert\Delta\rvert, a_s)$	.55 if $n_{\text{after}} > 0$ and $\Delta \geq .05$; else .45 if $\lvert g\rvert \geq .10$; else .15

Table 32: Deterministic evidence-value mapping. Here a denotes the predefined salience/recency weight for the raw event type.

The outcome-cue row describes category-level search–view cues. The separate historical dominant-category cue uses its viewing share for support, .30 for contrast, and .20 for falsification.

For trend evidence, $\phi = .55$ when $n_d > 0$ and $m \geq .03$; $\phi = .50$ when there is an immediate bridge but no delayed bridge; $\phi = .35$ when $\Delta \leq -.08$; $\phi = .45$ when searches occur without later viewing; and $\phi = .20$ otherwise. Its salience is $a = \max(|\Delta|, |m|, a_s, a_d)$.

For task-state evidence, define

$$
\begin{aligned}
L &= \begin{cases} 1, & 0 < n_q \le 3, \\ \max(0, 1 - \min(n_q, 10)/10), & \text{otherwise}, \end{cases} \\
N &= \mathbb{1}[n_d = 0 \wedge a_d < .03], \\
I &= \mathbb{1}[n_{\text{immediate}} > 0 \wedge n_d = 0], \\
T &= \text{clip}(.45\bar{e} + .25L + .20N + .10I), \\
U &= \text{clip}\big(.55(1 - \bar{e}) + .25\min(n_q/5, 1) \\
&\qquad + .20[1 - \min(4\max(a_s, a_d), 1)]\big), \\
C &= \text{clip}\big(.45\min(4a_d, 1) + .30\max(m, 0) \\
&\qquad + .25\min(n_d/2, 1)\big).
\end{aligned}
$$

Its salience is $\max(q, |g|, T, U)$. The falsification value is .20 if $g < .08$ and $q < .15$; .60 if $T \ge .72$ and $C < .45$; .55 if $U \ge .65$ and $C < .45$; .50 if $C \ge .45$; and .35 otherwise.

For habitual-anchor evidence, let H be the larger of the overall and late share of the dominant viewing category, E the normalized viewing entropy, $P = 1 - E$, $B = \mathbb{1}[n_d = 0 \wedge m \le .02]$, J indicate that the target is the dominant anchor, and $D = \max(0, q - b)$. Then

$$
\begin{aligned}
A &= \text{clip}(.40H + .25P + .20B + .15(1 - J)), \\
R &= \text{clip}\big(.55A + .25\min(2D, 1) \\
&\qquad + .20\max(\Delta_{\text{anchor}}, 0)\big).
\end{aligned}
$$

Its salience is $\max(A, R, |g|)$. The falsification value is .20 for $q < .10$ and $|g| < .08$; .30 when $J = 1$; .60 when $R \ge .58$, $n_d = 0$, and $a_d < .10$; .45 when $A \ge .55$; and .25 otherwise. The HDC score is exactly $d(e) = .50\phi + .35\kappa + .15\sigma + .05a$. Evidence in the contrast, trajectory, or decision rooms is protected, as is any item with $\phi \ge .25$ or $\kappa \ge .45$. Remaining items are grouped by room, channel, and top-level category; groups larger than three create an anchor with group size, aggregate values, and source IDs while retaining the three highest-scoring representatives before the global cap of 80.

H.4 Complete prompts, configurations, and dynamic inputs

The anonymous archive's `prompts/palace_agents_source.py` contains all four runtime prompts and their dynamic input assembly. The reported entry point is `PalaceAgentSystem.run_palace_outcome`: no cases or feedback, `agent_only` arbitration, and one synthesis. Other source modes are outside the reported pipeline. Tables 33–34 give settings and role boundaries; the archive is a prompt reference, not a complete reproduction package.

Setting	Fixed value
Qwen3-Max	temperature .1, `top_p` = .95
DeepSeek-V4-Flash	temperature .1, thinking disabled
External search	disabled for every role
HDC budget	80 cards; at most 3 representatives per group
Displayed room budget	at most 18 cards per room
Agent execution	one acyclic pass; no role reruns

Table 33: Inference configuration reported in the main paper.

Role	Dynamic input	Withheld input
Palace Scout	ordered room text, target context	all verdicts and future events
Falsification Judge	room text, Scout candidates, target context	flow-audit record and future events
Blind Flow Auditor	room text, target context, alphabetically sorted candidate IDs	Judge label, ranking, confidence, rationale, and evidence path
Synthesizer	room text, Scout candidates, isolated Judge record, isolated Auditor record, target context	future events

Table 34: Runtime dynamic-input contract for the four logical calls.