SAME LABELS, DIFFERENT FORMATS: AUDITING CONVERSATIONAL PERSONALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Giving a conversational protocol and an in-context learning (ICL) baseline the same labels does not control how those labels are presented. We audit this distinction in individual purchase prediction with frozen language models. On a 200-person cohort from Twin-2K-500, DeepSeek-V3.2 scores 0.477 F_1 with block-format ICL (block ICL) and 0.699 with a conversational outcome-feedback protocol. Presenting the same seven labelled examples as a fixed dialogue within one request yields 0.697, without generated context turns or feedback on model predictions; the paired protocol-minus-dialogue difference is +0.002 (95% participant-bootstrap interval [-0.031, 0.033]). A disjoint 200-person cohort reproduces the pattern (0.429, 0.703 and 0.705 for block ICL, dialogue ICL and the protocol). The direction depends on the model: Qwen-2.5-72B scores 0.705, 0.668 and 0.657; a wording-matched single-message control shows that the lever is message format (dialogue minus control +0.311, [0.241, 0.382]). Re-running three conditions changes F_1 by less than 0.01. Further controls show that the conversation's advantage is carried by the participant's own history rather than the demographic profile, and that accuracy and prediction propensity must accompany F_1; a nine-model panel, two slate-choice panels and an acknowledgment intervention set the audit's scope. Across the three models run under matched conditions, a three-call non-interactive baseline performs at or above the 49-call conversational protocol, which shows no resolved advantage on any of them; presentation, not only labels, is therefore the baseline against which claims about interaction should be judged.

1 INTRODUCTION

A conversational prediction protocol can receive exactly the same labelled examples as a single-request baseline while presenting them differently. This creates an attribution problem: an advantage over that baseline may reflect the presentation of available evidence, alongside any benefit of generated interaction or outcome feedback. We examine this problem in individual purchase prediction, where language models are routinely asked to stand in for particular people (Argyle et al., 2023; Horton et al., 2024; Brand et al., 2023; Park et al., 2023; Hewitt et al., 2024). A frozen model receives a participant's profile and seven past purchase decisions, then predicts three held-out decisions. In the *conversational outcome-feedback protocol*, the model predicts each context product before its label is revealed; *block ICL* receives those labels together in one prompt. Our central control, *dialogue ICL*, presents the same labelled examples as a fixed sequence of user and assistant turns within one request, without generating the context conversation. Comparing these three conditions tests whether an observed protocol advantage persists against a dialogue baseline.

Table 1 shows why this control changes the interpretation. On DeepSeek-V3.2, block ICL scores 0.477 F_1, dialogue ICL 0.697, and the protocol 0.699; a disjoint participant cohort reproduces the pattern at 0.429, 0.703 and 0.705. On Qwen-2.5-72B, dialogue serialization instead lowers performance: block ICL scores 0.705, dialogue ICL 0.668, and the protocol 0.657. Changing only the presentation can thus move block ICL up or down to the protocol's level. We use these results to organize an audit of conversational personalization: compare alternative presentations of the same labels, distinguish participant-history matching from persona matching, and examine accuracy and prediction propensity alongside F_1. The resulting recommendation is to include a fixed-dialogue baseline before attributing a protocol-minus-block advantage to interaction.

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.

Contributions.

1. **A replicated attribution counterexample** (§3). On V3.2, the same seven labelled examples presented as a fixed dialogue in one request raise F_1 from 0.477 to 0.697, against 0.699 for the 49-call conversational protocol; a disjoint cohort reproduces the pattern (0.429, 0.703, 0.705), and three repeated cells move by less than 0.01. The protocol's advantage over block ICL therefore does not by itself identify a benefit of interaction.
2. **Evidence that the lever is message format** (§3). A wording-matched single-message control scores 0.387, and the dialogue exceeds it by +0.311 ([0.241, 0.382]); across six model configurations the dialogue–block difference is resolved (interval excludes zero) on V3.2 and Qwen and points in opposite directions, so the control has to be run on the model under evaluation.
3. **An operational audit** (§4, §5). Replacing the participant's history with another participant's lowers protocol F_1 (matched minus swapped +0.060, [0.020, 0.102]), while swapping the demographic profile produces no resolved difference (+0.001, [-0.030, 0.033]); accuracy and prediction propensity reported alongside F_1 show that label-free conversation raises F_1 by predicting purchase more often; the protocol and dialogue ICL score above a tuned collaborative filter and logistic regression (0.652, 0.626) on the identical decisions; a nine-model panel, two slate-choice panels, trajectory diagnostics and an acknowledgment intervention establish the audit's scope.

2 Evaluation design and attribution controls

Task and cohort. Twin-2K-500 (Toubia et al., 2025) records, for 2,058 U.S. respondents, binary purchase decisions on 40 grocery products at stated prices together with 14 demographic fields. We use a frozen cohort of $N{=}200$ participants; for each participant a fixed random subset of 10 products is split into 7 *context* products and 3 *held-out* products (the split file and its hash are released). A *cell* is one model $\times$ method $\times$ demographics condition on one cohort. Every method predicts the same 600 held-out decisions. Purchase is the positive class (prevalence 0.455) and F_1 is the primary metric; accuracy and the predicted-positive rate are reported alongside so that a change in F_1 is never read in isolation. Uncertainty is a participant-level bootstrap (10,000 resamples of participants with all their predictions, percentile 95% intervals; paired contrasts resample the same participants in both arms); a difference whose interval excludes zero is *resolved*; an interval covering zero marks an unresolved difference, not equivalence. This applies to every contrast below. All 130 reported paired contrasts are exploratory; Appendix L lists each with a two-sided bootstrap p-value and a Benjamini–Hochberg adjustment at $q{=}0.05$ over the whole family (61 intervals exclude zero, 57 survive the adjustment).

Persona rendering. When supplied, demographics are rendered as the 14 fields in a question–answer list under a "Persona Profile" header.

Methods. All methods use temperature 0 (the GPT-5 and V4-pro-thinking APIs accept no temperature parameter) and the same product rendering. *Single-call zero-shot*: one call per held-out product. *Block ICL*: one call per held-out product whose prompt lists the 7 context products with their true labels in one user message; this matches the labelled information the protocol receives. *Dialogue ICL*: the same 7 labelled examples in one request, rendered as a fixed sequence of user turns (each context product with the prediction question) and assistant turns holding the participant's actual decision as a bare digit, followed by the held-out product; no turn is generated by the model and no feedback on a model prediction occurs (Appendix N). *Warm-up only*: a conversation that introduces the persona, receives an acknowledgment, asks for the 7 context predictions in sequence with no feedback, then the 3 held-out predictions. *Self-feedback*: the full protocol structure, but after each committed context prediction the model is asked to reflect on its own answer instead of being told the label. *Conversational protocol*: after each committed context prediction the true label is revealed ("Correct answer: You purchased (1) / did not purchase (0) this product."), the model acknowledges, predicts the 3 held-out products, and those held-out turns are removed from the history before the next context product; the held-out predictions made after the seventh context round (*Round 7*) are the output. Per participant the protocol makes about 49 calls, single-call methods

Table 1: Serialization check, with demographics, Round-7 held-out F_1 with participant-bootstrap 95% intervals: block ICL, the same labelled examples as a fixed dialogue in one request, and the conversational protocol, with the two paired contrasts. September 2026 cells; the Qwen protocol cell is the intervention control arm of §5.

Model	block ICL	dialogue ICL	protocol	dialogue − block	protocol − dialogue
DeepSeek-V3.2 (pinned host)	0.477 [0.406, 0.544]	0.697 [0.654, 0.737]	0.699 [0.653, 0.741]	+0.220 [0.151, 0.293]	+0.002 [-0.031, 0.033]
Qwen-2.5-72B (pinned bf16 host)	0.705 [0.660, 0.747]	0.668 [0.622, 0.711]	0.657 [0.611, 0.701]	-0.038 [-0.066, -0.010]	-0.010 [-0.042, 0.022]
DeepSeek-V3.2, fresh cohort	0.429 [0.355, 0.499]	0.703 [0.656, 0.746]	0.705 [0.655, 0.750]	+0.274 [0.199, 0.350]	+0.002 [-0.037, 0.038]
DeepSeek V4-pro (GA, thinking on)	0.714 [0.669, 0.755]	0.705 [0.661, 0.744]	0.692 [0.646, 0.733]	-0.009 [-0.034, 0.016]	-0.013 [-0.044, 0.019]
DeepSeek V4-pro (thinking off)	0.686 [0.642, 0.726]	0.667 [0.620, 0.709]	0.688 [0.644, 0.730]	-0.019 [-0.062, 0.023]	+0.022 [-0.005, 0.048]
Llama 3.3 70B (September ICL cells; protocol April)	0.673 [0.631, 0.713]	0.673 [0.628, 0.715]	–	-0.001 [-0.025, 0.024]	–
Claude Haiku 4.5 (September ICL cells; protocol April)	0.688 [0.643, 0.729]	0.663 [0.617, 0.706]	–	-0.025 [-0.054, 0.004]	–

Table 2: Accuracy / predicted-positive rate (PPR) for every cell of Table 1; prevalence 0.455 (frozen cohort) and 0.405 (fresh cohort).

Model	block ICL Acc / PPR	dialogue ICL Acc / PPR	protocol Acc / PPR
DeepSeek-V3.2 (pinned host)	0.660 / 0.195	0.693 / 0.558	0.702 / 0.537
Qwen-2.5-72B (pinned bf16 host)	0.695 / 0.580	0.677 / 0.518	0.668 / 0.513
DeepSeek-V3.2, fresh cohort	0.680 / 0.155	0.725 / 0.520	0.732 / 0.503
DeepSeek V4-pro (GA, thinking on)	0.722 / 0.517	0.707 / 0.538	0.673 / 0.605
DeepSeek V4-pro (thinking off)	0.612 / 0.780	0.683 / 0.495	0.698 / 0.513
Llama 3.3 70B (September ICL cells; protocol April)	0.613 / 0.728	0.630 / 0.675	– / –
Claude Haiku 4.5 (September ICL cells; protocol April)	0.688 / 0.543	0.625 / 0.657	– / –

3. Verbatim prompts are in Appendix N. Every arm is parsed by one rule, fixed in April 2026 before the September runs (the *April rule*): a response containing "1" scores 1; otherwise a response containing "0" scores 0. September cells archive raw responses, and Appendix O re-scores them under first-digit and strict rules, which matters for the GA build of V4-pro (§5); April cells kept only parsed predictions.

Models and endpoints. The panel (§5) uses nine models on the identical cohort, eight run in April 2026 and DeepSeek-V3.2 re-run in September 2026: DeepSeek-V3.2, GPT-4o-mini, Claude Haiku 4.5, Gemini 2.0 Flash, Llama 3.3 70B, Qwen-2.5-72B, and three reasoning-mode models (Gemini 3.1 Pro, DeepSeek V4-pro, GPT-5). The contemporaneous cells (§3, §5) were run in September 2026: DeepSeek-V3.2 through a pinned third-party host of the public V3.2 weights (the first-party endpoint no longer serves V3.2), DeepSeek V4-pro through the first-party endpoint (post August-2026 general-availability (GA) release), and Qwen-2.5-72B through a pinned bf16 host. Model identifiers, providers, finish reasons and token counts are logged per request for the September cells; no main-text contrast mixes cells from different collection periods.

3 SAME LABELS, DIFFERENT BASELINES

Table 1 compares three ways of using the same seven labelled examples: block ICL, dialogue ICL, and the conversational protocol. On V3.2 their F_1 scores are 0.477, 0.697 and 0.699; the disjoint cohort gives 0.429, 0.703 and 0.705. On Qwen the corresponding scores are 0.705, 0.668 and 0.657. The presentation change thus moves F_1 in opposite directions across models, while the protocol stays close to dialogue ICL on both. These comparisons evaluate a presentation package that changes message roles and wording while preserving the labelled examples; all September V3.2 cells were run within 48 hours on the pinned host, and each contrast compares cells on the same cohort. We first examine replication and repeated cells, then compare the scale of performance against statistical baselines.

Other models and a wording-matched block prompt. The same dialogue and block ICL cells were run on four further model configurations in September (Table 1): Llama 3.3 70B and Claude Haiku 4.5 (dialogue minus block -0.001, [-0.025, 0.024]; -0.025, [-0.054, 0.004]; their protocol cells are April measurements and are not paired here), and DeepSeek V4-pro with thinking on (-0.009, [-0.034, 0.016]; protocol minus dialogue -0.013, [-0.044, 0.019]) and off (-0.019, [-0.062, 0.023]; +0.022, [-0.005, 0.048]). To compare message formatting while holding the example wording fixed, a further single-message variant places the dialogue's exact text (intro, "Understood.", per-example question and bare-digit answer) inside one user turn with no assistant turns: on V3.2 it scores

Table 3: Statistical baselines on the identical 600 held-out decisions. The logistic regression uses one-hot encodings of the 14 demographics, is trained per product on the 1,858 participants outside the frozen cohort; the collaborative filter's donors are the 1,658 participants outside both evaluation cohorts (Appendix I).

Baseline	F_1	95% CI	Acc	PPR
L1 logistic on 14 demographics (participant-disjoint, per product)	0.571	[0.513, 0.625]	0.627	0.415
L1 logistic on demographics + product + participant history rate, F1-tuned threshold	0.626	[0.576, 0.674]	0.588	0.647
kNN collaborative filtering on the 7 context labels (participant-disjoint donors, K and threshold tuned)	0.652	[0.606, 0.695]	0.577	0.762
always predict purchase	0.625	[0.582, 0.667]	0.455	1.000
product base rate (non-cohort participants)	0.489	[0.427, 0.546]	0.620	0.288
participant's own context purchase rate > 0.5	0.528	[0.461, 0.590]	0.580	0.435

F_1 0.387 (-0.090 relative to block ICL, [-0.162, -0.018]; dialogue ICL minus the single-message variant +0.311, [0.241, 0.382]), and on Qwen 0.681 (-0.024, [-0.047, -0.002]). On V3.2 the dialogue's wording without its message roles therefore falls below block ICL and predicts a purchase for only 0.148 of held-out items (dialogue ICL: 0.558); the V3.2 results support an advantage of dialogue message formatting when wording is held fixed, though this formatting change alters message roles and message boundaries together. On Qwen the variant falls slightly below block ICL (interval excludes zero before multiplicity adjustment; Appendix L); the other four configurations show no resolved dialogue–block difference. Pooling the two cohorts (n=400), the protocol-minus-dialogue-ICL difference is +0.002 ([-0.023, 0.026]); the protocol exceeds warm-up-only by +0.046 ([0.019, 0.072]), and dialogue ICL exceeds block ICL by +0.245 ([0.194, 0.298]).

Replication, repeated cells and splits. Dialogue ICL raises F_1 with demographics from 0.477 to 0.697 ([0.654, 0.737]; accuracy 0.693, PPR 0.558; +0.220 over block ICL, [0.151, 0.293]). It also exceeds block ICL with the persona placed in the system message (+0.132, [0.073, 0.194]) and reaches 0.717 without a profile under the participant-specific instruction (instruction variants: §4). The gain also appears in accuracy (Table 2: block ICL 0.660, dialogue ICL 0.693; fresh cohort 0.680 to 0.725), and every V3.2 headline cell is unchanged under the first-digit and strict scoring rules of Appendix O, since every response is a bare digit. The protocol's paired difference from dialogue ICL is +0.002 ([-0.031, 0.033]). On this pinned V3.2 endpoint, dialogue ICL therefore recovers the observed advantage over block ICL using one request per held-out prediction, leaving the protocol's incremental benefit unresolved. The direction of the serialization effect is model-specific (Table 1): on Qwen-2.5-72B, where block ICL outperforms the protocol, the same dialogue serialization *lowers* ICL from 0.705 to 0.668 (-0.038, [-0.066, -0.010]), and the protocol's difference from dialogue ICL is again unresolved (-0.010, [-0.042, 0.022]). On both models a single request that serializes the same labels as a dialogue reaches a similar aggregate level to the protocol, in either direction relative to block ICL; the intervals permit residual differences of a few hundredths. The direction is model-specific, which is why the control has to be run on the model under evaluation. Run-to-run variation at temperature 0 is small relative to these contrasts: repeating the V3.2 protocol, the V3.2 dialogue ICL and the Qwen dialogue ICL cells on the same cohort gives paired rerun-minus-original differences of 0.000 ([-0.024, 0.023]), -0.007 ([-0.026, 0.009]) and -0.006 ([-0.019, 0.008]), and the repeated protocol minus the repeated dialogue ICL is +0.009 ([-0.023, 0.040]). A fresh cohort of 200 participants disjoint from the frozen cohort (Appendix K, Table 1) replicates the pattern: protocol minus block ICL +0.276 ([0.201, 0.350]), dialogue minus block ICL +0.274 ([0.199, 0.350]), protocol minus dialogue ICL +0.002 ([-0.037, 0.038]). On this cohort the protocol's increment over the label-free warm-up conversation is resolved (+0.069, [0.032, 0.105]; frozen cohort +0.024), so a label-free conversation does not always recover the whole gain, whereas the label-carrying single request does on both cohorts. Re-drawing the 7/3 product split for the same 200 participants with 5 further seeds (Appendix J) gives protocol-minus-block-ICL differences from +0.197 to +0.254 (5 of 5 intervals exclude zero); these splits replicate the protocol-versus-block comparison only, not the dialogue comparison.

Against statistical baselines. On the identical 600 decisions (Table 3), a participant-disjoint L1 logistic regression on the 14 demographics reaches 0.571, a tuned demographics-plus-product-plus-history model 0.626, always predicting purchase 0.625, and a k-nearest-neighbour collaborative filter that matches each participant to the k most similar participant-disjoint donors on the seven context labels (k and threshold tuned on validation participants) 0.652 ([0.606, 0.695]). The proto-

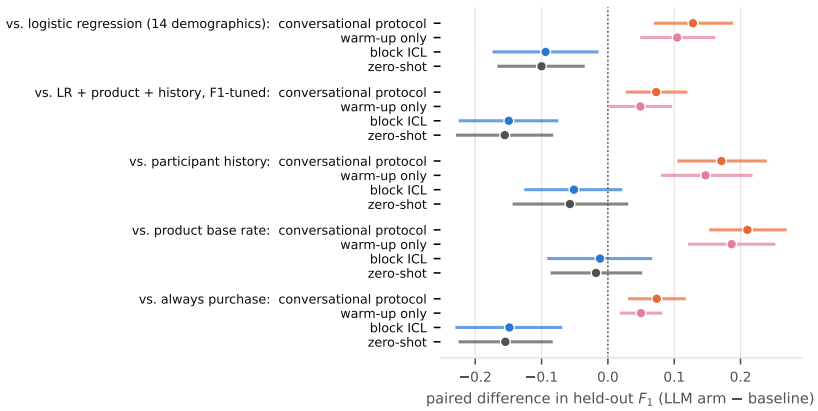

Figure 1: Four selected September V3.2 arms (with demographics) versus the matched statistical baselines on the identical 600 decisions: exploratory paired differences in held-out F_1 with pointwise participant-bootstrap 95% intervals, without multiplicity adjustment.

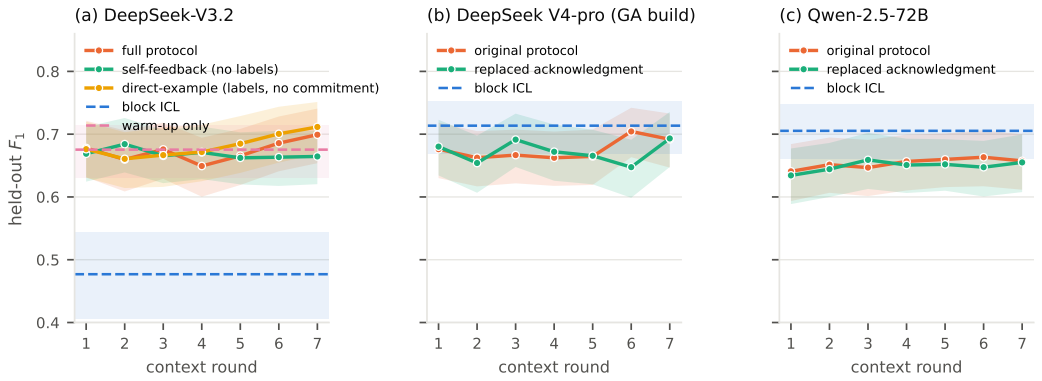

Figure 2: Held-out F_1 after each context round (pointwise participant-bootstrap 95% bands), September-2026 cells with demographics. (a) DeepSeek-V3.2: full protocol, self-feedback and direct-example conversations, with block ICL and the warm-up-only conversation (Round-7 predictions only) as horizontal references. (b) DeepSeek V4-pro (GA build) and (c) Qwen-2.5-72B: original versus replaced-acknowledgment protocol with block ICL as reference.

col exceeds the collaborative filter by +0.047 ([0.008, 0.087]) and the tuned regression by +0.073 ([0.030, 0.117]); block ICL and zero-shot fall below all of them (Appendix I). The protocol and dialogue ICL score above every statistical comparator, including a collaborative filter with 1,658 donors tuned on 400 validation participants; dialogue ICL reaches that level with three requests per participant against about 49 for the protocol (paired differences in Figure 1).

4 Participant History, Persona and Prediction Propensity

Label-free conversation versus the full protocol. The protocol reaches 0.699 with demographics, +0.222 over block ICL ([0.154, 0.293]). The warm-up-only conversation, which reveals no label, reaches 0.675 (protocol minus warm-up +0.024, [-0.013, 0.061]; resolved on the fresh cohort, +0.069, [0.032, 0.105], and pooled, +0.046, [0.019, 0.072]); self-feedback reaches 0.665 (the protocol exceeds it by +0.035, [-0.002, 0.072]); the direct-example conversation, which shows the labels without a prior commitment, reaches 0.712 (protocol minus direct-example -0.012, [-0.042, 0.018]). Accuracy and the predicted-positive rate qualify the label-free result: warm-up-only reaches its F_1 by predicting purchase for 70% of decisions at a *lower* accuracy (0.627) than block ICL (0.660),

Table 4: DeepSeek-V3.2, contemporaneous September-2026 cells, Round-7 held-out F_1 with participant-bootstrap 95% intervals; PPR = predicted-positive rate (prevalence 0.455).

	no demographics				with demographics			
Method	F_1	95% CI	Acc	PPR	F_1	95% CI	Acc	PPR
Single-call zero-shot	0.626	[0.573, 0.676]	0.640	0.508	0.471	[0.411, 0.528]	0.637	0.232
Block ICL (7 labelled examples)	0.565	[0.498, 0.627]	0.690	0.258	0.477	[0.406, 0.544]	0.660	0.195
Warm-up only (no labels)	0.678	[0.632, 0.721]	0.602	0.783	0.675	[0.633, 0.715]	0.627	0.695
Direct-example multi-turn (labels shown, no commitment)	0.675	[0.625, 0.721]	0.708	0.443	0.712	[0.668, 0.751]	0.728	0.487
Self-feedback (commit, no labels)	0.665	[0.618, 0.708]	0.635	0.633	0.665	[0.620, 0.706]	0.655	0.573
Conversational protocol (Round 7)	0.689	[0.641, 0.731]	0.683	0.562	0.699	[0.653, 0.741]	0.702	0.537

whereas the protocol and direct-example improve both (0.702 and 0.728) at a predicted-positive rate near the prevalence. Part of the label-free "gain" is thus a shift in response propensity. Figure 2 shows that on V3.2 the protocol's advantage over block ICL is present from the first held-out prediction and that the intervention arms of §5 track each other round by round.

Does the conversation use the individual's information? Two controls answer this on the same cohort (Appendix H, Table 12). In the *shuffled-persona* cells each participant's conversation opens with another participant's 14-field profile (a fixed derangement), and in the *shuffled-history* cell the seven revealed context products and labels belong to another participant while the held-out products remain the participant's own. The protocol's paired difference over its shuffled-persona counterpart is +0.001 ([-0.030, 0.033]), the warm-up-only conversation's is +0.009 ([-0.021, 0.037]), and the protocol's difference over the shuffled-history version is +0.060 ([0.020, 0.102]); the shuffled-persona protocol still exceeds block ICL by +0.221 ([0.154, 0.290]). Matching the persona to the correct participant therefore yields no resolved improvement over another participant's persona, in either conversation type. Replacing the participant's context products and labels with another participant's reduces protocol F_1 resolvably, which demonstrates a benefit from participant-matched history as a bundle, without separating product selection, purchase propensity and product-specific preferences. The protocol's advantage over block ICL survives an incorrect persona assignment.

Instruction sensitivity. The no-demographics condition of Table 4 also changes the system instruction ("a typical American shopper" instead of "THIS participant"), so the with-/without-demographics comparison bundles profile availability with an instruction change and does not identify the effect of supplying demographics; matched-instruction cells are in Appendix H. The apparent demographic penalty therefore cannot be attributed to demographics; rendering the persona inside the system message raises block ICL by +0.088 ([0.044, 0.135]), and dialogue ICL still exceeds that rendering by +0.132.

5 Supporting checks on feedback attribution

The remaining experiments establish the scope of the audit across models, tasks, mechanism and measurement.

Nine-model panel. Table 5 and Figure 3 report the protocol and block ICL on the identical cohort for nine models, eight run in April 2026 and V3.2 in September; the panel lacks the dialogue control, so it measures cross-model spread but cannot test serialization. The protocol-minus-block difference ranges from +0.222 on V3.2 to -0.083 ([-0.120, -0.047]) on Gemini 3.1 Pro, the only April loss that excludes zero; Qwen's September loss (-0.048, [-0.082, -0.016]) also excludes zero, and reasoning-mode models do not systematically score higher or lower.

Slate-choice panels. To test whether the difference is specific to survey purchase intent, we reformulate two revealed-preference panels as four-option slate choice: for each held-out purchase the model picks the bought item among three distractors from the same taxonomy, year and price band (Amazon-Study and H&M, 200 users each, 7 context and 3 held-out events, chance 0.25; Appendix M). Table 6 reports four arms, all with the profile, run on `deepseek-flash` (DeepSeek V4.1). The protocol-minus-ICL difference is unresolved on both panels (+0.013, [-0.022, 0.048]; +0.033, [0.000, 0.068]), the direct-example conversation exceeds ICL on both

Table 5: Round-7 held-out F_1 on the fixed 200-participant cohort (600 predictions per cell) with participant-bootstrap 95% intervals (accuracy and predicted-positive rate per cell in Appendix E); April-2026 panel except the DeepSeek-V3.2 row (September 2026, pinned host, 14-field rendering). Lift = protocol minus block ICL, with demographics, paired bootstrap. (R) = reasoning mode. †: per-participant files for these six ICL cells (three models, both conditions) were not retained, so the point estimate is reported without an interval. The Claude Haiku protocol cells are incomplete (n with/without demographics shown); paired contrasts use the participants present in both arms.

	no demographics		with demographics		
Model	ICL F_1 [CI]	protocol F_1 [CI]	ICL F_1 [CI]	protocol F_1 [CI]	lift (with demo)
DeepSeek-V3.2 (Sep 2026)	0.565 [0.498, 0.627]	0.689 [0.641, 0.731]	0.477 [0.406, 0.544]	0.699 [0.653, 0.741]	+0.222 [0.154, 0.293]
Llama 3.3 70B	0.685 [0.640, 0.727]	0.688 [0.641, 0.729]	0.672 [0.629, 0.714]	0.693 [0.648, 0.735]	+0.021 [-0.009, 0.050]
Claude Haiku 4.5 (n=196/200)	0.693†	0.682 [0.639, 0.724]	0.679†	0.673 [0.629, 0.715]	-0.006†
GPT-5 (R)	0.697 [0.648, 0.740]	–	0.710 [0.664, 0.752]	0.701 [0.655, 0.743]	-0.008 [-0.037, 0.020]
Qwen-2.5-72B	0.712 [0.666, 0.755]	0.684 [0.641, 0.726]	0.710 [0.665, 0.751]	0.685 [0.639, 0.727]	-0.025 [-0.058, 0.006]
GPT-4o-mini	0.691†	0.692 [0.648, 0.733]	0.715†	0.667 [0.617, 0.713]	-0.048†
DeepSeek V4-pro (R)	0.671†	0.643 [0.590, 0.691]	0.691†	0.622 [0.567, 0.671]	-0.069†
Gemini 3.1 Pro (R)	0.690 [0.641, 0.736]	0.665 [0.615, 0.712]	0.700 [0.653, 0.743]	0.617 [0.562, 0.669]	-0.083 [-0.120, -0.047]
Gemini 2.0 Flash	–	0.703 [0.658, 0.745]	–	0.671 [0.626, 0.714]	–

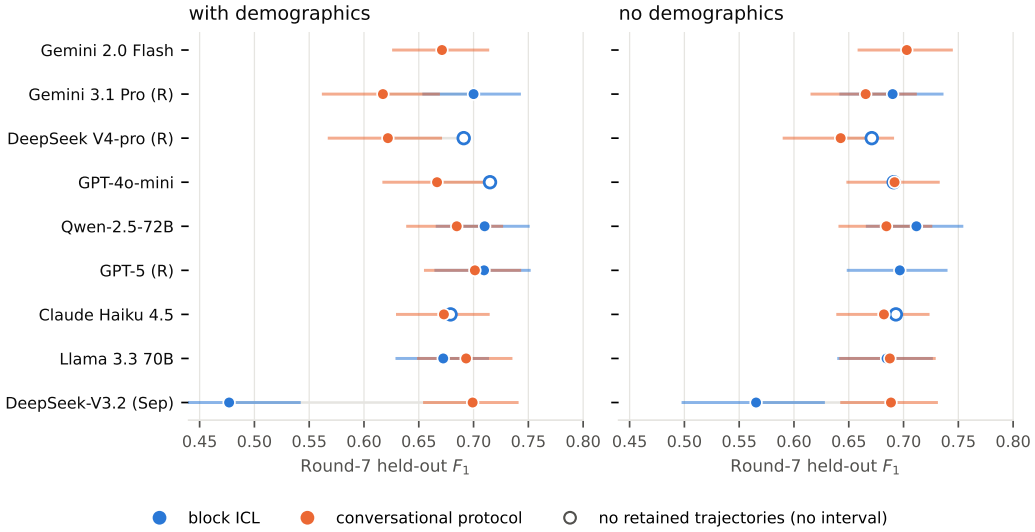

Figure 3: Round-7 held-out F_1 of block ICL (blue) and the conversational protocol (orange) per model, with participant-bootstrap 95% intervals, ordered by the with-demographics lift. Hollow markers: April ICL cells without retained per-participant files (no interval). The DeepSeek-V3.2 row is the September-2026 contemporaneous cell; all other rows are the April-2026 panel.

(+0.042, [0.005, 0.077]; +0.052, [0.013, 0.090]) with the protocol not resolvably different from it (-0.028, [-0.057, 0.000]; -0.018, [-0.055, 0.018]), and the label-free self-feedback conversation falls *below* block ICL on both (-0.098, [-0.145, -0.053]; -0.102, [-0.155, -0.050]). The role of revealed labels thus differs by model and task: on V3.2 purchase intent a label-free conversation recovers most of the gain, whereas on these panels the direct-example conversation exceeds ICL on both panels before multiplicity adjustment (only H&M survives it), committing before each reveal adds no resolved difference, and a reflective conversation without labels is harmful.

Trajectory diagnostic. A protocol loss could arise because the model treats each revealed label as a rule and re-labels the person's other decisions. On every panel model's held-out trajectories we measure how often a correct held-out prediction flips to an incorrect prediction equal to the label just revealed (Appendix F, Figure 4). The April build of DeepSeek V4-pro has by far the highest label-directed error rate (0.222), and its protocol loss is consistent with this mechanism; the rate does not track the other models' lifts (GPT-4o-mini has the lowest rate, 0.039, and Gemini 3.1 Pro's

Table 6: Slate-choice panels: Round-7 top-1 accuracy (chance 0.25) with participant-bootstrap 95% intervals, 200 users per dataset, with profile, `deepseek-flash` (V4.1), September 2026.

Dataset	ICL	protocol	direct-example	self-feedback	protocol − ICL
Amazon	0.390 [0.348, 0.433]	0.403 [0.360, 0.447]	0.432 [0.388, 0.475]	0.292 [0.250, 0.333]	+0.013 [-0.022, 0.048]
H&M	0.440 [0.395, 0.487]	0.473 [0.430, 0.518]	0.492 [0.447, 0.537]	0.338 [0.298, 0.380]	+0.033 [0.000, 0.068]

resolved loss comes with a rate, 0.119, comparable to V3.2's 0.113), so the diagnostic is consistent with V4-pro's loss and does not identify a general mechanism.

Acknowledgment intervention. Replacing the model's free-text acknowledgment of each revealed label with the fixed turn "Acknowledged." tests whether verbose acknowledgments drive protocol losses; the decision rule was specified before launch (Appendix B) and the cells ran on DeepSeek V4-pro (GA build) and Qwen-2.5-72B with interleaved controls and fresh block ICL (Appendix D, Table 8). Replacement produces no resolved change in F_1 or in the synchronized-error rate (the fraction of participant-rounds in which at least two held-out predictions flip from correct to incorrect toward the label just revealed) on either model (V4-pro +0.001, [-0.030, 0.032]; Qwen -0.002, [-0.030, 0.024]) under every scoring rule (Appendix D).

6 Related work

LLMs as individual predictors and recommenders. Silicon samples and persona-conditioned simulation (Argyle et al., 2023; Aher et al., 2023; Horton et al., 2024; Park et al., 2023; Brand et al., 2023; Hewitt et al., 2024) motivate the task; digital twins correlate only weakly with the people they imitate (Peng et al., 2025), and persona variables explain little annotation variance (Santurkar et al., 2023; Dominguez-Olmedo et al., 2024). Predicting held-out choices from a person's labelled history is also a recommendation problem: LLMs rank items conditioned on interaction histories (Hou et al., 2024; Yang et al., 2023), LaMP evaluates generation personalized with retrieved user-profile entries (Salemi et al., 2024), and our kNN comparator follows neighbourhood collaborative filtering (Breese et al., 1998).

Prompt format and chat serialization. ICL is sensitive to calibration, example order, label content, position and surface formatting (Zhao et al., 2021; Lu et al., 2022; Min et al., 2022; Liu et al., 2024; Sclar et al., 2024), and single-template evaluation can change the relative standing of models (Mizrahi et al., 2024). Closest to our setting, Voronov et al. (2024) show that template choice can reduce strong ICL methods to chance; optimal templates also fail to transfer across models, complicating comparisons of ICL improvements. Sensitivity studies estimate the spread over formats for a fixed method; here the format difference sits *inside* a baseline-versus-method comparison, so an unmeasured sensitivity can be read as a benefit of interaction. We extend the concern to dialogue serialization, which changes message roles and wording together; genuine multi-turn exchanges are known to degrade performance (Laban et al., 2026).

Feedback, self-correction and baseline attribution. Self-Refine iterates on self-generated feedback (Madaan et al., 2023) and Reflexion on verbal reflections over internal or external task feedback (Shinn et al., 2023), and Huang et al. (2024) and Kamoi et al. (2024) show that intrinsic self-correction often fails while reported gains frequently depended on oracle labels. Apparent gains have also vanished against properly tuned baselines in language modelling (Melis et al., 2018) and recommendation (Ferrari Dacrema et al., 2019); we apply this logic to conversational feedback and supply a three-request control against it.

7 Discussion and limitations

Matching labels does not control a baseline when their serialization changes performance by this much. On V3.2, dialogue ICL recovers the observed protocol gain without generated interaction, and on Qwen the same serialization falls to the protocol's lower level; the protocol's advantage over collaborative filtering is likewise matched by the one-request baseline. The model-dependent

direction motivates evaluating the baseline format on the target model before attributing an advantage to interaction. Where the protocol wins, participant-matched history contributes and the full protocol modestly outperforms the label-free warm-up when both cohorts are pooled (+0.046, [0.019, 0.072]), while swapping the demographic profile produces no resolved difference. The practical recommendation is correspondingly plain: before running the feedback conversation (about 49 calls per participant as evaluated here, of which 18 are intermediate held-out diagnostics that a deployment would omit), run the one-request dialogue serialization of the same labels, and treat any protocol advantage that does not exceed the dialogue baseline as unattributed.

Limitations. The serialization controls, trajectory diagnostics and intervention use one dataset and task, and all cells except the three repeated ones of §3 are single runs; the V3.2 cells are measurements on the public V3.2 weights served by a third-party host; and the block–dialogue comparison changes message roles and wording together, so the remaining protocol-minus-dialogue contrast is not an isolated estimate of feedback. *Broader impact:* individual-level purchase prediction supports preference elicitation but also profiling and targeted pricing; Appendix G audits error rates by income, politics and region, and deployment should be limited to consented settings with such auditing.

AI USE STATEMENT

We used generative AI tools to implement the experiment runners and analysis scripts, to clean and audit prior result files, to refine hypotheses and give feedback on the experimental and intervention design, to interpret results, and to draft and edit the text; we did not use them to generate synthetic data, to develop theory or proofs, or for translation. The outputs of the evaluated language models are the experimental data and were produced by the released runners. The authors reviewed all AI-generated code, re-executed every analysis from the raw result files with the released scripts, verified every citation, and take full responsibility for the content.

REPRODUCIBILITY STATEMENT

The frozen and fresh participant splits (with hashes), every per-prediction result file (raw model responses for every September cell; the April files contain parsed predictions where retained), per-request metadata for every cell except the two self-feedback cells, whose runner logged call counts only (model identifier, serving provider, finish reason, token counts), the runner and analysis code, and the intervention plan are released at `https://anonymous.4open.science/r/same-labels-different-formats-E8FB`; one script there regenerates every table and figure from the result files. Every table of September-2026 and April-2026 cells with retained files is generated by one script from those files; six April ICL cells (three models, both demographic conditions) lack retained per-participant files (marked †), and per-request metadata was recorded for the September cells. Appendix O lists the endpoints and dates of every September cell.

REFERENCES

Gati V. Aher, Rosa I. Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, 2023.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351, 2023.

Alex Berke, Dan Calacci, Robert Mahari, Takahiro Yabe, Kent Larson, and Sandy Pentland. Open e-commerce 1.0, five years of crowdsourced U.S. Amazon purchase histories with user demographics. *Scientific Data*, 11(1):491, 2024.

James Brand, Ayelet Israeli, and Donald Ngwe. Using GPT for market research. Working Paper 23-131, Marketing Science Institute, 2023.

John S. Breese, David Heckerman, and Carl Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, 1998.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. In *Advances in Neural Information Processing Systems*, 2024.

Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys)*, 2019.

Luke Hewitt, Ashwini Ashokkumar, Isaias Ghezae, and Robb Willer. Predicting results of social science experiments using large language models. *Working paper*, 2024.

H&M Group. H&m personalized fashion recommendations. Kaggle competition dataset, `https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations`, 2022.

John J. Horton, Apostolos Filippas, and Benjamin S. Manning. Large language models as simulated economic agents: What can we learn from *Homo Silicus*? In *Proceedings of the 25th ACM Conference on Economics and Computation (EC '24)*, 2024.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems. In *Advances in Information Retrieval (ECIR)*, 2024.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations (ICLR)*, 2024.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. When can LLMs actually correct their own mistakes? a critical survey of self-correction of LLMs. *Transactions of the Association for Computational Linguistics*, 12, 2024.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2505.06120.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12, 2024.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of ACL*, 2022.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023.

Gábor Melis, Chris Dyer, and Phil Blunsom. On the state of the art of evaluation in neural language models. In *International Conference on Learning Representations (ICLR)*, 2018.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of EMNLP*, 2022.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? A call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of UIST*, 2023.

Tianyi Peng, George Z. Gui, Melanie Brucks, et al. Digital twins as funhouse mirrors: Five key distortions. *arXiv preprint arXiv:2509.19088*, 2025.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. LaMP: When large language models meet personalization. In *Proceedings of ACL*, 2024.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. Whose opinions do language models reflect? In *International Conference on Machine Learning*, 2023.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Olivier Toubia, George Z. Gui, Tianyi Peng, Daniel J. Merlau, Ang Li, and Haozhe Chen. Twin-2k-500: A dataset for building digital twins of over 2,000 people based on their answers to over 500 questions. *Marketing Science*, 44(6):1446–1455, 2025.

Anton Voronov, Lena Wolf, and Max Ryabinin. Mind your format: Towards consistent evaluation of in-context learning improvements. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6287–6310, 2024.

Fan Yang, Zheng Chen, Ziyan Jiang, Eunah Cho, Xiaojiang Huang, and Yanbin Lu. PALR: Personalization aware LLMs for recommendation. *arXiv preprint arXiv:2305.07622*, 2023.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, 2021.

A Historical February-2026 cells and the persona-rendering correction

Correction and comparability. The earlier version's headline comparison combined a protocol run whose persona block contained up to 50 cross-block survey answers (February 2026) with single-prompt baselines that used the 14 demographic fields, and its demographic comparison also changed the system instruction; neither contrast isolated what it claimed to. The September method comparisons use contemporaneous cells with matched persona rendering and system instruction within each condition (§3); historical results are dated explicitly, and comparisons across demographic conditions also change the instruction unless identified as system-matched; the February cells, a re-run under their rendering, and the April panel are labelled as historical wherever they appear (Appendix A).

Rendering comparison. On the September endpoint the richer February rendering (up to 50 survey answers) gives no resolvable difference from the 14-field rendering for the protocol (-0.004, [-0.033, 0.023]) or warm-up-only (-0.002, [-0.027, 0.022]); this neither establishes equivalence nor validates the earlier comparison (Appendix A).

The February-2026 runs behind the earlier version of this study built the persona from every non-purchase question of the full Twin-2K-500 persona, keeping the first 50 answers across survey blocks (demographics, personality inventories, cognitive tests, economic-preference games) and rendering them as bullet lines; the paper described this as 14 demographic fields. The May-2026 single-call and self-feedback cells and all April-2026 panel cells used the 14-field question–answer rendering.

Table 7 reproduces the February point estimates with intervals recomputed from the retained per-participant files, and the contemporaneous re-runs under the February rendering.

Table 7: DeepSeek-V3.2 historical cells (Round-7 held-out F_1, 95% intervals). February cells use the 50-answer persona; the September re-runs use the February rendering on the pinned V3.2 host.

	no demographics		with persona	
Method	F_1	CI	F_1	CI
Conversational protocol (Feb 2026)	0.690	[0.644, 0.732]	0.716	[0.672, 0.757]
Warm-up only (Feb 2026)	0.671	[0.625, 0.713]	0.661	[0.617, 0.702]
Direct-example multi-turn, labels shown, no commit (Feb 2026)	0.660	[0.611, 0.706]	0.663	[0.613, 0.709]
Mirror-RL: multi-turn commit, no feedback (Feb 2026)	0.670	[0.623, 0.713]	0.628	[0.581, 0.672]
Self-feedback (May 2026, 14-field Q/A)	0.666	[0.620, 0.707]	0.670	[0.625, 0.711]
Conversational protocol, Feb persona rendering, re-run Sep 2026	–	–	0.695	[0.649, 0.738]
Warm-up only, Feb persona rendering, re-run Sep 2026	–	–	0.674	[0.629, 0.715]

B INTERVENTION PLAN

The following text was written into the project's internal study plan on 2026-09-22, before any intervention cell was launched. It was not submitted to an external registry and is not a formal pre-registration; the launch record and per-cell source hashes are released with the code. Verbatim:

> We evaluate the original protocol and acknowledgment replacement on the same frozen 200 participants, seven revealed products and three held-out products, with and without demographics, using contemporaneous DeepSeek V4-pro calls with matched settings and interleaved conditions. Replacement changes every retained feedback acknowledgment to "Acknowledged."; labels, message roles and task instructions are unchanged, and subsequent commitments evolve naturally. The primary contrast is replacement minus original, with demographics. Co-primary outcomes are pooled Round-7 F1 and the participant-round frequency, over rounds 2–7, of at least two held-out predictions changing from correct to incorrect toward the latest revealed label. We use 10,000 paired participant bootstrap resamples. Success requires the two-sided 95% interval for the F1 difference to lie above zero and that for the synchronized-error difference to lie below zero. A repair claim additionally requires the contemporaneous original-minus-ICL F1 interval to lie below zero. Conditional on primary success, the same joint rule tests replication without demographics. Demographic interaction, accuracy, positive-prediction rate and individual transitions are secondary. Failure of either primary criterion means joint success was not established, not equivalence. The intervention identifies acknowledgment replacement, not semantic content independently of history length.

Amendment (pre-launch, 2026-09-22). A second-model replication on Qwen-2.5-72B (pinned bf16 host) with the identical three arms (contemporaneous original protocol, replacement, fresh single-prompt ICL) was added before launch, using the same contrasts, outcomes and decision rule; the no-demographics V4-pro replication in the original text became optional and is reported if run. The parser was frozen as the April rule (a response containing "1" scores 1, otherwise one containing "0" scores 0, with a keyword fallback and one in-conversation nudge on unparseable replies); raw responses are archived for first-digit and strict re-scoring.

C NINE-MODEL PANEL: DETAILS

Figure 3 and Table 5 report Round-7 held-out F_1 for the protocol and single-prompt ICL across the panel; the DeepSeek-V3.2 row is the contemporaneous September cell of §3, the other rows are the April-2026 panel. Three regularities survive the intervals. First, the large lift is specific to DeepSeek-V3.2 under the evaluated prompts; its apparent single-prompt demographic penalty reverses when the system instruction is held fixed (§4); on every other model with a usable ICL cell the lift lies within $[-0.09, +0.03]$. Second, the point estimates do not sort by reasoning mode: GPT-4o-mini's protocol point estimate is lower without a reasoning mode (-0.048†; no retained paired files), while GPT-5 in reasoning mode has an unresolved difference (-0.008 [-0.037, 0.020]). Third, only Gemini 3.1 Pro's loss (-0.083 [-0.120, -0.047]) excludes zero among the negative cells; Qwen's

(-0.025 [-0.058, 0.006]) does not. The with-/without-demographics difference inside the protocol (a bundled prompt-condition difference, since the instruction also changes) is not resolvable on any model (the largest, Gemini 3.1 Pro's -0.048, has interval [-0.096, -0.0003]); the sign flip reported for V3.2 in the earlier version ($+0.026$, [-0.004, 0.056]) is not resolvable at this sample size and we no longer report it as a finding.

Fresh product splits. Re-drawing the 7/3 product split for the same 200 participants with 5 further seeds (September 2026, V3.2, with demographics; Appendix J) gives protocol-minus-ICL differences from +0.197 to +0.254 (5 of 5 intervals exclude zero) and protocol-minus-warm-up differences from +0.017 to +0.046 (1 of 5 exclude zero): the gain over the single prompt is not specific to the fixed split, whereas the protocol's increment over the label-free conversation is small and resolvable in only one split.

D INTERVENTION DETAILS

Design. In the protocol, after each revealed label the model produces a free-text acknowledgment that stays in the history. The intervention replaces every free-text feedback acknowledgment with the fixed assistant turn "Acknowledged."; labels, roles and prompt rules are unchanged, while subsequent generated turns evolve naturally. The study plan (Appendix B) specified the primary contrast (replacement minus original, with demographics), two co-primary outcomes (Round-7 F_1 and the synchronized-error rate of Table 10), a joint success rule on the paired bootstrap intervals, and a repair gate requiring that the contemporaneous original protocol be below matched single-prompt ICL. DeepSeek V4-pro (first-party, thinking on) and Qwen-2.5-72B (pinned bf16 host) ran with interleaved contemporaneous controls and fresh single-prompt ICL.

Table 8: Intervention cells, September 2026, with demographics: Round-7 held-out F_1 with 95% intervals, and the synchronized-error rate (sync) of the original and replaced protocols; n = participants completed in the control arm. The thinking-off V4-pro row is a control-versus-ICL cell only (no replacement arm).

Model	n	block ICL	protocol (original)	protocol (replaced ack.)	Acc (ICL / orig. / repl.)	PPR (ICL / orig. / repl.)	sync orig.	sync repl.
DeepSeek V4-pro (GA, thinking on)	200	0.714 [0.669, 0.755]	0.692 [0.646, 0.733]	0.693 [0.649, 0.734]	0.722 / 0.673 / 0.678	0.517 / 0.605 / 0.593	0.014	0.018
Qwen-2.5-72B (pinned bf16 host)	200	0.705 [0.660, 0.747]	0.657 [0.611, 0.701]	0.655 [0.609, 0.698]	0.695 / 0.668 / 0.670	0.580 / 0.513 / 0.502	0.005	0.006
DeepSeek V4-pro, thinking off	200	0.686 [0.642, 0.726]	0.688 [0.644, 0.730]	–	0.612 / 0.698 / –	0.780 / 0.513 / –	0.007	–

Result on Qwen-2.5-72B. The original protocol underperformed contemporaneous single-prompt ICL (-0.048, [-0.082, -0.016]), satisfying the planned repair gate. Acknowledgment replacement produced neither a resolvable F_1 change (-0.002, [-0.030, 0.024]) nor a resolvable reduction in synchronized errors (+0.001, [-0.004, 0.007]); the replaced protocol also remains below ICL (-0.050, [-0.086, -0.016]). Under the planned decision rule, joint success and repair are not established.

Result on DeepSeek V4-pro. On the GA build served in September, the original protocol was not resolvably worse than contemporaneous single-prompt ICL (-0.022, [-0.054, 0.010]), so the planned repair gate was not met under the April parser; the April build's loss of -0.069 is not pooled with these cells, and these measurements do not establish a temporal change in harm or attribute one to the build change. The gate conclusion is parser-sensitive: under a first-digit rule, which does not favour the positive class when the GA build emits several digits, the same paired contrast is -0.044 ([-0.078, -0.011]); the replacement contrast is +0.002 ([-0.034, 0.039]) under that rule. Acknowledgment replacement produced neither a resolvable F_1 change (+0.001, [-0.030, 0.032], n=200) nor a resolvable change in synchronized errors (+0.003, [-0.006, 0.013]). Joint success and repair are therefore not established on either model. With V4-pro's thinking disabled (Table 8) control minus ICL is +0.003 ([-0.035, 0.041]), unresolved; without a paired interaction contrast this does not establish that thinking leaves the protocol effect unchanged.

Scoring-rule robustness. Because the GA build of V4-pro sometimes answers a held-out product with several digits, Table 16 (Appendix O) repeats the two intervention contrasts under three scoring rules: the April rule, a first-digit rule, and a strict rule that keeps only bare-digit responses present in both arms, which conditions on the generated outputs and changes the estimand to a contrast-specific

subset of decisions. The replacement contrast is unresolvable under every rule on both models, and Qwen's repair gate is met under every rule; V4-pro's gate is met only under the first-digit rule, so the planned conclusion for V4-pro depends on how multi-digit answers are scored and is reported as such.

The GA release of V4-pro already produces terse acknowledgments ("Got it. Ready for the next product."), so on that build the replacement removes little text.

E PANEL ACCURACY AND PREDICTED-POSITIVE RATES

Table 9: Accuracy / predicted-positive rate for every cell of Table 5 (prevalence 0.455). Cells without retained per-participant files are omitted.

	no demographics		with demographics	
Model	ICL Acc / PPR	protocol Acc / PPR	ICL Acc / PPR	protocol Acc / PPR
DeepSeek-V3.2 (Sep 2026)	0.690 / 0.258	0.683 / 0.562	0.660 / 0.195	0.702 / 0.537
Llama 3.3 70B	0.647 / 0.665	0.650 / 0.665	0.613 / 0.725	0.692 / 0.550
Claude Haiku 4.5	–	0.657 / 0.625	–	0.644 / 0.634
GPT-5 (R)	0.718 / 0.473	–	0.722 / 0.503	0.712 / 0.510
Qwen-2.5-72B	0.702 / 0.580	0.643 / 0.675	0.703 / 0.568	0.688 / 0.533
GPT-4o-mini	–	0.687 / 0.562	–	0.685 / 0.490
DeepSeek V4-pro (R)	–	0.653 / 0.515	–	0.653 / 0.462
Gemini 3.1 Pro (R)	0.720 / 0.448	0.700 / 0.442	0.750 / 0.378	0.690 / 0.355
Gemini 2.0 Flash	–	0.700 / 0.555	–	0.675 / 0.533

F TRAJECTORY DIAGNOSTICS TABLE AND LEARNING CURVES

A loss under the protocol could arise because the model treats each revealed label as a rule and re-labels the person's other decisions; we test this on the held-out trajectories of every panel model. For rounds $r = 2, \ldots, 7$, let L_r be the label revealed in round r. Among held-out predictions that differed from L_r in round $r-1$, we record the fraction that flip to L_r in round r (*toward*), the fraction of correct predictions that flip to an incorrect L_r (*correct→wrong toward label*), and, as a control, the fraction of predictions equal to L_r that flip away. We also report the Round-7 rate at which all three held-out predictions coincide, next to the rate at which the three true labels coincide; Figure 4 plots the label-directed error rate against each model's lift and Table 10 (Appendix F) gives all quantities, with eligibility counts and an interval for the label-directed error rate.

Two facts stand out. DeepSeek V4-pro is the only model whose Round-7 predictions coincide far more often than its labels do (0.530 vs. 0.36) and it has the highest label-directed error rate (0.222; Figure 4); its protocol loss is consistent with this candidate explanation. But the signature does not generalize: GPT-4o-mini has a lower protocol point estimate (paired uncertainty unavailable) and the *lowest* label-directed error rate (0.039), Claude Haiku's lift is near zero (no retained paired comparison) yet it has one of the highest (0.158), and Gemini 3.1 Pro's resolvable loss comes with a coincidence rate at the label baseline and a label-directed error rate comparable to V3.2's. Label-directed flipping is therefore a retrospective diagnostic consistent with V4-pro's loss that does not explain the other models'; §5 uses the synchronized-error rate as a planned co-primary outcome.

When does the gain appear? Figure 2 plots held-out F_1 after each context round. On V3.2 the protocol's Round-1 predictions, made after the persona turn and a single revealed label, already sit at the level of the warm-up-only conversation and far above single-prompt ICL; the paired Round-7-minus-Round-1 difference is +0.021 ([-0.018, 0.061]) for the protocol, -0.004 ([-0.031, 0.022]) for self-feedback and +0.035 ([-0.001, 0.072]) for direct-example. The gain over the single prompt is present from the first held-out prediction; later improvement is unresolved rather than excluded, and because Round 1 follows one revealed label this does not by itself establish the warm-up turn as the mechanism. On V4-pro and Qwen both the original and the replaced-acknowledgment protocol have Round-7 point estimates below single-prompt ICL (Round-7-minus-Round-1: +0.016 and +0.013 on V4-pro, +0.017 and +0.021 on Qwen); the plotted trajectories establish neither roundwise equivalence of the two arms nor a persistent separation from ICL.

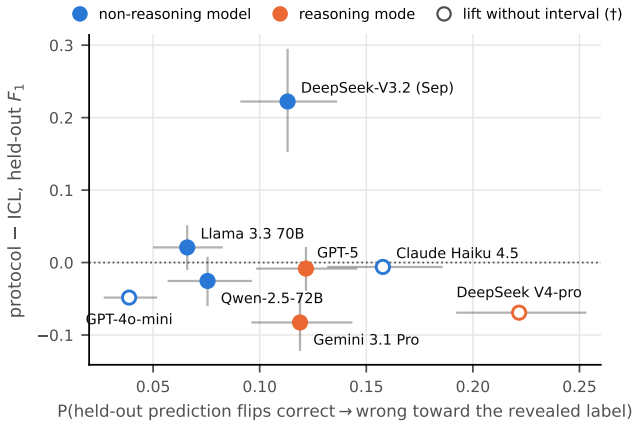

Figure 4: Rate at which a held-out prediction flips from correct to incorrect toward the label just revealed (with participant-bootstrap 95% interval) against each model's protocol-minus-ICL lift, with demographics. Hollow markers: lifts without an interval (†).

Table 10: Trajectory diagnostics on held-out predictions, with demographics; April panel except the September V3.2 row. Lift as in Table 5. n_{elig} = number of eligible (correct, differing from the revealed label) prediction-rounds behind the label-directed error rate. "sync" = fraction of participant-rounds with ≥ 2 correct→wrong flips toward the revealed label. "all-same" = fraction of participants whose three Round-7 predictions coincide; "labels" = the same for the true labels.

Model	lift	P(flip toward L_r)	P(flip away)	P(correct→wrong toward L_r)	95% CI	n_{elig}	sync	all-same	labels
DeepSeek-V3.2 (Sep 2026)	+0.222	0.156	0.043	0.113	[0.091, 0.136]	1123	0.012	0.285	0.360
Llama 3.3 70B	+0.021	0.112	0.065	0.066	[0.051, 0.082]	1121	0.004	0.280	0.360
Claude Haiku 4.5	-0.006†	0.197	0.073	0.158	[0.132, 0.185]	995	0.015	0.316	0.357
GPT-5 (R)	-0.008	0.161	0.069	0.122	[0.099, 0.145]	1101	0.009	0.345	0.360
Qwen-2.5-72B	-0.025	0.109	0.075	0.076	[0.057, 0.096]	1112	0.007	0.285	0.360
GPT-4o-mini	-0.048†	0.066	0.103	0.039	[0.027, 0.051]	1187	0.000	0.235	0.360
DeepSeek V4-pro (R)	-0.069†	0.288	0.231	0.222	[0.193, 0.253]	1051	0.022	0.530	0.360
Gemini 3.1 Pro (R)	-0.083	0.158	0.053	0.119	[0.097, 0.143]	1110	0.009	0.365	0.360
Gemini 2.0 Flash	–	0.114	0.100	0.082	[0.066, 0.100]	1091	0.005	0.305	0.360

G SUBGROUP ERROR AUDIT

Table 11: Error audit by income, political affiliation and region for the main V3.2 cells (held-out predictions): group size, prevalence, predicted-positive rate, false-positive and false-negative rates, F_1 with 95% interval.

Cell	Subgroup	n (participants)	prev.	PPR	FPR	FNR	F_1
V3.2 protocol with	income: $100,000 or more	50	0.493	0.607	0.447	0.230	0.691 [0.607, 0.767]
V3.2 protocol with	income: $30,000-$50,000	42	0.476	0.452	0.242	0.317	0.701 [0.593, 0.793]
V3.2 protocol with	income: $50,000-$75,000	49	0.442	0.503	0.268	0.200	0.748 [0.656, 0.824]
V3.2 protocol with	income: $75,000-$100,000	28	0.405	0.643	0.480	0.118	0.682 [0.552, 0.791]
V3.2 protocol with	income: Less than $30,000	31	0.430	0.495	0.340	0.300	0.651 [0.521, 0.762]
V3.2 protocol no	income: $100,000 or more	50	0.493	0.560	0.368	0.243	0.709 [0.621, 0.784]
V3.2 protocol no	income: $30,000-$50,000	42	0.476	0.563	0.409	0.267	0.672 [0.569, 0.761]
V3.2 protocol no	income: $50,000-$75,000	49	0.442	0.544	0.317	0.169	0.745 [0.652, 0.823]
V3.2 protocol no	income: $75,000-$100,000	28	0.405	0.619	0.480	0.176	0.651 [0.507, 0.773]
V3.2 protocol no	income: Less than $30,000	31	0.430	0.538	0.415	0.300	0.622 [0.488, 0.738]
V3.2 zero-shot with	income: $100,000 or more	50	0.493	0.287	0.224	0.649	0.444 [0.333, 0.545]
V3.2 zero-shot with	income: $30,000-$50,000	42	0.476	0.190	0.091	0.700	0.429 [0.282, 0.561]
V3.2 zero-shot with	income: $50,000-$75,000	49	0.442	0.238	0.098	0.585	0.540 [0.426, 0.640]
V3.2 zero-shot with	income: $75,000-$100,000	28	0.405	0.369	0.220	0.412	0.615 [0.475, 0.735]
V3.2 zero-shot with	income: Less than $30,000	31	0.430	0.065	0.000	0.850	0.261 [0.100, 0.415]
V3.2 ICL with	income: $100,000 or more	50	0.493	0.253	0.132	0.622	0.500 [0.373, 0.602]
V3.2 ICL with	income: $30,000-$50,000	42	0.476	0.159	0.030	0.700	0.450 [0.289, 0.597]
V3.2 ICL with	income: $50,000-$75,000	49	0.442	0.218	0.037	0.554	0.598 [0.444, 0.729]
V3.2 ICL with	income: $75,000-$100,000	28	0.405	0.190	0.140	0.735	0.360 [0.143, 0.562]
V3.2 ICL with	income: Less than $30,000	31	0.430	0.118	0.038	0.775	0.353 [0.186, 0.509]
V3.2 protocol with	politics: Democrat	86	0.419	0.543	0.373	0.222	0.677 [0.607, 0.741]
V3.2 protocol with	politics: Independent	53	0.465	0.522	0.318	0.243	0.713 [0.625, 0.789]
V3.2 protocol with	politics: Republican	52	0.506	0.551	0.351	0.253	0.715 [0.626, 0.793]
V3.2 protocol with	politics: Something else	9	0.444	0.481	0.267	0.250	0.720 [0.476, 0.889]
V3.2 protocol no	politics: Democrat	86	0.419	0.589	0.447	0.213	0.654 [0.579, 0.722]
V3.2 protocol no	politics: Independent	53	0.465	0.572	0.365	0.189	0.727 [0.645, 0.800]
V3.2 protocol no	politics: Republican	52	0.506	0.526	0.338	0.291	0.696 [0.602, 0.776]
V3.2 protocol no	politics: Something else	9	0.444	0.444	0.200	0.250	0.750 [0.556, 0.880]
V3.2 zero-shot with	politics: Democrat	86	0.419	0.244	0.107	0.565	0.550 [0.462, 0.629]
V3.2 zero-shot with	politics: Independent	53	0.465	0.189	0.094	0.703	0.423 [0.314, 0.525]
V3.2 zero-shot with	politics: Republican	52	0.506	0.263	0.195	0.671	0.433 [0.305, 0.545]
V3.2 zero-shot with	politics: Something else	9	0.444	0.185	0.200	0.833	0.235 [0.000, 0.471]
V3.2 ICL with	politics: Democrat	86	0.419	0.205	0.060	0.593	0.547 [0.433, 0.647]
V3.2 ICL with	politics: Independent	53	0.465	0.164	0.059	0.716	0.420 [0.280, 0.549]
V3.2 ICL with	politics: Republican	52	0.506	0.237	0.130	0.658	0.466 [0.333, 0.580]
V3.2 ICL with	politics: Something else	9	0.444	0.037	0.000	0.917	0.154 [0.000, 0.400]
V3.2 protocol with	region: Midwest	39	0.427	0.496	0.328	0.280	0.667 [0.562, 0.759]
V3.2 protocol with	region: Northeast	27	0.358	0.519	0.346	0.172	0.676 [0.536, 0.786]
V3.2 protocol with	region: Pacific	1	0.333	0.333	0.000	0.000	1.000 –
V3.2 protocol with	region: South	84	0.488	0.567	0.341	0.195	0.744 [0.682, 0.800]
V3.2 protocol with	region: West	49	0.476	0.531	0.390	0.314	0.649 [0.541, 0.740]
V3.2 protocol no	region: Midwest	39	0.427	0.521	0.328	0.220	0.703 [0.594, 0.797]
V3.2 protocol no	region: Northeast	27	0.358	0.642	0.519	0.138	0.617 [0.448, 0.756]
V3.2 protocol no	region: Pacific	1	0.333	0.667	0.500	0.000	0.667 –
V3.2 protocol no	region: South	84	0.488	0.587	0.395	0.211	0.716 [0.652, 0.774]
V3.2 protocol no	region: West	49	0.476	0.503	0.338	0.314	0.667 [0.569, 0.752]
V3.2 zero-shot with	region: Midwest	39	0.427	0.222	0.060	0.560	0.579 [0.438, 0.700]
V3.2 zero-shot with	region: Northeast	27	0.358	0.160	0.077	0.690	0.429 [0.216, 0.612]
V3.2 zero-shot with	region: Pacific	1	0.333	0.333	0.000	0.000	1.000 –
V3.2 zero-shot with	region: South	84	0.488	0.258	0.178	0.659	0.447 [0.360, 0.523]
V3.2 zero-shot with	region: West	49	0.476	0.231	0.143	0.671	0.442 [0.309, 0.562]
V3.2 ICL with	region: Midwest	39	0.427	0.197	0.030	0.580	0.575 [0.431, 0.698]
V3.2 ICL with	region: Northeast	27	0.358	0.111	0.077	0.828	0.263 [0.000, 0.478]
V3.2 ICL with	region: Pacific	1	0.333	0.333	0.000	0.000	1.000 –
V3.2 ICL with	region: South	84	0.488	0.226	0.078	0.618	0.522 [0.411, 0.619]
V3.2 ICL with	region: West	49	0.476	0.184	0.104	0.729	0.392 [0.247, 0.517]

H PERSONALIZATION AND INSTRUCTION CONTROLS

Instruction sensitivity of single-prompt prediction. With the 14-field persona, single-call zero-shot drops from 0.626 to 0.471 (-0.155, [-0.219, -0.091]) and block ICL reaches only 0.477. The no-demographics condition of Tables 4 and 5, however, also changes the system instruction ("a typical American shopper" instead of "THIS participant"), so the with-/without-demographics comparison bundles profile availability with an instruction change and does not identify the effect of supplying demographics. Removing the profile while retaining the instruction to use the provided persona

information gives zero-shot 0.077 (PPR 0.018; with-profile minus this cell +0.393, [0.325, 0.459]), ICL 0.400 (+0.077, [0.028, 0.129]) and protocol 0.669 (+0.030, [-0.005, 0.065]). This underspecified condition demonstrates prompt sensitivity, not a benefit attributable to demographic information; because the original comparison changed both profile availability and instruction, it does not identify whether demographics help or hurt on this model, and we withdraw the earlier demographic-penalty interpretation on that ground. Rendering the persona inside the system message instead of the user turn changes single-call zero-shot by +0.020 ([-0.031, 0.075]) and single-prompt ICL by +0.088 ([0.044, 0.135]); the protocol nevertheless retains a lead of +0.134 ([0.073, 0.195]) over that rendering.

Table 12: Personalization and prompt controls on DeepSeek-V3.2 (September 2026, same cohort): Round-7 held-out F_1 with participant-bootstrap 95% intervals, accuracy and predicted-positive rate.

Cell (DeepSeek-V3.2, September 2026)	F_1	95% CI	Acc	PPR	n
Conversational protocol, own persona (Table 3)	0.699	[0.653, 0.741]	0.702	0.537	200
Conversational protocol, another participant's persona	0.698	[0.652, 0.740]	0.700	0.538	200
Conversational protocol, another participant's context history	0.639	[0.591, 0.683]	0.655	0.500	200
Warm-up only, own persona (Table 3)	0.675	[0.633, 0.715]	0.627	0.695	200
Warm-up only, another participant's persona	0.667	[0.622, 0.708]	0.618	0.690	200
Conversational protocol, no persona, with-demo system prompt	0.669	[0.620, 0.715]	0.683	0.502	200
Single-call zero-shot, no persona, with-demo system prompt	0.077	[0.036, 0.122]	0.563	0.018	200
Block ICL, no persona, with-demo system prompt	0.400	[0.318, 0.476]	0.635	0.153	200
Single-call zero-shot, persona in system message	0.491	[0.429, 0.552]	0.658	0.217	200
Block ICL, persona in system message	0.565	[0.502, 0.624]	0.690	0.258	200

I PAIRED GAINS OVER THE MATCHED STATISTICAL BASELINES

Baseline construction. The two logistic regressions are fitted on the 1,858 Twin-2K-500 participants outside the frozen cohort; the collaborative filter's donor pool additionally excludes the fresh cohort of Appendix K (1,658 donors), and its 400 validation participants are drawn from that pool and excluded from their own donor sets. *Demographics-only*: one L1 logistic regression per product on the 14 one-hot demographic fields (C=1, default 0.5 threshold). *Demographics + product + history*: one L1 logistic regression on the demographic one-hots, a product one-hot and the participant's purchase rate over their other observed products (39 for training participants, the 7 context products for cohort participants); its decision threshold is the value on a 0.05 grid that maximizes F_1 on 400 held-out non-cohort participants (selected 0.35; validation F_1 0.649). *kNN collaborative filter*: donors are all non-cohort participants with 40 focal answers; similarity is the agreement rate on the participant's seven context labels; the score for a held-out product is the purchase rate among the k most similar donors; $k \in \{5, 10, 20, 50, 100, 200\}$ and threshold $\in \{0.30, \ldots, 0.60\}$ are chosen to maximize F_1 on 400 validation donors, each with a simulated 7/3 split and excluded from their own donor set (selected k=100, threshold 0.30; validation F_1 0.650). *Participant history*: the participant's own context purchase rate thresholded at 0.5. *Product base rate*: the non-cohort purchase rate of the product thresholded at 0.5.

J FRESH PRODUCT SPLITS

Table 13: Five further random 7/3 product splits for the same 200 participants (DeepSeek-V3.2, September 2026, with demographics): Round-7 held-out F_1 of the protocol, single-prompt ICL and the warm-up-only conversation, with paired participant-bootstrap 95% intervals for the two contrasts. Split labels are the first eight characters of each split file's hash.

Split	protocol F_1	ICL F_1	warm-up F_1	protocol − ICL	protocol − warm-up
split 180af2c2	0.701	0.482	0.678	+0.219 [0.155, 0.287]	+0.023 [-0.014, 0.059]
split 2c68ebb8	0.712	0.473	0.685	+0.240 [0.179, 0.303]	+0.027 [-0.009, 0.060]
split 73ad3ab4	0.679	0.425	0.662	+0.254 [0.187, 0.326]	+0.017 [-0.020, 0.055]
split 83cd1932	0.670	0.433	0.624	+0.237 [0.170, 0.308]	+0.046 [0.009, 0.081]
split a2a80ad5	0.712	0.515	0.685	+0.197 [0.138, 0.260]	+0.027 [-0.007, 0.061]

K FRESH COHORT

Table 14: Fresh cohort of 200 Twin-2K-500 participants disjoint from the frozen cohort (seed 20260924, split hash 272c6067; held-out prevalence 0.405), DeepSeek-V3.2 on the pinned host, September 2026: Round-7 held-out F_1 with participant-bootstrap 95% intervals, accuracy and predicted-positive rate.

Cell (fresh cohort, DeepSeek-V3.2, September 2026)	F_1	95% CI	Acc	PPR	n
Conversational protocol	0.705	[0.655, 0.750]	0.732	0.503	200
Warm-up only	0.636	[0.589, 0.682]	0.605	0.680	200
Single-prompt ICL (block format)	0.429	[0.355, 0.499]	0.680	0.155	200
Single-prompt ICL (dialogue format)	0.703	[0.656, 0.746]	0.725	0.520	200
Single-prompt ICL (block format), no profile	0.514	[0.444, 0.581]	0.692	0.230	200
Single-prompt ICL (dialogue format), no profile, with-demo system prompt	0.686	[0.634, 0.732]	0.730	0.455	200

L MULTIPLICITY ACCOUNTING

Every paired contrast for which a participant bootstrap was computed, with its point estimate and interval, a two-sided percentile-bootstrap p-value (floored at $1/B$) and the Benjamini–Hochberg adjusted value at q=0.05 computed over all reported contrasts, including the intervention contrasts and their synchronized-error outcomes, as one exploratory family. Contrasts reported without a computed bootstrap are not in this family: the April ICL cells without retained files. 4 of the 61 exploratory intervals that exclude zero do not survive the adjustment.

Table 15: Multiplicity accounting for every paired contrast with a computed participant bootstrap (continued across pages).

Family	Contrast	diff	95% CI	p_{boot}	p_{BH}	BH q=0.05
April panel / historical	V3.2 (Feb 2026) with-demo: protocol minus mirror-RL	+0.088	[0.047, 0.129]	<0.001	<0.001	yes
April panel / historical	Gemini 3.1 Pro (R) with-demo: protocol minus ICL	-0.083	[-0.120, -0.047]	<0.001	<0.001	yes
April panel / historical	V3.2 (Feb 2026) with-demo: protocol minus direct-example	+0.053	[0.014, 0.094]	0.008	0.021	yes
April panel / historical	V3.2 (Feb 2026) with-demo: protocol minus warm-up-only	+0.055	[0.016, 0.095]	0.009	0.021	yes
April panel / historical	V3.2 (Feb 2026 protocol / May 2026 self-feedback) with-demo: protocol minus self-feedback	+0.046	[0.009, 0.084]	0.015	0.035	yes
April panel / historical	Gemini 3.1 Pro (R) protocol: with-demo minus no-demo	-0.048	[-0.096, -0.0003]	0.049	0.104	no
April panel / historical	V3.2 (Feb 2026) protocol: with-demo minus no-demo	+0.026	[-0.004, 0.056]	0.089	0.173	no
April panel / historical	Qwen-2.5-72B no-demo: protocol minus ICL	-0.027	[-0.058, 0.005]	0.092	0.174	no
April panel / historical	Gemini 2.0 Flash protocol: with-demo minus no-demo	-0.032	[-0.070, 0.006]	0.097	0.179	no
April panel / historical	Qwen-2.5-72B with-demo: protocol minus ICL	-0.025	[-0.058, 0.006]	0.118	0.213	no
April panel / historical	GPT-4o-mini protocol: with-demo minus no-demo	-0.025	[-0.058, 0.007]	0.132	0.231	no
April panel / historical	Gemini 3.1 Pro (R) no-demo: protocol minus ICL	-0.025	[-0.058, 0.008]	0.134	0.232	no
April panel / historical	Llama 3.3 70B with-demo: protocol minus ICL	+0.021	[-0.009, 0.050]	0.158	0.263	no
April panel / historical	DeepSeek V4-pro (R) protocol: with-demo minus no-demo	-0.021	[-0.066, 0.023]	0.355	0.524	no
April panel / historical	GPT-5 (R) with-demo: protocol minus ICL	-0.008	[-0.037, 0.020]	0.584	0.697	no
April panel / historical	V3.2 (Feb 2026 warm-up-only / May 2026 self-feedback) with-demo: self-feedback minus warm-up-only	+0.009	[-0.028, 0.046]	0.618	0.718	no
April panel / historical	Claude Haiku 4.5 protocol: with-demo minus no-demo	-0.008	[-0.043, 0.026]	0.619	0.718	no
April panel / historical	Llama 3.3 70B protocol: with-demo minus no-demo	+0.006	[-0.027, 0.039]	0.723	0.832	no
April panel / historical	Llama 3.3 70B no-demo: protocol minus ICL	+0.003	[-0.023, 0.029]	0.844	0.938	no
April panel / historical	Qwen-2.5-72B protocol: with-demo minus no-demo	0.000	[-0.032, 0.033]	0.982	0.988	no
September (F1)	V3.2 with-demo: protocol minus ICL	+0.222	[0.154, 0.293]	<0.001	<0.001	yes
September (F1)	V3.2 with-demo: protocol minus zero-shot	+0.228	[0.165, 0.293]	<0.001	<0.001	yes
September (F1)	V3.2 with-demo: zero-shot minus no-demo zero-shot	-0.155	[-0.219, -0.091]	<0.001	<0.001	yes
September (F1)	V3.2 no-demo: protocol minus ICL	+0.123	[0.061, 0.186]	<0.001	<0.001	yes
September (F1)	V3.2 with-demo: shuffled-persona protocol minus ICL	+0.221	[0.154, 0.290]	<0.001	<0.001	yes
September (F1)	V3.2 zero-shot: with-demo minus no-demo (matched system prompt)	+0.393	[0.325, 0.459]	<0.001	<0.001	yes
September (F1)	V3.2 with-demo: ICL (persona in system) minus ICL	+0.088	[0.044, 0.135]	<0.001	<0.001	yes
September (F1)	V3.2 with-demo: protocol minus ICL (persona in system)	+0.134	[0.073, 0.195]	<0.001	<0.001	yes
September (F1)	V3.2 with-demo: protocol minus ICL [split 180af2c2]	+0.219	[0.155, 0.287]	<0.001	<0.001	yes
September (F1)	V3.2 with-demo: protocol minus ICL [split 2c68ebb8]	+0.240	[0.179, 0.303]	<0.001	<0.001	yes
September (F1)	V3.2 with-demo: protocol minus ICL [split 73ad3ab4]	+0.254	[0.187, 0.326]	<0.001	<0.001	yes
September (F1)	V3.2 with-demo: protocol minus ICL [split 83cd1932]	+0.237	[0.170, 0.308]	<0.001	<0.001	yes
September (F1)	V3.2 with-demo: protocol minus ICL [split a2a80ad5]	+0.197	[0.138, 0.260]	<0.001	<0.001	yes
September (F1)	V3.2 with-demo: dialogue ICL minus ICL qablock	+0.311	[0.241, 0.382]	<0.001	<0.001	yes
September (F1)	V3.2 with-demo: dialogue ICL minus ICL [pooled cohorts]	+0.245	[0.194, 0.298]	<0.001	<0.001	yes
September (F1)	V3.2 with-demo: protocol minus ICL [fresh]	+0.276	[0.201, 0.350]	<0.001	<0.001	yes
September (F1)	V3.2 with-demo: dialogue ICL minus ICL [fresh]	+0.274	[0.199, 0.350]	<0.001	<0.001	yes
September (F1)	V3.2 with-demo: dialogue ICL minus ICL	+0.220	[0.151, 0.293]	<0.001	<0.001	yes
September (F1)	V3.2 with-demo: dialogue ICL minus ICL (persona in system)	+0.132	[0.073, 0.194]	<0.001	<0.001	yes
September (F1)	V3.2 no-demo (matched system prompt): dialogue ICL minus ICL	+0.317	[0.244, 0.392]	<0.001	<0.001	yes
September (F1)	V3.2 protocol with minus logistic regression	+0.128	[0.071, 0.186]	<0.001	<0.001	yes
September (F1)	V3.2 protocol with minus participant history	+0.171	[0.107, 0.237]	<0.001	<0.001	yes
September (F1)	V3.2 protocol with minus product base rate	+0.210	[0.155, 0.267]	<0.001	<0.001	yes
September (F1)	V3.2 warm-up with minus participant history	+0.147	[0.082, 0.216]	<0.001	<0.001	yes
September (F1)	V3.2 warm-up with minus product base rate	+0.187	[0.123, 0.250]	<0.001	<0.001	yes
September (F1)	V3.2 zero-shot with minus logistic regression + history (tuned)	-0.155	[-0.227, -0.085]	<0.001	<0.001	yes
September (F1)	V3.2 zero-shot with minus kNN collaborative filtering (tuned)	-0.181	[-0.253, -0.111]	<0.001	<0.001	yes
September (F1)	V3.2 zero-shot with minus always purchase	-0.155	[-0.223, -0.085]	<0.001	<0.001	yes
September (F1)	V3.2 ICL with minus logistic regression + history (tuned)	-0.149	[-0.223, -0.077]	<0.001	<0.001	yes

Table 15 (continued)

Family	Contrast	diff	95% CI	p_{boot}	p_{BH}	BH q=0.05
September (F1)	V3.2 ICL with minus kNN collaborative filtering (tuned)	-0.175	[-0.252, -0.099]	<0.001	<0.001	yes
September (F1)	V3.2 ICL with minus always purchase	-0.149	[-0.228, -0.071]	<0.001	<0.001	yes
September (F1)	V3.2 with-demo: protocol minus warm-up-only [fresh]	+0.069	[0.032, 0.105]	<0.001	0.001	yes
September (F1)	V3.2 warm-up with minus logistic regression	+0.104	[0.051, 0.160]	<0.001	0.001	yes
September (F1)	V3.2 with-demo: protocol minus warm-up-only [pooled cohorts]	+0.046	[0.019, 0.072]	<0.001	0.003	yes
September (F1)	V3.2 protocol with minus logistic regression + history (tuned)	+0.073	[0.030, 0.117]	0.001	0.003	yes
September (F1)	V3.2 protocol with minus always purchase	+0.074	[0.033, 0.115]	0.001	0.003	yes
September (F1)	V3.2 warm-up with minus always purchase	+0.050	[0.020, 0.080]	0.001	0.004	yes
September (F1)	V3.2 zero-shot with minus logistic regression	-0.100	[-0.164, -0.037]	0.001	0.004	yes
September (F1)	V3.2 ICL: with-demo minus no-demo (matched system prompt)	+0.077	[0.028, 0.129]	0.002	0.005	yes
September (F1)	Qwen with-demo: control minus ICL (repair gate)	-0.048	[-0.082, -0.016]	0.003	0.008	yes
September (F1)	Qwen with-demo: replaced-ack minus ICL	-0.050	[-0.086, -0.016]	0.003	0.009	yes
September (F1)	V3.2 with-demo: protocol minus shuffled-history protocol	+0.060	[0.020, 0.102]	0.004	0.011	yes
September (F1)	Qwen with-demo: dialogue ICL minus ICL	-0.038	[-0.066, -0.010]	0.008	0.021	yes
September (F1)	V3.2 with-demo: ICL qablock minus ICL	-0.090	[-0.162, -0.018]	0.014	0.034	yes
September (F1)	V3.2 with-demo: protocol minus warm-up-only [split 83cd1932]	+0.046	[0.009, 0.081]	0.014	0.034	yes
September (F1)	V3.2 ICL with minus logistic regression	-0.094	[-0.172, -0.016]	0.018	0.041	yes
September (F1)	V3.2 protocol with minus kNN collaborative filtering (tuned)	+0.047	[0.008, 0.087]	0.019	0.042	yes
September (F1)	V3.2 warm-up with minus logistic regression + history (tuned)	+0.049	[0.004, 0.095]	0.034	0.074	no
September (F1)	Qwen with-demo: ICL qablock minus ICL	-0.024	[-0.047, -0.002]	0.036	0.077	no
September (F1)	V3.2 direct-example with: round 7 minus round 1	+0.035	[-0.001, 0.072]	0.058	0.120	no
September (F1)	V3.2 with-demo: protocol minus self-feedback	+0.035	[-0.002, 0.072]	0.062	0.124	no
September (F1)	V3.2 protocol: with-demo minus no-demo (matched system prompt)	+0.030	[-0.005, 0.065]	0.089	0.173	no
September (F1)	Haiku with-demo: dialogue ICL minus ICL	-0.025	[-0.054, 0.004]	0.091	0.174	no
September (F1)	V4-pro nothink with-demo: control minus dialogue ICL	+0.022	[-0.005, 0.048]	0.104	0.190	no
September (F1)	V3.2 with-demo: protocol minus warm-up-only [split a2a80ad5]	+0.027	[-0.007, 0.061]	0.122	0.218	no
September (F1)	V3.2 with-demo: protocol minus warm-up-only [split 2c68ebb8]	+0.027	[-0.009, 0.060]	0.143	0.245	no
September (F1)	V3.2 ICL with minus participant history	-0.051	[-0.124, 0.019]	0.155	0.262	no
September (F1)	V4-pro with-demo: control minus ICL (repair gate)	-0.022	[-0.054, 0.010]	0.184	0.300	no
September (F1)	V3.2 warm-up with minus kNN collaborative filtering (tuned)	+0.023	[-0.012, 0.058]	0.185	0.300	no
September (F1)	V3.2 zero-shot with minus participant history	-0.057	[-0.141, 0.028]	0.187	0.300	no
September (F1)	V4-pro with-demo: replaced-ack minus ICL	-0.020	[-0.052, 0.011]	0.199	0.315	no
September (F1)	Qwen replaced-ack with: round 7 minus round 1	+0.021	[-0.011, 0.052]	0.204	0.319	no
September (F1)	V3.2 with-demo: protocol minus warm-up-only [split 180af2c2]	+0.023	[-0.014, 0.059]	0.215	0.332	no
September (F1)	V3.2 with-demo: protocol minus warm-up-only	+0.024	[-0.013, 0.061]	0.219	0.335	no
September (F1)	V3.2 protocol with: round 7 minus round 1	+0.021	[-0.018, 0.061]	0.290	0.435	no
September (F1)	Qwen protocol with: round 7 minus round 1	+0.017	[-0.015, 0.049]	0.291	0.435	no
September (F1)	V3.2 with-demo: protocol minus warm-up-only [split 73ad3ab4]	+0.017	[-0.020, 0.055]	0.369	0.533	no
September (F1)	V4-pro nothink with-demo: dialogue ICL minus ICL	-0.019	[-0.062, 0.023]	0.381	0.544	no
September (F1)	V3.2 with-demo: dialogue ICL rerun minus dialogue ICL	-0.007	[-0.026, 0.009]	0.412	0.580	no
September (F1)	V3.2 with-demo: protocol minus direct-example	-0.012	[-0.042, 0.018]	0.416	0.580	no
September (F1)	V4-pro protocol with: round 7 minus round 1	+0.016	[-0.023, 0.054]	0.431	0.580	no
September (F1)	V3.2 with-demo: zero-shot (persona in system) minus zero-shot	+0.020	[-0.031, 0.075]	0.431	0.580	no
September (F1)	V3.2 with-demo: self-feedback minus warm-up-only	-0.011	[-0.039, 0.016]	0.432	0.580	no
September (F1)	V4-pro with-demo: control minus dialogue ICL	-0.013	[-0.044, 0.019]	0.432	0.580	no
September (F1)	V3.2 no-demo: protocol minus direct-example	+0.013	[-0.022, 0.049]	0.442	0.584	no
September (F1)	Qwen with-demo: dialogue ICL rerun minus dialogue ICL	-0.006	[-0.019, 0.008]	0.445	0.584	no
September (F1)	V4-pro with-demo: dialogue ICL minus ICL	-0.009	[-0.034, 0.016]	0.472	0.610	no
September (F1)	V4-pro replaced-ack with: round 7 minus round 1	+0.013	[-0.022, 0.049]	0.477	0.610	no
September (F1)	V3.2 protocol: with-demo minus no-demo	+0.011	[-0.020, 0.042]	0.511	0.645	no
September (F1)	Qwen with-demo: control minus dialogue ICL	-0.010	[-0.042, 0.022]	0.536	0.670	no
September (F1)	V3.2 with-demo: warm-up-only minus shuffled-persona warm-up-only	+0.009	[-0.021, 0.037]	0.569	0.697	no
September (F1)	V3.2 with-demo: protocol rerun minus dialogue ICL rerun	+0.009	[-0.023, 0.040]	0.581	0.697	no
September (F1)	V3.2 no-demo: protocol minus warm-up-only	+0.010	[-0.026, 0.046]	0.585	0.697	no
September (F1)	V3.2 zero-shot with minus product base rate	-0.018	[-0.084, 0.049]	0.596	0.704	no
September (F1)	V3.2 ICL with minus product base rate	-0.012	[-0.089, 0.064]	0.755	0.857	no
September (F1)	V3.2 self-feedback with: round 7 minus round 1	-0.004	[-0.031, 0.022]	0.758	0.857	no
September (F1)	V3.2 with-demo: 50-answer persona protocol minus protocol (persona richness)	-0.004	[-0.033, 0.023]	0.772	0.865	no
September (F1)	V3.2 protocol no: round 7 minus round 1	-0.004	[-0.044, 0.036]	0.860	0.946	no
September (F1)	V4-pro with-demo: thinking-off control minus thinking-off ICL	+0.003	[-0.035, 0.041]	0.882	0.956	no
September (F1)	V3.2 with-demo: protocol minus dialogue ICL [pooled cohorts]	+0.002	[-0.023, 0.026]	0.890	0.956	no
September (F1)	V3.2 with-demo: protocol minus dialogue ICL	+0.002	[-0.031, 0.033]	0.910	0.964	no
September (F1)	V3.2 with-demo: 50-answer persona warm-up minus warm-up	-0.002	[-0.027, 0.022]	0.912	0.964	no
September (F1)	V3.2 with-demo: protocol minus dialogue ICL [fresh]	+0.002	[-0.037, 0.038]	0.941	0.978	no
September (F1)	V3.2 with-demo: protocol minus shuffled-persona protocol	+0.001	[-0.030, 0.033]	0.948	0.978	no
September (F1)	Llama with-demo: dialogue ICL minus ICL	-0.001	[-0.025, 0.024]	0.964	0.986	no
September (F1)	V3.2 with-demo: protocol rerun minus protocol	0.000	[-0.024, 0.023]	0.988	0.988	no
intervention	V4-pro with-demo: replaced-ack minus control (synchronized-error rate)	+0.003	[-0.006, 0.013]	0.547	0.677	no
intervention	Qwen with-demo: replaced-ack minus control	-0.002	[-0.030, 0.024]	0.866	0.946	no
intervention	V4-pro with-demo: replaced-ack minus control	+0.001	[-0.030, 0.032]	0.921	0.966	no
intervention	Qwen with-demo: replaced-ack minus control (synchronized-error rate)	+0.001	[-0.004, 0.007]	0.971	0.986	no
slate choice (accuracy)	Amazon: protocol minus self-feedback (with demo) [flash]	+0.112	[0.068, 0.155]	<0.001	<0.001	yes
slate choice (accuracy)	H&M: protocol minus self-feedback (with demo) [flash]	+0.135	[0.080, 0.188]	<0.001	<0.001	yes
slate choice (accuracy)	H&M: self-feedback minus ICL (with demo) [flash]	-0.102	[-0.155, -0.050]	<0.001	<0.001	yes
slate choice (accuracy)	Amazon: self-feedback minus ICL (with demo) [flash]	-0.098	[-0.145, -0.053]	<0.001	<0.001	yes
slate choice (accuracy)	H&M: direct-example minus ICL (with demo) [flash]	+0.052	[0.013, 0.090]	0.008	0.021	yes
slate choice (accuracy)	Amazon: direct-example minus ICL (with demo) [flash]	+0.042	[0.005, 0.077]	0.026	0.057	no
slate choice (accuracy)	Amazon: protocol minus direct-example (with demo) [flash]	-0.028	[-0.057, 0.000]	0.056	0.117	no
slate choice (accuracy)	H&M: protocol minus ICL (with demo) [flash]	+0.033	[0.000, 0.068]	0.060	0.122	no
slate choice (accuracy)	H&M: protocol minus direct-example (with demo) [flash]	-0.018	[-0.055, 0.018]	0.359	0.525	no
slate choice (accuracy)	Amazon: protocol minus ICL (with demo) [flash]	+0.013	[-0.022, 0.048]	0.479	0.610	no

M CROSS-DOMAIN TASK EXTENSIONS

Two revealed-preference panels are reformulated as $K=4$ slate choice (one true purchase plus three distractors from the same taxonomy, year and price band; correct letter uniform across positions): Amazon-Study (Berke et al., 2024), 200 users, and the H&M Personalized Fashion Recommendations panel (H&M Group, 2022), 200 users, each with 7 context and 3 held-out events. The demographic schemas differ from Twin-2K-500, and these panels do not carry the personalization controls of §4. Table 6 reports the September-2026 four-arm run on `deepseek-flash` (V4.1).

N PROMPT TEMPLATES AND EXCERPTS

The turns below are the templates of the released runners, with product and persona blocks abbreviated by ellipses.

System prompt (with demographics).

```
You are an expert at predicting consumer purchase behavior for American
grocery shoppers. Analyze the product and price, and decide whether THIS
participant would purchase it, using the provided persona information about
their characteristics and preferences. Base your decision on what this
person would realistically do. All prices are in US dollars (USD).
```

Without demographics: "You are an expert at predicting consumer purchase behavior for American shoppers. Analyze the product and price, then decide if a typical American shopper would purchase it. All prices are in US dollars (USD)."

Protocol warm-up turn.

```
Persona Profile:
Q: Which part of the United States do you currently live in?
A: <answer>
... (14 demographic Q/A pairs) ...

I will show you a series of products. For each product, first tell me
whether you would purchase it (answer only 0 or 1), and then I will tell
you what you actually did. After each learning example, I will ask you
to predict some test products to see how well you've learned. Learn from
the feedback to improve your predictions.

Ready to begin?
```

Prediction turn (context and held-out).

```
Product:
- Category: <category>
- Product Name: <name>
- Price: $<price>.

Would you purchase this product? Answer 1 for Yes (would purchase),
0 for No (would not purchase). Answer only 0 or 1.
```

Feedback turn and acknowledgment. "Correct answer: You purchased (1) this product." / "Correct answer: You did not purchase (0) this product." The model's free-text reply is kept in the history. Under the intervention of §5 the reply is not generated; the fixed assistant turn "Acknowledged." is inserted instead.

Held-out phase and transition. "Now, based on what you've learned so far, predict whether you would purchase these test products:" followed by the three prediction turns, all removed from the history before "Good. Let's continue with the next learning example."

Warm-up-only. Same system prompt. First turn: "Persona Profile: … I will show you a series of products. For each product, tell me whether you would purchase it (answer only 0 or 1). Ready to begin?"; the seven context products are asked in sequence with the transition "Next product." and no feedback; then "Now predict whether you would purchase these products:" and the three held-out products.

Self-feedback. Identical to the protocol except for the warm-up sentence ("… first tell me whether you would purchase it (answer only 0 or 1), and then I will ask you to reflect on that prediction. After each example, I will ask you to predict some test products to see how well your reasoning is converging. Use your own self-reflection to refine your predictions.") and the feedback turn, which never reveals the label: "Take a moment to reflect on that prediction. Looking at the persona profile, are you confident your answer matches what this person would actually do, or would you reconsider?"

Direct-example. Same system prompt. First turn: "Persona Profile: … I will show you examples of this participant's purchase decisions. After each example, I will ask you to predict whether they would purchase some test products. Use the examples to learn their preferences. Ready to begin?" Each context turn: "Here is an example of this participant's purchase behavior: [product] The participant purchased / did not purchase this product." followed by "Now predict whether you would purchase these test products:" and the three held-out products; transition: "Good. Let's continue with the next example." No prediction is requested before the label is shown.

Dialogue-format single-call ICL. One API request. System message as for single-call ICL (with the persona appended to it only in the persona-in-system condition; otherwise the persona opens the first user turn). First user turn: "[Persona Profile: …] I will show you this participant's past purchase decisions as a series of products, each followed by the participant's actual decision. Then I will ask about a new product." Fixed assistant turn: "Understood." Then, for each of the seven context products in the protocol's order, a user turn "Product: … Would you purchase this product? Answer 1 for Yes (would purchase), 0 for No (would not purchase). Answer only 0 or 1." followed by a fixed assistant turn containing the participant's actual decision as a bare digit ("1" or "0"). Final user turn: "Now a new product. [held-out product] Would you purchase this product? … Answer only 0 or 1." The no-profile cells use the with-demographics system prompt. The same one-shot "Please answer with just a single digit" nudge as block ICL applies to unparseable responses; no response required it in any dialogue cell.

Single-call ICL. One user message containing the persona block (when used), "Here are 7 prior purchase decisions by this participant:" with the seven products rendered as above each followed by "Decision: purchased (1)" or "Decision: did not purchase (0)", then "Now predict the participant's decision for the following new product." and the prediction turn. Single-call zero-shot omits the examples.

O ENDPOINTS, DATES AND RE-SCORING

Table 16: The two intervention contrasts under three scoring rules (paired participant-bootstrap 95% intervals; "gate" = whether control is resolvably below single-prompt ICL). Strict-rule contrasts condition on jointly bare-digit outputs in both arms, use contrast-specific subsets (n_d = retained decisions out of 600), and do not estimate the full-cohort effect.

	DeepSeek V4-pro (GA)			Qwen-2.5-72B		
Scoring rule	control − ICL	gate	replaced − control	control − ICL	gate	replaced − control
April rule	-0.022 [-0.054, 0.010]	not met	+0.001 [-0.030, 0.032]	-0.048 [-0.082, -0.016]	met	-0.002 [-0.030, 0.024]
first digit	-0.044 [-0.078, -0.011]	met	+0.002 [-0.034, 0.039]	-0.049 [-0.083, -0.017]	met	-0.001 [-0.028, 0.025]
strict, common bare-digit decisions	-0.021 [-0.055, 0.013] (n_d=489)	not met	+0.005 [-0.033, 0.041] (n_d=413)	-0.049 [-0.082, -0.017] (n_d=594)	met	+0.001 [-0.025, 0.027] (n_d=593)

Table 18 lists, for every September-2026 cell including the slate-choice four-arm run, the requested model identifier, endpoint, serving provider as returned per request, thinking setting, first-call date, number of calls and the distribution of finish reasons; the April-2026 panel cells were run against first-party endpoints (OpenAI, Anthropic, Google, DeepSeek) and OpenRouter (Llama, Qwen,

V3.2) as recorded in their result folders. Every September prediction row carries the raw model response. Table 17 re-scores each September cell under a first-digit rule and under a strict rule that counts any response other than a bare "0" or "1" (or its numeric rendering "0.0"/"1.0") as missing; the April rule (any "1" scores 1) was retained online so that the trajectories match the April panel.

Table 17: Strict-binary re-scoring of the September-2026 cells (Round-7 held-out predictions).

Cell	n	F_1 (April rule: any 1)	F_1 (first digit)	bare-digit fraction	dropped	F_1 (strict, dropped)
V4-pro protocol with	200	0.692	0.670	0.815	111	0.694
V4-pro replaced-ack with	200	0.693	0.672	0.823	106	0.701
V4-pro ICL with	200	0.714	0.714	1.000	0	0.714
V4-pro ICL no	200	0.726	0.726	1.000	0	0.726
V4-pro nothink protocol with	200	0.688	0.688	1.000	0	0.688
V4-pro nothink ICL with	200	0.686	0.686	1.000	0	0.686
Qwen protocol with	200	0.657	0.656	0.990	6	0.659
Qwen replaced-ack with	200	0.655	0.655	0.998	1	0.656
Qwen ICL with	200	0.705	0.705	1.000	0	0.705
Qwen ICL no	200	0.716	0.716	1.000	0	0.716
V3.2 protocol with	200	0.699	0.699	1.000	0	0.699
V3.2 protocol 50-answer persona with	200	0.695	0.695	1.000	0	0.695
V3.2 protocol no	200	0.689	0.689	1.000	0	0.689
V3.2 warm-up with	200	0.675	0.675	1.000	0	0.675
V3.2 warm-up 50-answer persona with	200	0.674	0.674	1.000	0	0.674
V3.2 warm-up no	200	0.678	0.678	1.000	0	0.678
V3.2 self-feedback with	200	0.665	0.665	1.000	0	0.665
V3.2 self-feedback no	200	0.665	0.665	1.000	0	0.665
V3.2 direct-example with	200	0.712	0.712	1.000	0	0.712
V3.2 direct-example no	200	0.675	0.675	1.000	0	0.675
V3.2 ICL with	200	0.477	0.477	1.000	0	0.477
V3.2 ICL no	200	0.565	0.565	1.000	0	0.565
V3.2 zero-shot with	200	0.471	0.471	1.000	0	0.471
V3.2 zero-shot no	200	0.626	0.626	1.000	0	0.626
V3.2 protocol shuffled-persona with	200	0.698	0.698	1.000	0	0.698
V3.2 warm-up shuffled-persona with	200	0.667	0.667	1.000	0	0.667
V3.2 protocol shuffled-history with	200	0.639	0.639	1.000	0	0.639
V3.2 protocol no (matched system prompt)	200	0.669	0.669	1.000	0	0.669
V3.2 zero-shot no (matched system prompt)	200	0.077	0.077	1.000	0	0.077
V3.2 ICL no (matched system prompt)	200	0.400	0.400	1.000	0	0.400
V3.2 zero-shot with (persona in system)	200	0.491	0.491	1.000	0	0.491
V3.2 ICL with (persona in system)	200	0.565	0.565	1.000	0	0.565
V3.2 protocol with [fresh]	200	0.705	0.705	1.000	0	0.705
V3.2 warm-up with [fresh]	200	0.636	0.636	1.000	0	0.636
V3.2 ICL with [fresh]	200	0.429	0.429	1.000	0	0.429
V3.2 ICL no [fresh]	200	0.514	0.514	1.000	0	0.514
V3.2 ICL dialogue with [fresh]	200	0.703	0.703	1.000	0	0.703
V3.2 ICL dialogue no (matched system prompt) [fresh]	200	0.686	0.686	1.000	0	0.686
Qwen ICL dialogue with	200	0.668	0.668	1.000	0	0.668
Qwen ICL dialogue no (matched system prompt)	200	0.685	0.685	1.000	0	0.685
V3.2 ICL dialogue with	200	0.697	0.697	1.000	0	0.697
V3.2 ICL dialogue no (matched system prompt)	200	0.717	0.717	1.000	0	0.717
V3.2 protocol with [rerun]	200	0.699	0.699	1.000	0	0.699
V3.2 ICL dialogue with [rerun]	200	0.690	0.690	1.000	0	0.690
Qwen ICL dialogue with [rerun]	200	0.662	0.662	1.000	0	0.662
Llama ICL dialogue with	200	0.673	0.673	1.000	0	0.673
Llama ICL dialogue no (matched system prompt)	200	0.685	0.685	1.000	0	0.685
Haiku ICL dialogue with	200	0.663	0.663	1.000	0	0.663
Haiku ICL dialogue no (matched system prompt)	200	0.673	0.673	0.998	1	0.674
V4-pro ICL dialogue with	200	0.705	0.705	1.000	0	0.705
V4-pro ICL dialogue no (matched system prompt)	200	0.702	0.702	1.000	0	0.702
V4-pro nothink ICL dialogue with	200	0.667	0.667	1.000	0	0.667
V4-pro nothink ICL dialogue no (matched system prompt)	200	0.663	0.663	1.000	0	0.663
Llama ICL with	200	0.673	0.673	1.000	0	0.673
Llama ICL no	200	0.679	0.679	1.000	0	0.679
Haiku ICL with	200	0.688	0.688	1.000	0	0.688
Haiku ICL no	200	0.696	0.696	1.000	0	0.696
V3.2 ICL qablock with	200	0.387	0.387	1.000	0	0.387
V3.2 ICL qablock no (matched system prompt)	200	0.351	0.351	1.000	0	0.351
Qwen ICL qablock with	200	0.681	0.681	0.993	4	0.686
V3.2 protocol with [split 180af2c2]	200	0.701	0.701	1.000	0	0.701
V3.2 warm-up with [split 180af2c2]	200	0.678	0.678	1.000	0	0.678
V3.2 ICL with [split 180af2c2]	200	0.482	0.482	1.000	0	0.482
V3.2 protocol with [split 2c68ebb8]	200	0.712	0.712	1.000	0	0.712
V3.2 warm-up with [split 2c68ebb8]	200	0.685	0.685	1.000	0	0.685
V3.2 ICL with [split 2c68ebb8]	200	0.473	0.473	1.000	0	0.473
V3.2 protocol with [split 73ad3ab4]	200	0.679	0.679	1.000	0	0.679
V3.2 warm-up with [split 73ad3ab4]	200	0.662	0.662	1.000	0	0.662
V3.2 ICL with [split 73ad3ab4]	200	0.425	0.425	1.000	0	0.425
V3.2 protocol with [split 83cd1932]	200	0.670	0.670	1.000	0	0.670
V3.2 warm-up with [split 83cd1932]	200	0.624	0.624	1.000	0	0.624
V3.2 ICL with [split 83cd1932]	200	0.433	0.433	1.000	0	0.433
V3.2 protocol with [split a2a80ad5]	200	0.712	0.712	1.000	0	0.712
V3.2 warm-up with [split a2a80ad5]	200	0.685	0.685	1.000	0	0.685
V3.2 ICL with [split a2a80ad5]	200	0.515	0.515	1.000	0	0.515

Table 18: Endpoint provenance of the September-2026 cells (generated from per-request metadata). *: the self-feedback runner recorded call counts and the launch date only, not per-request finish reasons.

Cell	requested model	endpoint	provider	thinking	date (UTC)	calls	finish reasons
baseline_deepseek-v3.2	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-23	4000	stop:4000
baseline_deepseek-v3.2_personainsystem	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-23	4000	stop:4000
baseline_deepseek-v3.2_sysmatched	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-23	4000	stop:4000
direct_example_deepseek-v3.2_no_demo	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-23	8400	length:7, stop:8393
direct_example_deepseek-v3.2_with_demo	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-23	8400	stop:8400
icl_claude-haiku-openrouter	anthropic/claude-haiku-4.5	OpenRouter	Anthropic		2026-09-24	1200	stop:1200
icl_claude-haiku-openrouter_sysmatched_dialogue	anthropic/claude-haiku-4.5	OpenRouter	Anthropic		2026-09-24	1201	length:2, stop:1199
icl_deepseek-v3.2	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-23	1200	stop:1200
icl_deepseek-v3.2_personainsystem	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	1200	stop:1200
icl_deepseek-v3.2_split180af2c2	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	1200	stop:1200
icl_deepseek-v3.2_split272c6067 [fresh cohort]	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	1200	stop:1200
icl_deepseek-v3.2_split2c68ebb8	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	1200	stop:1200
icl_deepseek-v3.2_split73ad3ab4	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	1200	stop:1200
icl_deepseek-v3.2_split83cd1932	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	1200	stop:1200
icl_deepseek-v3.2_splita2a80ad5	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	1200	stop:1200
icl_deepseek-v3.2_sysmatched_dialogue	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	1200	stop:1200
icl_deepseek-v3.2_sysmatched_dialogue [repeat]	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	1200	stop:1200
icl_deepseek-v3.2_sysmatched_dialogue_split272c6067 [fresh cohort]	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	1200	stop:1200
icl_deepseek-v3.2_sysmatched	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	1200	stop:1200
icl_deepseek-v3.2_sysmatched_qablock	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	1200	stop:1200
icl_deepseek-v4-pro-nothink	deepseek-v4-pro	DeepSeek	first-party	disabled	2026-09-23	1200	stop:1200
icl_deepseek-v4-pro-nothink_sysmatched_dialogue	deepseek-v4-pro	DeepSeek	first-party	disabled	2026-09-24	1200	stop:1200
icl_deepseek-v4-pro	deepseek-v4-pro	DeepSeek	first-party	enabled	2026-09-23	1205	stop:1205
icl_deepseek-v4-pro_sysmatched_dialogue	deepseek-v4-pro	DeepSeek	first-party	enabled	2026-09-24	1200	stop:1200
icl_llama-3.3-70b	meta-llama/llama-3.3-70b-instruct	OpenRouter	DeepInfra / Novita / AkashML		2026-09-24	1202	length:2, stop:1200
icl_llama-3.3-70b_sysmatched_dialogue	meta-llama/llama-3.3-70b-instruct	OpenRouter	DeepInfra / Novita / AkashML		2026-09-24	1200	stop:1200
icl_qwen-2.5-72b-openrouter	qwen/qwen-2.5-72b-instruct	OpenRouter	Novita		2026-09-23	1210	stop:1210
icl_qwen-2.5-72b-openrouter_sysmatched_dialogue	qwen/qwen-2.5-72b-instruct	OpenRouter	Novita		2026-09-24	1200	stop:1200
icl_qwen-2.5-72b-openrouter_sysmatched_dialogue [repeat]	qwen/qwen-2.5-72b-instruct	OpenRouter	Novita		2026-09-24	1200	stop:1200
icl_qwen-2.5-72b-openrouter_sysmatched_qablock	qwen/qwen-2.5-72b-instruct	OpenRouter	Novita		2026-09-24	1200	length:4, stop:1196
rl_self_feedback_deepseek-v3.2-openrouter_no_demo	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-23*	9800	–*
rl_self_feedback_deepseek-v3.2-openrouter_with_demo	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-23*	9800	–*
rl_v1_deepseek-v3.2_no_demo	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-23	9800	stop:9800
rl_v1_deepseek-v3.2_no_demo_sysmatched	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-23	9800	stop:9800
rl_v1_deepseek-v3.2_with_demo_feb_bullets50	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-23	9800	stop:9800
rl_v1_deepseek-v3.2_with_demo	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-23	9800	stop:9800
rl_v1_deepseek-v3.2_with_demo [repeat]	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	9800	stop:9800
rl_v1_deepseek-v3.2_with_demo_shuffledhistory	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-23	9800	stop:9800
rl_v1_deepseek-v3.2_with_demo_shuffledpersona	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-23	9800	stop:9800
rl_v1_deepseek-v3.2_with_demo_split180af2c2	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	9800	stop:9800
rl_v1_deepseek-v3.2_with_demo_split272c6067 [fresh cohort]	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	9800	stop:9800
rl_v1_deepseek-v3.2_with_demo_split2c68ebb8	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	9800	stop:9800
rl_v1_deepseek-v3.2_with_demo_split73ad3ab4	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	9800	stop:9800
rl_v1_deepseek-v3.2_with_demo_split83cd1932	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	9800	stop:9800
rl_v1_deepseek-v3.2_with_demo_splita2a80ad5	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	9800	stop:9800
rl_v1_deepseek-v4-pro-nothink_with_demo	deepseek-v4-pro	DeepSeek	first-party	disabled	2026-09-23	9800	stop:9800
rl_v1_deepseek-v4-pro_with_demo	deepseek-v4-pro	DeepSeek	first-party	enabled	2026-09-23	9803	stop:9803
rl_v1_deepseek-v4-pro_with_demo_scriptedack	deepseek-v4-pro	DeepSeek	first-party	enabled	2026-09-23	8402	stop:8402
rl_v1_qwen-2.5-72b-openrouter_with_demo	qwen/qwen-2.5-72b-instruct	OpenRouter	Novita		2026-09-23	9799	length:101, stop:9698
rl_v1_qwen-2.5-72b-openrouter_with_demo_scriptedack	qwen/qwen-2.5-72b-instruct	OpenRouter	Novita		2026-09-23	8398	length:337, stop:8061
warmup_only_deepseek-v3.2_no_demo	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-23	3600	stop:3600
warmup_only_deepseek-v3.2_with_demo_feb_bullets50	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-23	3600	stop:3600
warmup_only_deepseek-v3.2_with_demo	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-23	3600	stop:3600
warmup_only_deepseek-v3.2_with_demo_shuffledpersona	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-23	3600	stop:3600
warmup_only_deepseek-v3.2_with_demo_split180af2c2	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	3600	stop:3600
warmup_only_deepseek-v3.2_with_demo_split272c6067 [fresh cohort]	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	3600	stop:3600
warmup_only_deepseek-v3.2_with_demo_split2c68ebb8	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	3600	stop:3600
warmup_only_deepseek-v3.2_with_demo_split73ad3ab4	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	3600	stop:3600
warmup_only_deepseek-v3.2_with_demo_split83cd1932	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	3600	stop:3600
warmup_only_deepseek-v3.2_with_demo_splita2a80ad5	deepseek/deepseek-v3.2	OpenRouter	Novita		2026-09-24	3600	stop:3600
slate four-arm (Amazon, H&M; ICL, protocol, direct-example, self-feedback)	deepseek-flash	DeepSeek	first-party	disabled	2026-09-23	57200	not aggregated by that runner (35 length-bounded outputs of 57200 attempts)