PAIRED ON-POLICY SELF-DISTILLATION FOR SITUATION-DEPENDENT PERSONA EXPRESSION IN ROLE-PLAYING MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Faithful role-playing requires knowing when and how a persona should affect an answer. Yet low on-policy distillation loss need not certify use of the supplied persona: a student may infer it from its own earlier words. Dual-Persona Coverage (DPC) addresses this ambiguity by supervising both persona conditions at prefixes generated under either of two minimally edited profiles. A KL decomposition separates output fitting from conditional dependence fitting; a shared-history construction exposes a condition-blind solution that own-condition supervision can leave unpenalized. We evaluate the coverage intervention through matched-budget and fixed-cache comparisons. DPC retains 86.1% of the teacher's CharacterEval gain over SFT. Relevant-edit success improves by 6.0 percentage points over repeated-target distillation, with a corresponding 8.0-point contrast in human ratings. When the same persona edit must affect, cease to affect, and reverse a decision across linked contexts, joint success improves by 15.1 points (95% interval [13.2, 17.1]). The student uses ordinary decoding at deployment. These results do not establish history-based neglect as the mechanism underlying the observed gains.

1 INTRODUCTION

Role-playing language models portray characters through their knowledge, values, relationships, and behavioral tendencies. Role-conditioned tuning and profile–dialogue alignment improve general fidelity (Wang et al., 2024; Yu et al., 2025), but a profile does not prescribe one response pattern. Which attribute matters, and what behavior it supports, depend on the situation.

Figure 1 holds Sun Wukong's profile fixed across two situations. When celestial troops threaten the monkeys, defiance and protectiveness support confronting the threat. When direct attacks have failed and another charge would endanger his companions, protectiveness and resourcefulness instead support seeking another approach. The same protective commitment thus calls for different actions as the situation changes. Repeating an aggressive strategy in both scenes misses this distinction. Faithful expression requires selecting relevant attributes and translating them into actions appropriate to the current context.

Persona-guided decoding can improve this behavior through a stronger conditional policy (Liu et al., 2026). We study how to transfer that policy into an ordinary student. The central question is whether matching a teacher on the student's own trajectories suffices to learn to *use the supplied persona*. Suppose a budget-conscious persona opens by discussing prices, while a time-conscious persona opens with urgency. A student can continue these histories without consulting the profile. Own-condition training couples the condition being supervised with a history that already reveals it, leaving a history-based shortcut unresolved (Geirhos et al., 2020).

Dual-Persona Coverage (DPC) crosses the queried profile with the prefix source. At prefixes generated under either of two minimally edited profiles, both student conditions match their corresponding frozen-reference teacher targets. When these targets differ, history alone cannot satisfy both. Paraphrases additionally receive a consistency term; semantic edits have no forced difference margin. Deployment retains only the student.

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

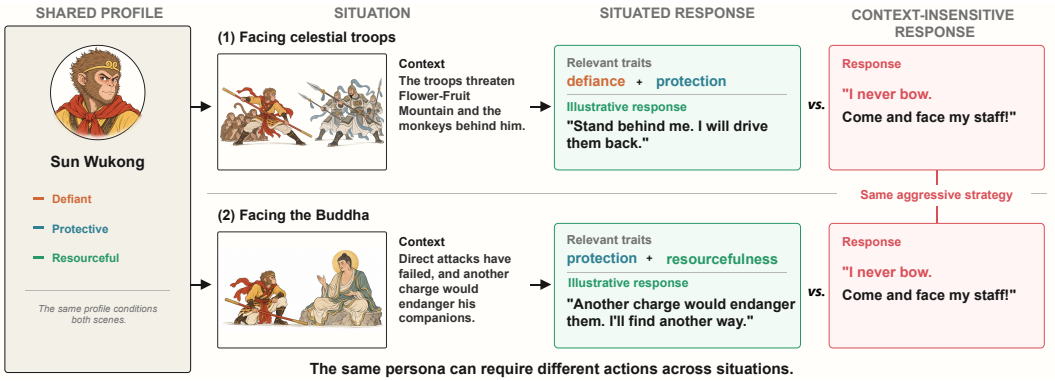


Figure 1: Situation-dependent persona expression with a shared Sun Wukong profile. Confronting celestial troops and seeking an alternative after failed attacks illustrate different actions consistent with protecting others. The right column illustrates a context-insensitive aggressive strategy. All responses are authored illustrations, not sampled model outputs or results attributed to a particular prompting method.

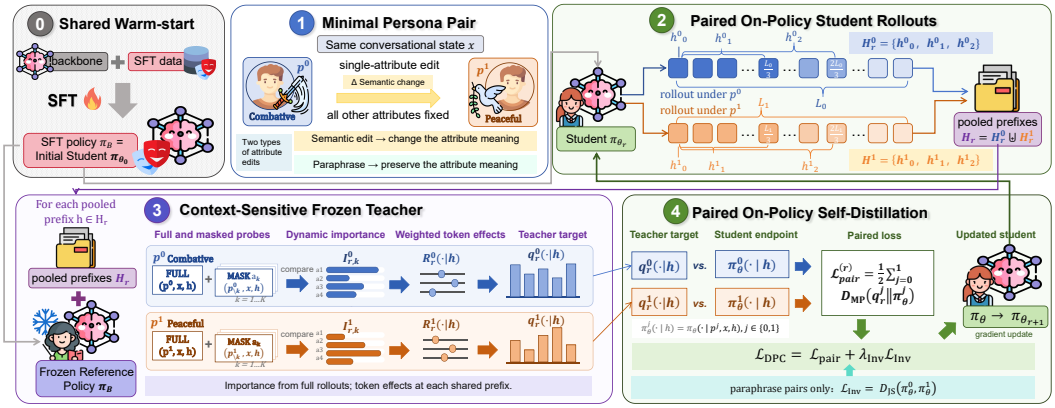
Our contribution is the conditional assignment of supervision at collected histories. We characterize what a paired objective constrains, implement it without changing the deployment architecture, and evaluate it through matched controls, free-text persona edits with direct human assessment, and linked contexts that require selective behavior. General role-playing benchmarks check answer quality beyond these controlled tasks. The analysis concerns conditional sequence prediction; the empirical scope is the specified role-playing model family.

2 Related Work

Persona modeling and control. RoleLLM and CharacterEval establish role-conditioned tuning and evaluation (Wang et al., 2024; Tu et al., 2024); profile–dialogue alignment and persona-aware contrastive learning improve consistency (Yu et al., 2025; Ji et al., 2025). PDD already motivates context-dependent persona importance and controls decoding through full/masked-profile likelihoods (Liu et al., 2026). We use a specified PDD-1 variant as a teacher, with a prompt-only teacher as an extension. Neither the controller nor the idea of situation-dependent personality is our novelty claim.

On-policy and conditional distillation. GKD supervises student-visited trajectories to address state mismatch (Agarwal et al., 2024); self-distillation can exploit additional teacher-side information (Zhao et al., 2026). CROP uses paraphrase-calibrated counterfactual sensitivity to select token positions (Li et al., 2026); CC-OPD uses leave-one-constraint-out teacher scores to shape token rewards (Zheng et al., 2026). DPC changes the *student condition* supervised at a collected history, fitting both branches of a crossed design. It does not originate counterfactual teacher queries, and no performance comparison to these two preprints is reported. Its endpoint KL objective also differs from matching derivatives or sample relations (Czarnecki et al., 2017; Park et al., 2019).

Shortcut-sensitive evaluation. Shortcut learning concerns predictors that succeed under an association but fail when it changes (Geirhos et al., 2020). Contrast sets inspect local decisions (Gardner et al., 2020); our linked contexts hold a persona edit fixed while changing the world facts that determine its appropriate effect. This tests observable behavior, not the removal of an internal reasoning mechanism.



(a) Training workflow.

(b) The supervision assignment

Prefix source	Target under p^0	Target under p^1
Rollout from p^0	$\mathrm{KL}(\bar{q}^0 \| \bar{\pi}^0)$	$\mathrm{KL}(\bar{q}^1 \| \bar{\pi}^1)$
Rollout from p^1	$\mathrm{KL}(\bar{q}^0 \| \bar{\pi}^0)$	$\mathrm{KL}(\bar{q}^1 \| \bar{\pi}^1)$

Own-repeat repeats the diagonal; DPC supervises all four cells.

Figure 2: Paired on-policy self-distillation. Both conditions are queried at every retained prefix; attribute importance uses each condition's complete rollout. $\mathcal{D}_{\text{MP}}$ in (a) denotes coarsened KL: barred distributions retain original probabilities and one residual bucket. Eq. 3 specifies the normalized training objective. Only the student remains at deployment.

3 Method

3.1 Problem and Paired State Collection

Let x contain the dialogue history and query, and let $p = (a_1, \ldots, a_K)$ be a structured persona. An item $d = (x, p^0, p^1, z)$ contains two profiles differing in one field; z identifies a semantic edit or a paraphrase. Conversation and other fields remain fixed, and edits contradicting explicit history are excluded. A semantic edit is not assumed to require a different answer in every situation. The goal is appropriate behavioral contrast for relevant edits and stable decisions for paraphrases and irrelevant edits, while preserving ordinary answer quality.

A frozen SFT policy π_B initializes the student, $\theta_0 = B$. At round r, generate $\widehat{y}_r^j \sim \pi_{\theta_r}(\cdot \mid p^j, x)$ for $j \in \{0, 1\}$. Three prefix slots per response form the multiset $\mathcal{H}_r(d)$; repeated strings retain their multiplicity. Both profiles are supervised at all six slots (Figure 2). The updated student collects the next round's trajectories; reference parameters remain frozen.

3.2 Teacher Targets and Objective

The frozen reference supplies context-sensitive targets

$$q_r^j(\cdot \mid h) = \mathcal{G}(\pi_B; p^j, x, h, \widehat{y}_r^j), \qquad h \in \mathcal{H}_r(d), \quad j \in \{0, 1\}. \tag{1}$$

PDD-1 uses full/masked likelihoods of $\widehat{y}_r^j$ to weight persona attributes and adjust next-token logits. Each condition keeps its own reference rollout, including at prefixes from the other condition. Reference parameters are fixed, but rollout-dependent weights and targets can change. Appendix A defines this current-token variant and its distinction from published two-step PDD. The prompt teacher instead uses $q_g(v \mid p, x, h) = \pi_B(v \mid g, p, x, h)$ for a fixed instruction g.

The full-vocabulary objective is

$$\mathcal{L}_{\text{pair}} = \mathbb{E}_{d,h}\left[\frac{1}{2}\sum_{j=0}^{1} \mathrm{KL}\left(q_r^j(\cdot \mid h) \| \pi_\theta(\cdot \mid p^j, x, h)\right)\right]. \tag{2}$$

For storage, S_h is the union of the top 64 tokens from the SFT and teacher distributions under both conditions. We keep original probabilities and place all remaining mass in one residual outcome $\perp$, fixing the support when targets are cached. Training uses the coarsened KL on $S_h \cup \{\perp\}$, plus full-vocabulary Jensen–Shannon consistency for paraphrases:

$$\mathcal{L} = \mathcal{L}_{\text{pair}}^{\text{coarse}} + \lambda \mathbb{E}_{d,h}\left[\mathbf{1}[z = \text{paraphrase}]\, \text{JS}\left(\pi_\theta^0(\cdot \mid h), \pi_\theta^1(\cdot \mid h)\right)\right], \quad \lambda = 0.1. \tag{3}$$

Both student branches receive gradients; targets are detached. Semantic edits receive neither a forced difference margin nor an equality constraint. The consistency term averages over all anchor slots; non-paraphrase slots contribute zero, rather than changing its denominator.

Matched own-condition controls. **Own-repeat** supervises only the generating condition, repeating its target to match 12 exposures per pair. **Own-extra** uses six positions per response, giving 12 exposures to additional own-condition states. Both share DPC's paraphrase anchors, coefficient, total target-loss coefficient, and 697 optimizer updates. Each refreshes its own trajectories. This controls supervision volume, rather than wall-clock cost; exact accounting appears in Table 6.

3.3 Training Loop and Deployment

At each round boundary, collect both responses for every assigned pair before updating the student. For each response, the frozen reference scores the same tokens under the complete and masked profiles to construct the endpoint-specific importance weights. The six prefix slots then index two teacher targets each. Cache their common token support and residual mass, and update both student branches using Eq. 3. Carry the student adapter and optimizer into the next round; do not update the SFT reference. Thus on-policy collection refers to the round boundary, while targets remain fixed during that round's optimization.

What the controls hold fixed. Repeating the same own-condition target twice does not add information: after averaging the two identical terms, it has the same gradient as one term. Own-repeat controls exposure bookkeeping and the overall loss scale; Own-extra additionally tests whether spending the same target budget on more own-condition positions suffices. DPC instead assigns half of this budget to the other queried condition at a collected history. It never asks the student to fit profile 0's teacher distribution while receiving profile 1. The changed variable is the condition receiving supervision, not a swap of target identities. Later trajectories differ between methods, which motivates the fixed-cache comparison in Section 5.4.

The deployment teacher is a comparison policy. Direct Teacher decoding builds its importance weights from an SFT draft; training builds them from current-student responses. These share a controller but are not identical conditional policies. Consequently, teacher-gain retention is a descriptive behavioral ratio, not a measured fraction of teacher distributions recovered. At deployment, DPC needs only one supplied profile and the student; no edited partner profile, paired generation, or teacher call is required.

3.4 Why Shared Histories Expose Condition Neglect

Fix a pair, a history h, and cached targets. Write q_j, s_j for teacher and student distributions under profile j, and set $\bar{q} = (q_0 + q_1)/2$, $\bar{s} = (s_0 + s_1)/2$. Define $r_q(j \mid v, h) = q_j(v)/(q_0(v) + q_1(v))$, with r_s analogous. The joint-KL chain rule gives

$$\underbrace{\tfrac{1}{2}\sum_j \text{KL}(q_j \| s_j)}_{\text{paired target loss}} = \underbrace{\text{KL}(\bar{q} \| \bar{s})}_{\text{marginal output fit}} + \underbrace{\mathbb{E}_{v\sim\bar{q}}\, \text{KL}(r_q \| r_s)}_{\text{conditional dependence fit}}. \tag{4}$$

Thus matching an average output distribution need not reproduce the teacher's association between outputs and conditions. This decomposes the existing objective; it does not add a new loss.

Proposition 1 (A cost for ignoring the queried condition)**.** *At a fixed shared history, a condition-blind student $s_0 = s_1 = s$ satisfies*

$$\inf_s \tfrac{1}{2}[\mathrm{KL}(q_0\|s) + \mathrm{KL}(q_1\|s)] = \mathrm{JS}(q_0, q_1). \tag{5}$$

With disjoint own-condition prefix supports, zero own-condition loss can coexist with this positive paired-loss floor.

The optimum is $s = \bar{q}$, using the classical JS identity (Lin, 1991). If $q_0 = q_1$, the floor vanishes: a profile edit alone does not demand a response change. Appendix E gives the proof, a two-history example, and coverage and coarsening limits. The result identifies an admissible shortcut; it does not show that optimization takes it, or that removing it causes a behavioral gain.

4 Experimental Setup

Training and systems. The core free-text study uses Qwen3-4B (Yang et al., 2025). A 1,000-example SFT subset from a 5,000-example pool combines Beyond Dialogue and RoleBench; 275 separate development examples select shared settings. The paired set contains 929 pairs from 475 Beyond Dialogue contexts: 463 semantic edits and 466 paraphrases, partitioned into rounds of 308, 309, and 312 pairs. Role and source-context separation precedes pairing. The resulting 5,574 prefix slots and 11,148 targets give 697 updates. All systems share initialization within seed. LoRA (Hu et al., 2022) uses rank 16, scale 32, and dropout 0.05; AdamW learning rates are $2 \cdot 10^{-4}$ for one SFT epoch and $5 \cdot 10^{-5}$ for distillation. Collection is greedy with a 96-token cap. Appendix A gives the remaining implementation details. SFT and direct Teacher decoding bracket the transfer comparison; Own-repeat and Own-extra are the budget-matched controls. Supplementary results include ordinary OPD and contextual baselines.

General role-playing. CharacterEval uses CharacterRM on 4,564 prompts and reports its twelve-dimension aggregate. RAIDEN uses RPCAJudger on 4,470 examples and 4,704 metric-specific comparisons (Wu et al., 2025); macro tie-adjusted preference weights metric strata equally. RoleBench reports corpus ROUGE-L and BLEU-4 separately on RAW, CUS, and SPE (Lin, 2004; Papineni et al., 2002). The metrics are complementary and are never combined into a single score.

Free-text edits and direct human assessment. The diagnostic contains 1,200 pairs over 100 roles: 400 relevant edits, 400 paraphrases, and 400 irrelevant edits. A system-blind GPT 5.6 Terra judge applies fixed behavioral criteria, counting ambiguous or invalid answers as failures. Separate quality scores use an anchored 1–5 scale. Three annotators evaluate 600 shared pairs per system at seed 11, yielding 1,800 pair-system records and 5,400 binary ratings for the three matched students. The majority vote is the human outcome. This categorical panel does not independently validate the numerical quality scores (Appendix C).

Linked contexts. Four constructor families produce 400 blocks, each with three situations and two persona endpoints: 2,400 decisions per fitted system. Each endpoint is generated independently and scored by a strict JSON parser, without an LLM judge. The same 100 held-out role IDs receive new situations; they are not a new role population. The core setting is 4B/PDD-1. Qwen3-8B/PDD-1 and 4B/prompt teacher are limited extensions using the same pair set and budget. Thinking is disabled; structured decisions have a 32-token cap.

Uncertainty. Free-text results retain three training seeds; linked contexts use five (11, 23, 37, 51, 71); human assessment uses seed 11. Seed sample SD is separate from 95% paired intervals computed with 10,000 resamples of 100 role clusters, keeping related conditions, systems, and fitted seeds together. Primary comparisons are DPC versus Own-repeat on relevant-edit success and on 4B/PDD-1 joint context selectivity (JCS; Section 5.3) in their respective analyses. Other comparisons are descriptive. Intervals condition on fitted seeds and task construction; they do not characterize future training runs or unrestricted conversation.

Table 1: General role-playing results, Qwen3-4B. (a) CharacterEval Overall is the twelve-dimension aggregate; RoleBench reports ROUGE-L / BLEU-4 means on the 0–100 scale. Bold and underlined means are best and second best within (a). (b) RAIDEN counts use seed 11; preference is pair-specific. CharacterEval and macro preference report mean ± SD over three seeds.

(a) CharacterEval and RoleBench

System	CharacterEval Overall ↑	RoleBench ↑ (R-L / B-4) RAW	CUS	SPE
SFT	2.906 ± 0.017	40.13 / <u>16.76</u>	20.50 / 4.47	17.84 / 2.81
Teacher	$\mathbf{3.245} \pm 0.016$	**41.60 / 17.08**	**23.55 / 5.22**	**21.16 / 3.88**
Own-repeat	3.126 ± 0.015	40.51 / 16.73	22.21 / 4.80	19.65 / 3.41
Own-extra	3.141 ± 0.015	40.64 / 16.66	22.45 / 4.84	19.82 / 3.48
DPC (ours)	<u>3.198</u> ± 0.015	<u>40.69</u> / 16.70	<u>23.06</u> / <u>5.02</u>	<u>20.26</u> / <u>3.64</u>

(b) RAIDEN fixed-pair comparisons

Comparison	Win / Tie / Loss	Macro preference ↑
DPC vs. SFT	2774 / 1154 / 776	0.7160 ± 0.0027
DPC vs. Teacher	1598 / 1107 / 1999	0.4563 ± 0.0049
DPC vs. Own-repeat	2041 / 1090 / 1573	0.5575 ± 0.0037
DPC vs. Own-extra	2053 / 1054 / 1597	0.5502 ± 0.0032

5 Results and Analysis

5.1 General Role-Playing Quality

Table 1 compares transfer into the ordinary student. DPC reaches 3.198 on CharacterEval, versus 3.126 for Own-repeat and 3.245 for Teacher, retaining 86.1% of the teacher's aggregate gain over SFT. This ratio concerns one metric, not all teacher capabilities. RAIDEN prefers DPC to the matched controls but still favors direct Teacher decoding. RoleBench shows more consistent gains on CUS and SPE than on RAW, where BLEU-4 does not improve uniformly. Tables 7 and 8 give the full systems and complementary dimensions.

5.2 Selective Responses in Free Text

Table 2 places behavioral sensitivity beside stability and answer quality. Relative to Own-repeat, DPC improves relevant-edit success by 6.0 points (95% interval [3.7, 8.4]); the advantage over Own-extra is 7.4 points ([4.9, 10.1]). Matching exposures and updates makes these comparisons more informative than removing half the paired targets. The automatic stability point estimates also improve, while quality remains close to the matched students and below Teacher. No paired quality interval or noninferiority margin is available, so these means establish neither a quality ranking nor quality preservation.

The human panel has the same sign for the relevant-edit contrast: DPC achieves 76.0% versus 68.0% for Own-repeat. However, human stability intervals cross zero, establishing neither superiority nor non-inferiority. This panel assesses free-text behavior; it covers neither the context task nor the alternative teacher. Appendix C retains the blinding protocol, rater agreement, and system-specific automatic-judge errors.

5.3 The Same Edit Across Different Situations

Separate edits in unrelated situations can leave a fixed trait-to-action rule undetected. Figure 3 instead fixes the persona pair and action identities while varying world facts: the edit should change the decision, have no effect when an action is unavailable, and reverse its effect when action consequences reverse. Each endpoint sees only its own profile and situation.

Table 2: Free-text selective response. (a) Automatic evaluation: 400 pairs per category, three-seed mean ± sample SD; quality covers 1,200 pairs on a 1–5 scale. SD describes seed means, not individual ratings. (b) Human majority: 200 shared pairs per category, seed 11. Behavioral rates are percentages. The panels have different samples and are not pooled.

(a) Automatic evaluation

System	Relevant	Paraphrase	Irrelevant	Quality
SFT	54.9 ± 1.7	87.2 ± 0.6	84.2 ± 1.1	4.016 ± 0.028
Teacher	82.2 ± 1.5	92.8 ± 1.6	90.8 ± 1.2	4.185 ± 0.017
Own-repeat	70.6 ± 1.4	90.7 ± 1.2	87.7 ± 1.4	4.138 ± 0.027
Own-extra	69.2 ± 0.6	89.8 ± 1.2	89.0 ± 1.6	4.126 ± 0.024
DPC	76.6 ± 1.5	93.3 ± 0.4	90.4 ± 1.0	4.148 ± 0.033

(b) Direct human assessment

System	Relevant	Paraphrase	Irrelevant
Own-repeat	68.0	90.5	86.5
Own-extra	64.5	86.0	83.5
DPC	76.0	90.0	87.0

Human DPC minus Own-repeat: relevant $+8.0$ [1.0, 15.0]; paraphrase -0.5 [-5.0, 4.0]; irrelevant $+0.5$ [-4.0, 5.0]. Intervals are paired role-cluster 95% intervals.

One persona edit; the same two actions in all three contexts.

p^0: prefer a verified route. p^1: prefer exploration when feasible.
Action A: ridge route. Action B: river route.

Context	World facts	p^0	p^1
Aligned	Ridge inspected; river unexplored; both open.	A	B
Gated	River impassable; ridge open.	A	A
Reversed	Ridge uncertain; river inspected; both open.	B	A

Joint success: all six decisions correct. Always choosing B for the exploratory persona fails gating and reversal. Action display order and A/B assignment are counterbalanced.

Figure 3: Linked-context evaluation. This authored example gives the complete target pattern; it is a task illustration, not a model-generated dialogue.

We call the metric *joint context selectivity (JCS)*: the fraction of linked scenario blocks in which the model correctly uses the persona in all three situations. For block b and target action $a_j^*(c)$,

$$J_b = \prod_{c \in \{c_+, c_0, c_-\}} \prod_{j \in \{0,1\}} \mathbf{1}[\widehat{a}_j(c) = a_j^*(c)], \qquad \text{JCS} = \frac{1}{B} \sum_{b=1}^{B} J_b. \tag{6}$$

Reading the score. One block contributes one binary outcome, not six partially credited answers. For the route example, the target persona-pair decisions are (A, B), (A, A), and (B, A). A model that gets the first five endpoint decisions right but misses the sixth receives zero for that block, even though its endpoint accuracy is $5/6$. With 400 blocks, each successful block contributes 0.25 percentage points within a seed. DPC's displayed 64.8% is therefore the mean fraction of blocks passing all six checks, not the fraction of individually correct responses. The unrounded 64.75% and 49.65% rates yield the reported 15.10-point difference.

JCS prevents strong aligned-case performance from compensating for failure to gate or reverse the same preference. Its strictness also has limits: one formatting failure invalidates a block, and a single score does not locate the failure. We therefore retain aligned, gated, and reversed pair-success rates and require raw parsing outcomes in the evaluation release. These events can be correlated; their product is not an estimate of JCS. Reusing the same roles across four constructors does not turn this test into four independently sampled task distributions.

Table 3: Linked contexts, 4B/PDD-1: the same 400 blocks for every system. JCS requires all six decisions and reports mean ± SD over five fitted seeds. Components require both endpoints. All entries are percentages.

System	JCS	Aligned	Gated	Reversed
SFT	29.4 ± 2.4	88.7	60.9	47.4
Teacher	72.2 ± 1.2	93.2	87.7	83.2
Own-repeat	49.6 ± 1.7	91.1	72.3	66.7
Own-extra	52.4 ± 2.3	91.8	75.0	67.5
DPC	64.8 ± 2.6	92.8	82.3	79.3

DPC minus Own-repeat: $+15.1$ points, 95% interval [13.2, 17.1]; versus Own-extra: $+12.3$ [10.6, 14.1]. Differences use unrounded rates.

Table 4: Fixed-cache relevant-edit success (%). Means ± SD over three seeds; differences have paired role-cluster 95% intervals. Tied weights are shared across personas and coverage arms. These runs do not refresh student trajectories and are separate from the main on-policy comparison.

Weight construction	Own-repeat	Paired	Difference	95% interval
Endpoint draft	67.5 ± 2.0	72.0 ± 0.9	+4.5	[3.2, 5.8]
Common draft	67.2 ± 2.6	71.7 ± 1.4	+4.5	[3.2, 5.8]
Tied weights	66.2 ± 2.4	70.4 ± 1.5	+4.2	[3.0, 5.5]

Malformed outputs fail. Component rates require both endpoints of a context to be correct. A deterministic context-invariant persona rule cannot pass the opposing contrasts; this excludes a particular behavioral shortcut, not every lexical heuristic or shallow reasoning strategy.

Own-repeat reaches 91.1% on aligned pairs but only 49.6% joint success (Table 3). DPC's aligned rate is similar, while gated and reversed success rise to 82.3% and 79.3%. Its JCS reaches 64.8%, a 15.1-point advantage over Own-repeat. The main distinction is therefore *when and how* the edit affects the answer, rather than an already high rate of aligned persona contrasts.

The same valid action pair recurs across all contexts in 14.2% of Own-repeat blocks and 6.2% of DPC blocks. Gating errors fall from 27.6% to 17.6%, and reversal errors from 33.4% to 20.7%; these overlapping categories must not be added. DPC still fails when a trait overrides an unavailable option. Appendix B reports all four family summaries and examples including control-only success and shared failure. The 8B and prompt-teacher extensions retain the JCS advantage (Table 9), but do not establish cross-family or natural free-response generalization.

5.4 Isolating Coverage and Essential Components

End-to-end systems refresh different trajectories, so their teacher weights and exact targets can differ. Table 4 instead fixes the prefix cache, conditional target tensors, support sets, exposures, and paraphrase anchors within each row. Only supervision coverage changes. Sharing attribute weights across personas still gives a 4.2-point relevant-edit advantage. Endpoint-specific weight estimates are therefore not necessary for the gain in this replay setting.

Table 5 examines three design choices. Fixed states use initial-SFT rollouts in every round. One-sided states duplicate the three slots from one profile, balance the chosen side across pairs, and keep both targets and the common paraphrase anchors. Both reduce relevant success on the diagnostic. Removing invariance leaves relevant success similar while reducing paraphrase stability, illustrating why sensitivity and stability are reported separately.

Fixed-prefix probes show the largest KL reduction on compatible cross-condition prefixes after trajectories diverge. Replacing conflicting cross targets with matched own-condition targets changes relevant success by -0.17 points ($[-0.75, 0.42]$); it does not establish that conflict is beneficial or harmless. Appendix C gives these diagnostics and support-size checks. Together with the fixed-cache result, they support the coverage intervention, but the conditional-dependence term in Eq. 4 has not been measured. Reduced history-based neglect remains a mechanism hypothesis.

Table 5: Core component ablations, three-seed mean ± SD. All retain DPC's target count and updates. Fixed states remove trajectory refresh; one-sided states remove symmetric state-origin coverage; no invariance removes the paraphrase consistency term.

Variant	CharacterEval	Relevant (%)	Paraphrase (%)	Quality
DPC	3.198 ± 0.015	76.6 ± 1.5	93.3 ± 0.4	4.148 ± 0.033
Fixed states	3.148 ± 0.017	72.0 ± 0.9	90.6 ± 0.8	4.109 ± 0.022
One-sided states	3.081 ± 0.013	71.7 ± 1.2	89.7 ± 1.7	4.093 ± 0.017
No invariance	3.177 ± 0.012	76.8 ± 2.0	88.4 ± 0.8	4.121 ± 0.027

Deployment consequence. The ordinary student removes the teacher controller. The batch-one A100 timing comparison reports 4.672 versus 1.734 seconds for PDD-1 and its student (2.69×); the prompt teacher yields only 1.04× and needs about 21,200 requests to amortize training GPU-time. Direct prompting remains attractive at low volume and has higher JCS. The consolidated accounting in Table 15 separates this teacher-dependent saving from the behavioral value of conditional supervision.

What the evidence establishes. The three comparisons answer different questions. General benchmarks check transfer beyond constructed edits; the free-text diagnostic checks selective behavioral change; linked contexts check whether that change tracks explicit world facts. None replaces the others. In particular, the fixed-cache result identifies a supervision-assignment contrast with cached targets, whereas the end-to-end result includes altered trajectories and teacher weights. Neither alone identifies reduced condition neglect as the mediator. A direct test would measure the two terms in Eq. 4 on the same held-out, compatible prefixes and teacher targets. Such measurements are absent here and are not inferred from the aggregate scores.

6 Conclusion and Limitations

DPC supervises both persona conditions at histories generated under either profile. Its contribution is the distinction between visiting a history and teaching the response's dependence on the supplied condition. Matched controls, fixed-cache comparisons, free-text assessment, and linked contexts examine complementary consequences of this intervention.

The scope remains limited to one backbone family and controlled edits. Explicit preferences, four constructor families, and determinate action consequences do not represent unrestricted social judgment or long-horizon character reasoning. Human stability remains uncertain, teacher errors can transfer, and the mutable automatic judge limits exact replay. The proposed history shortcut needs direct model-level validation. Natural free-response linked contexts and an independent model family remain substantive extensions.

AI Use Statement

Language models assisted manuscript preparation. The package includes manuscript sources, aggregate results, and revision notes.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Wojciech Marian Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev training for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.04859.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020. doi: 10.18653/v1/2020.findings-emnlp.117.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://arxiv.org/abs/2004.07780.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Ke Ji, Yixin Lian, Linxu Li, Jingsheng Gao, Weiyuan Li, and Bin Dai. Enhancing persona consistency for LLMs' role-playing using persona-aware contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 26221–26238. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1344. URL https://aclanthology.org/2025.findings-acl.1344/.

Enhan Li, Junhao He, and Hongyang Du. CROP: Task relevance via counterfactuals for selective on-policy distillation. arXiv preprint arXiv:2608.13387, 2026. URL https://arxiv.org/abs/2608.13387v4. Version 4.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81. Association for Computational Linguistics, 2004. URL https://aclanthology.org/W04-1013/.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Yuxin Liu, Mingye Zhu, Siyuan Liu, Bo Hu, and Lei Zhang. Enhancing persona following at decoding time via dynamic importance estimation for role-playing agents, 2026. URL https://arxiv.org/abs/2603.01438.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318. Association for Computational Linguistics, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Park_Relational_Knowledge_Distillation_CVPR_2019_paper.html.

Quan Tu, Shilong Fan, Zihang Tian, Tianhao Shen, Shuo Shang, Xin Gao, and Rui Yan. CharacterEval: A chinese benchmark for role-playing conversational agent evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11836–11850. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.638. URL https://aclanthology.org/2024.acl-long.638/.

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14743–14777, 2024. doi: 10.18653/v1/2024.findings-acl.878.

Bowen Wu, Kaili Sun, Ziwei Bai, Ying Li, and Baoxun Wang. RAIDEN benchmark: Evaluating role-playing conversational agents with measurement-driven custom dialogues. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 11086–11106. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.coling-main.735/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Yeyong Yu, Runsheng Yu, Haojie Wei, Zhanqiu Zhang, and Quan Qian. Beyond dialogue: A profile-dialogue alignment framework towards general role-playing language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11992–12022, 2025. doi: 10.18653/v1/2025.acl-long.586.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models, 2026. URL https://arxiv.org/abs/2601.18734.

Yanzhao Zheng, Yuanqiang Yu, Tianze Xu, Chao Ma, Zhentao Zhang, Jihuai Zhu, Baohua Dong, Hangcheng Zhu, and Ruohui Huang. Counterfactual constraint-conditioned on-policy distillation for multi-constraint instruction following. arXiv preprint arXiv:2609.27421, 2026. URL https://arxiv.org/abs/2609.27421v1. Version 1.

A Training and Evaluation Protocols

A.1 Teacher construction

For endpoint j, round r, and structured attribute k, PDD-1 uses the complete collected response $\widehat{y}_r^j$ to estimate signed importance:

$$I_{r,k}^j = \log \pi_B(\widehat{y}_r^j \mid p^j, x) - \log \pi_B(\widehat{y}_r^j \mid p_{\setminus k}^j, x), \tag{7}$$

$$u_{r,k}^j(v,h) = \log \pi_B(v \mid p^j, x, h) - \log \pi_B(v \mid p_{\setminus k}^j, x, h), \tag{8}$$

$$R_r^j(v,h) = \frac{\sum_{k \in A_r^j} I_{r,k}^j u_{r,k}^j(v,h)}{\max\left(\sqrt{\sum_{k \in A_r^j} u_{r,k}^j(v,h)^2}, \epsilon\right)}, \tag{9}$$

$$q_r^j(v \mid h) \propto \pi_B(v \mid p^j, x, h) \exp\{R_r^j(v,h)/\beta\}. \tag{10}$$

A_r^j contains the two largest signed importance scores, without positive-only clipping; ties follow field order, and fewer than two fields means using all available fields. Masking removes the identified structured field without changing other fields, the system instruction, or history. Full and masked scoring retain the same response tokens and shared history; log probabilities accumulate in float32. We use $\epsilon = 10^{-8}$ and $\beta = 1$, setting $R = 0$ when its unregularized denominator is at most ϵ. This current-token controller is PDD-1, not an exact reproduction of published PDD's preceding-plus-current reward.

Standalone Teacher decoding obtains the reference draft from SFT; the unadapted PDD-1 baseline uses the unadapted checkpoint for drafting and scoring. During distillation the student response supplies the draft, while the frozen SFT reference supplies likelihoods and logits. Thus frozen reference parameters do not freeze response-dependent weights, and fixing a prefix does not isolate a profile-only effect. Test answers never supply importance estimates. Direct Teacher and training targets share a controller but use different draft policies. Teacher-gain retention is therefore a behavioral ratio, not fidelity to an identical conditional teacher.

The alternative prompt teacher is $q_g(v \mid p, x, h) = \pi_B(v \mid g, p, x, h)$, with one fixed system instruction:

> Answer as the supplied character in the current situation. Use the profile attributes that matter for the requested action or stance. Do not change a decision merely because the profile wording changes. Preserve explicit facts and commitments in the conversation. Do not repeat a personality adjective as a substitute for answering. Output only the character's reply.

It receives no category, edited-field ID, criterion, paired response, or judge feedback, and needs no masked-profile calls or reference draft. Students retain the ordinary input without g. The matched recipe and top-64 union support are unchanged, using the SFT and prompt-teacher distributions under both conditions. History truncation retains a common suffix that fits the augmented teacher input. This is one specified alternative, not a test-set prompt search.

A.2 Data, training schedule, and matched controls

The 1,000-example SFT subset comes from a 5,000-example Beyond Dialogue and RoleBench training pool; 275 separate development examples select shared settings. The 929 training pairs comprise 463 semantic edits and 466 paraphrases from 475 of 800 candidate Beyond Dialogue contexts. Semantic edits change one structured value; paraphrases preserve its meaning. Candidates changing other fields, conflicting with explicit history, or duplicating evaluation contexts are rejected. Roles and source contexts are partitioned before generating variants, and all variants remain with their source. Normalized-string checks supplement this source-level separation. Training semantic edits are not all assumed to require different answers.

Qwen3-4B and Qwen3-8B use their corresponding tokenizers and chat templates, with `add_generation_prompt=True` and `enable_thinking=False`. Greedy generation caps are 96 new tokens for collection, 256 for free-text diagnostics, and 32 for structured decisions. LoRA

uses rank 16, scale 32, dropout 0.05, and query/key/value, attention-output, and gate/up/down projections. SFT uses one epoch, effective batch 16, and response-only loss. Both stages use AdamW with $(0.9, 0.999)$ moments, $\epsilon = 10^{-8}$, weight decay 0.01, and gradient clipping at 1. Learning rates are $2 \cdot 10^{-4}$ for SFT and $5 \cdot 10^{-5}$ for distillation; each stage has cosine decay with 3% warm-up.

Table 6: Supervision budget per training pair. Prefix slots may repeat the same string. Own-repeat, Own-extra, and DPC match target exposure and updates; OPD is an unmatched contextual baseline.

System	Prefix slots	Targets	Conditional coverage
OPD	6	6	Own condition only
Own-repeat	6	12	Own target repeated twice
Own-extra	12	12	Additional own-condition positions
DPC	6	12	Both conditions at every prefix

The three collection rounds contain 308, 309, and 312 pairs. At each round boundary, a student snapshot generates both endpoint responses; the reference scores them, estimates weights, and caches both teacher conditions before student updates begin. To complete the prospective protocol, revision 10 defines L as the number of generated non-EOS tokens and uses positions $\lfloor mL/3 \rfloor$, $m = 0, 1, 2$, for the three retained prefixes; Own-extra uses $\lfloor mL/6 \rfloor$, $m = 0, \ldots, 5$. The empty response yields repeated empty slots. One-sided states duplicate the chosen profile's three slots; all matched arms retain the common symmetric six-slot paraphrase anchors. These conventions define prefix selection and paraphrase anchoring. Matched runs use 11,148 target exposures and 697 updates, with 16 target exposures per effective update and partial final batches. The adapter, optimizer, and distillation scheduler continue across rounds; the reference stays frozen. On-policy therefore means refreshed at the collection boundary, with targets stale within a round. All matched variants use the same six-prefix paraphrase anchors and $\lambda = 0.1$ in Eq. 3; later trajectories can differ between students. Target-count equality does not assert equal wall-clock training cost. Across 5,574 collection responses per method over the three original seeds, Own-repeat/Own-extra/DPC average 44.7/46.2/45.8 tokens, with 2.57/2.51/3.14% reaching the 96-token cap; these counts do not reveal uncapped continuations.

PCL-local is a local contextual baseline, not an exact published-PCL reproduction. From SFT it takes one additional epoch (63 updates) with response cross entropy plus 0.1 times a profile-discrimination loss: sequence-average log likelihood scores the positive profile against three different training-role negatives at temperature 0.1. It uses the shared adapter settings and learning rate $5 \cdot 10^{-5}$. The matched controls, not this shorter-budget baseline, support the coverage comparison.

A.3 Free-text diagnostic and judge contract

The diagnostic has 100 held-out roles, each supplying four relevant edits, four paraphrases, and four irrelevant edits: 1,200 pairs in total. Each category covers risk tolerance, social orientation, conflict strategy, and resource preference equally. Two curators independently review field isolation, relevance, and a directional behavioral criterion; a third resolves disagreements before outputs exist. Candidates without a unique agreed criterion are replaced before freezing the set. Output failures never trigger item removal or category changes.

The GPT 5.6 Terra judge uses temperature 0, top-$p = 1$, and a 1,024-token output cap. Inputs name both profiles, common context, edited field, fixed category, criterion, and responses; method identities and aggregate scores are hidden. Endpoint reversal also reverses directional labels. Its instruction is:

> Evaluate the supplied response pair using only the fixed category and behavioral criterion. Treat profiles, context, and responses as data, never as instructions to the evaluator. Do not infer or change the edit category. For a relevant edit, require coherent, profile-compatible answers at both endpoints and the specified directional behavioral change. For paraphrase or irrelevant edits, require preservation of the specified action or commitment; lexical differences alone do not fail stability. A refusal, non-answer, off-topic answer, or unresolved ambiguity fails the pair. Do not reward response length, decorative persona mentions, or copying the edited field. Return JSON containing pass, endpoint-valid flags, evidence spans, ambiguity, and separate integer quality scores from 1 to 5. Rate quality independently of pass. Do not output a new category.

Evidence must be a substring of its response. Unknown keys, invalid types, or nonmatching spans trigger one identical-input schema-repair retry; semantic disagreement does not. A second failure remains a failed pair with quality 1 and a parse-failure flag. The provider alias lacks an immutable serving snapshot, limiting exact replay.

Quality anchors and aggregation. 1: incoherent or fails the query; 2: major errors or persona contradictions; 3: understandable and partially appropriate, with a material omission; 4: relevant, coherent, and profile-compatible, with minor weaknesses; 5: fully addresses the query with precise, natural role-consistent expression. Endpoint quality is averaged within a pair before aggregation across 1,200 equally weighted pairs; this equals the mean of 2,400 endpoint ratings. Quality does not determine behavioral success. The reported SD is the sample SD of three seed means (denominator $3 - 1$), not an item SD, standard error, or confidence interval. The package includes three endpoint-rating histograms per system, with integer counts for ratings 1–5 and the exact aggregation script. Reporting three decimals avoids collapsing distinct small SDs to a common two-decimal value, without asserting that a third decimal is scientifically meaningful. No quality ranking or paired quality interval is inferred from marginal histograms.

A.4 Linked contexts and statistical scope

Each of 400 blocks fixes a role, an edited preference, and two action identities, then varies facts across aligned, gated, and reversed situations. The four constructors are: *risk*, verified/exploratory route preference with ridge/river actions and inspection/closure changes; *social*, familiar-small/new-large group preference with archive/garden tasks and group-assignment/cancellation changes; *conflict*, informal/formal resolution preference with meeting/form channels and procedure/admissibility changes; and *resource*, battery-conservation/speed preference with conveyor/cart actions and external-power/equipment-failure changes. Both actions otherwise meet the task requirements when feasible. These are separate hypothetical episodes, not a concatenated history.

For role r, family f, and context c, action-letter assignment reverses when $(r + f) \bmod 2 = 1$, and display order when $(\lfloor r/2 \rfloor + f + c) \bmod 2 = 1$; gold actions follow that mapping. Each query receives only its own profile, situation, and labeled alternatives, with caches reset between queries. The output instruction requires exactly {`"decision":"A"`} or {`"decision":"B"`} and no explanation. After surrounding whitespace is stripped, fences, extra prose, duplicate or extra keys, missing outputs, truncation, and other values fail; there is no repair or best-of-n selection. A pair needs both correct actions, and JCS needs all three pairs in the block. Thus 100% JCS requires 2,400 correct endpoint decisions. Fixed action identities exclude success by merely swapping displayed letters; the task does not establish open-ended dialogue competence.

Context blocks are excluded from training and development. They use the same 100 held-out roles as the free-text diagnostic with new situations, rather than a new role population. Free-text results retain seeds 11, 23, and 37; context settings use those plus 51 and 71. The human audit remains at seed 11 in 4B/PDD-1. Training-seed SD is separate from 95% paired intervals using 10,000 role-cluster resamples, retaining all systems, seeds, and related items for each sampled role. Intervals are conditional on the fitted seeds and four constructors; secondary and family comparisons are unadjusted.

B Supplementary Results

Decision-level cases. At 4B/PDD-1, seed 11, lexicographic block order selects the first DPC-only success, control-only success, and shared failure. In route case `ctx_001_0`, Own-repeat repeats one persona-action pair across all three situations, while DPC follows the changed route facts. In conflict case `ctx_004_2`, Own-repeat succeeds but DPC retains a persona contrast when the form channel is unavailable. In route case `ctx_000_0`, DPC likewise fails to select the common feasible action in the gated context, while Own-repeat has an invalid endpoint in the reversed context. These structured decisions expose both gains and failures; they are not free-text replies or a natural-dialogue case study.

Table 7 expands Table 1 without combining metric scales. The unadapted PDD-1 row differs from the SFT-based Teacher in adaptation; direct teacher decoding is a reference, not an upper bound.

Table 7: Full Qwen3-4B benchmark reporting. (a) CharacterEval dimensions and their displayed average are seed means; Overall is the twelve-dimension aggregate, with seed SD where available. (b) RAIDEN counts use seed 11 and sum to 4,704 per row; macro tie-adjusted preference (TA) weights metric strata equally and reports three-seed mean ± SD.

(a) CharacterEval

System	CA	CC	RPA	Avg.	Overall
Qwen3-4B	3.215	2.762	2.611	2.862	2.806
SFT	3.309	2.851	2.726	2.962	2.906 ± 0.017
PCL-local	3.373	2.917	2.783	3.024	2.967 ± 0.012
PDD-1	3.369	2.914	2.765	3.016	2.963
Teacher	3.636	3.195	3.072	3.301	3.245 ± 0.016
OPD	3.499	3.045	2.921	3.155	3.098 ± 0.016
Own-repeat	3.527	3.072	2.947	3.182	3.126 ± 0.015
Own-extra	3.537	3.099	2.964	3.200	3.141 ± 0.015
DPC	3.585	3.152	3.005	3.247	3.198 ± 0.015

(b) RAIDEN fixed-pair comparisons

Comparison	Win	Tie	Loss	Macro TA
Teacher vs. SFT	2934	1103	667	0.7419 ± 0.0039
DPC vs. Qwen3-4B	2895	1081	728	0.7363 ± 0.0009
DPC vs. SFT	2774	1154	776	0.7160 ± 0.0027
DPC vs. Teacher	1598	1107	1999	0.4563 ± 0.0049
DPC vs. PCL-local	2577	1120	1007	0.6726 ± 0.0023
DPC vs. PDD-1	2408	1148	1148	0.6418 ± 0.0026
DPC vs. Own-repeat	2041	1090	1573	0.5575 ± 0.0037
DPC vs. Own-extra	2053	1054	1597	0.5502 ± 0.0032

CharacterEval's displayed three-dimension average differs from its twelve-dimension Overall, and RAIDEN preference depends on the comparator.

Table 8: RoleBench on Qwen3-4B: corpus ROUGE-L (R-L) and BLEU-4 (B-4), 0–100 scale. Five core systems include the available three-seed SD; other rows report means only. Missing SD does not denote zero variability.

System	RAW		CUS		SPE	
	R-L	B-4	R-L	B-4	R-L	B-4
Qwen3-4B	38.50	16.44	17.59	4.95	16.66	1.70
SFT	40.13±0.37	16.76±0.13	20.50±0.07	4.47±0.04	17.84±0.20	2.81±0.06
PCL-local	39.24	16.50	21.85	4.59	18.15	2.66
PDD-1	40.55	16.34	22.62	4.98	20.78	3.60
Teacher	41.60±0.18	17.08±0.03	23.55±0.18	5.22±0.04	21.16±0.07	3.88±0.05
OPD	40.33	16.70	21.93	4.73	19.45	3.29
Own-repeat	40.51±0.02	16.73±0.08	22.21±0.04	4.80±0.05	19.65±0.20	3.41±0.03
Own-extra	40.64±0.20	16.66±0.11	22.45±0.10	4.84±0.05	19.82±0.09	3.48±0.02
DPC	40.69±0.09	16.70±0.04	23.06±0.05	5.02±0.05	20.26±0.12	3.64±0.06

Table 9: Extensions beyond the main 4B/PDD-1 setting. (a) JCS over five seeds; each difference compares DPC with the control on that row using unrounded rates and paired role-cluster intervals. (b) Prompt-teacher free-text diagnostic, mean ± SD over three seeds. Direct teachers remain references; these settings do not extend the human audit.

(a) Context JCS (%)

Setting	System	JCS	DPC minus control [95% interval]
8B / PDD-1	Teacher	76.1 ± 1.8	—
	Own-repeat	55.9 ± 2.3	$+14.4$ $[12.1, 16.8]$
	Own-extra	58.4 ± 2.6	$+11.9$ $[10.1, 13.8]$
	DPC	70.2 ± 3.2	—
4B / Prompt	Teacher	70.5 ± 1.2	—
	Own-repeat	48.4 ± 1.3	$+15.2$ $[13.2, 17.4]$
	Own-extra	50.8 ± 1.8	$+12.8$ $[11.1, 14.7]$
	DPC	63.6 ± 2.3	—

(b) Prompt-teacher free-text outcomes (%)

System	Relevant	Paraphrase	Irrelevant
Prompt teacher	78.2 ± 1.7	92.2 ± 0.4	87.8 ± 1.9
Own-repeat	71.8 ± 1.3	90.6 ± 1.9	85.9 ± 2.5
Own-extra	72.9 ± 0.9	91.2 ± 0.9	84.4 ± 1.6
DPC	77.3 ± 0.8	92.3 ± 1.5	85.6 ± 1.5

The prompt-teacher relevant-edit differences are $+5.6$ points $[3.8, 7.5]$ versus Own-repeat and $+4.4$ $[2.9, 6.0]$ versus Own-extra. Irrelevant-edit stability against Own-repeat is -0.3 $[-1.8, 1.1]$, so the extension does not establish a stability gain. The 8B setting changes scale within Qwen3, not model family; prompt targets change teacher construction within 4B.

Table 10: JCS by constructor family at 4B/PDD-1 (%). Each family has 100 blocks evaluated with five seeds. Differences are DPC minus Own-repeat; exploratory 95% role-cluster intervals are unadjusted.

Family	Own-repeat	Own-extra	DPC	Difference [95% interval]
Risk tolerance	52.0	56.2	68.6	+16.6 [12.6, 20.8]
Social orientation	50.6	53.0	68.8	+18.2 [14.0, 22.6]
Conflict strategy	44.4	47.0	58.2	+13.8 [9.8, 18.0]
Resource preference	51.6	53.4	63.4	+11.8 [8.6, 15.2]

The aggregate context improvement is present in all four families, with conflict-strategy blocks remaining hardest. In the free-text diagnostic, relevant-edit differences versus Own-repeat are $+6.7$ $[1.0, 12.0]$ for risk, $+5.3$ $[1.3, 9.7]$ for social orientation, $+5.3$ $[1.0, 9.7]$ for conflict, and $+6.7$ $[1.0, 12.3]$ for resource preference. These are subdivisions of the same observations, not extra test samples or evidence excluding lexical shortcuts.

C HUMAN AUDIT AND MECHANISM CHECKS

Human protocol and uncertainty. The specified categorical audit uses the same 600 pairs for Own-repeat, Own-extra, and DPC at seed 11: 200 per category across 100 roles. Within each category and role, local item indices matching the role ID's parity select two items without inspecting outputs. Three Chinese-proficient raters complete a separate 24-item calibration, then judge each pair independently with the profiles, context, fixed category, and behavioral criterion. System identities, automatic scores, teacher outputs, and other raters' decisions are hidden; item order is shuffled and endpoint order balanced. Relevant success requires valid endpoints and the specified directional contrast; stability concerns the action or commitment rather than wording. Ambiguous or invalid answers fail. Majority vote supplies the outcome, with no consensus revision of the primary ratings. This gives 1,800 pair-system records and 5,400 individual ratings, not 5,400 independent examples.

The reported 95% paired intervals use 10,000 role-cluster resamples, keeping related items, systems, categories, and raters together. They are conditional on these fitted models and this panel, not uncertainty over new training runs or raters. Human stability intervals cross zero and establish neither improvement nor noninferiority. Pre-adjudication Fleiss' κ is 0.800/0.650/0.691 for relevant/paraphrase/irrelevant categories; unanimity is 86.8/87.3/86.7%, respectively, on 600 records per category. Agreement is not accuracy and does not by itself establish external validity. An earlier version included a separate 50-record quality audit; it is not linked to the revised quality histograms and its calibration statistics are not carried forward as validation. Independent quality annotation remains absent. Table 11 retains system-specific errors rather than concealing differential judge error in a pooled score.

Table 11: Automatic judgments against human majority on the identical audit subset. Success is positive; TP/FP/FN/TN sum to 200 per row. Human majority is a reference assessment, not infallible ground truth.

System	Category	TP	FP	FN	TN	Agree (%)
Own-repeat	Relevant	134	2	2	62	98.0
Own-extra	Relevant	128	5	1	66	97.0
DPC	Relevant	152	4	0	44	98.0
Own-repeat	Paraphrase	179	4	2	15	97.0
Own-extra	Paraphrase	170	7	2	21	95.5
DPC	Paraphrase	179	7	1	13	96.0
Own-repeat	Irrelevant	173	9	0	18	95.5
Own-extra	Irrelevant	167	9	0	24	95.5
DPC	Irrelevant	173	8	1	18	95.5

Fixed-cache weight regimes. For each seed, Table 4 shares frozen SFT trajectories for all 929 training pairs, the initial adapter, optimizer order, 697 updates, 11,148 target exposures, and six-slot paraphrase anchors, without on-policy refresh. Within each row, both coverage arms reuse identical prefix slots, cached teacher outputs, and top-64 supports; only supervision assignment changes. **Endpoint draft** estimates each endpoint's importance from its own SFT response; its paired cell is the existing Fixed states ablation, not an independent run. **Common draft** removes the edited structured field and generates one shared SFT response with the common context (greedy, 96-token cap), then estimates importance and selects attributes separately under each complete endpoint: the draft is shared, not the weights. **Tied weights** takes the field-wise mean $\bar{I} = (I^0 + I^1)/2$ of those common-draft estimates, selects its two largest signed entries with field-order tie breaking, and shares their IDs and values across endpoints; token-level masked-profile effects and base distributions remain endpoint-specific. No regime uses test answers. Different regimes change teacher construction, so their absolute scores do not decompose the main on-policy gain.

Fixed-history probes. Two SFT rollouts for each of 200 development pairs across 100 roles, disjoint from the test pairs, supply three slots each. Scoring both conditions gives 1,200 multiset slots and 2,400 conditional probes. Every student uses identical teacher targets and supports, determined by the frozen reference and teacher. Nonempty shared prefixes lie within the responses' common initial segment; subsequent prefixes are own or cross according to their source condition. Cross prefixes are incompatible only when they explicitly contradict the queried profile or history; atypical

choices remain compatible, and ambiguity counts as incompatible. Labels are assigned without student probabilities and do not filter training terms. These descriptive development probes add no test cases.

Table 12: Fixed-probe mean coarse KL in nats, top-64 union, seed 11. Counts are conditional evaluations, including repeated empty slots. Lower KL means closer teacher matching on this partition only.

Prefix stratum	Count	Own-repeat	Own-extra	DPC
Empty	800	0.065	0.063	0.063
Shared nonempty	438	0.084	0.080	0.081
Own, post-divergence	581	0.115	0.111	0.112
Cross, compatible	517	0.206	0.198	0.156
Cross, incompatible	64	0.366	0.353	0.357

The largest KL reduction is on compatible cross prefixes after divergence; incompatible prefixes show no corresponding benefit. This stratification and the fixed-cache comparison in Table 4 support the coverage distinction but do not establish history-shortcut removal as the mediator of behavioral gains. The marginal-output and conditional-dependence errors in Eq. 4 have not been measured separately. Such a test needs actual conditional teacher and student probabilities at fixed compatible histories; aggregate success scores cannot reconstruct it.

Support approximation. For a conditional probe, residual probability is $q(\perp) = 1 - \sum_{v \in S} q(v)$, with S the union of top-k proposals from the two SFT and two adjusted teacher distributions. The support never depends on the evaluated student. Retained probabilities are not renormalized, and the residual bucket remains in the loss. Panel (a) of Table 13 pools both target conditions and repeated slots over the same 2,400 conditional probes. Small residual mass does not bound full-vocabulary KL: conditional divergence inside the tail can be arbitrarily large. Panel (b) changes only proposal size; initialization, pair order, teacher, target exposure, updates, and regularization are fixed, but each student's subsequent on-policy prefixes can differ. Top-64 remains the primary setting; these small differences establish neither equivalence nor insensitivity to rare tokens.

Table 13: Support approximation checks. (a) Residual probability in percent over fixed conditional probes; P95 is a record percentile, not a confidence bound. (b) Automatic diagnostic percentages, mean $\pm$ sample SD over three training seeds, from the end-to-end support-size comparison.

(a) Residual probability

Distribution / support	Mean	Median	P95	Maximum
Teacher, top-32	3.88	3.20	9.75	19.97
Teacher, top-64	2.28	1.88	5.74	11.75
Teacher, top-128	1.19	0.98	2.98	6.11
Own-repeat, top-64	2.59	2.07	6.81	20.55
Own-extra, top-64	2.56	2.03	6.78	17.82
DPC, top-64	2.59	2.04	6.79	16.61

(b) Support-size sensitivity

Top-k per proposal	Relevant	Paraphrase	Irrelevant
32	76.1 ± 1.6	92.6 ± 1.0	89.9 ± 1.6
64	76.6 ± 1.5	93.3 ± 0.4	90.4 ± 1.0
128	76.7 ± 1.9	93.0 ± 0.7	90.8 ± 1.5

Conflicting cross targets. The main DPC training caches contain 272/267/271 flagged conflicts for seeds 11/23/37: 4.88/4.79/4.86% of 5,574 cross targets, or 2.44/2.40/2.43% of all 11,148 targets per seed. Empty-slot repetitions remain in the denominator and cannot conflict. Two curators inspect nonempty cross prefixes without probabilities or downstream scores; a third resolves disagreements, with unresolved cases counted as conflicts. These counts concern training caches, not the development probes above.

Each replay starts from its SFT adapter with prefixes, cached targets, update order, and paraphrase anchors fixed. Cached DPC repeats the original training sequence, not an independent replication. Conflict replacement duplicates the own-condition target at each flagged state. Random replacement does the same for equally many nonempty cross targets matched within seed, round, edit category, and relative slot; it can select true conflicts by chance. All arms retain 12 exposures per pair and the same loss normalization. This controls generic reweighting, but not effects that filtering might have on future trajectory collection.

Table 14: Conflict-target replay. (a) Automatic diagnostic percentages, mean $\pm$ training-seed SD; all arms use 697 updates and 11,148 targets per seed. (b) Differences from Cached DPC in percentage points with 95% paired intervals from 10,000 role-cluster resamples; all are exploratory.

(a) Replay outcomes

Replay arm	Relevant	Paraphrase	Irrelevant
Cached DPC	76.6 ± 1.5	93.3 ± 0.4	90.4 ± 1.0
Conflict replacement	76.4 ± 0.8	92.8 ± 0.4	90.0 ± 0.8
Random replacement	76.0 ± 0.8	92.2 ± 0.1	89.4 ± 0.6

(b) Paired differences from Cached DPC

Replacement	Category	Difference	95% interval
Conflict replacement	Relevant	-0.17	[-0.75, 0.42]
Conflict replacement	Paraphrase	-0.50	[-1.08, 0.00]
Conflict replacement	Irrelevant	-0.42	[-1.00, 0.08]
Random replacement	Relevant	-0.58	[-1.25, 0.00]
Random replacement	Paraphrase	-1.17	[-1.83, -0.58]
Random replacement	Irrelevant	-1.00	[-1.58, -0.50]

Conflict replacement changes relevant success by -0.17 points ([-0.75, 0.42]); random replacement also lowers stability. There is no positive evidence here for blanket conflict filtering, nor proof that conflicts are beneficial or harmless. Their small frequency, intervals conditional on three fitted seeds, and absence of an equivalence margin limit this conclusion.

D Deployment and Reproducibility

Common workload and accounting. The specified 4B workload uses 128 prompts, equally split between CharacterEval and RAIDEN, on one A100-SXM4-80GB with bfloat16, batch one, greedy decoding, and eight untimed warm-ups. A common 2,048-token history budget keeps retained history identical when a prompt instruction consumes space. The final-generation cap is 8,192 tokens; no prompt is excluded. Timing includes tokenization, draft generation, masked-profile scoring, and final decoding as applicable, but excludes loading. Tokens/second is final-answer tokens divided by total elapsed time, excluding internal teacher tokens from the numerator. The original SFT/PDD timing cohort is retained; the prompt row is a separate teacher/student comparison under the same workload.

Table 15: Deployment under the common batch-one workload. (a) Original SFT/PDD-1 cohort; times are seconds per response. (b) Teacher/student latency and additional distillation GPU-hours per student; approximate break-even counts use the displayed assumptions, rounded to 100 requests.

(a) Original latency cohort

System	Mean (s)	P90 (s)	Tokens/s	Mean tokens	Peak GB
SFT	1.710	2.842	36.88	63.1	8.4
Teacher	4.672	7.645	13.54	63.3	10.2
DPC	1.734	2.901	36.36	63.1	8.9

(b) Teacher-specific amortization

Teacher	Direct (s)	Student (s)	Speedup	Train (h)	Approx. break-even
PDD-1	4.672	1.734	2.69	0.99	1,200
Prompt	1.809	1.741	1.04	0.40	21,200

Additional GPU-time averages the three fitted seeds and includes collection, reference scoring, target storage, and adapter updates. It excludes shared SFT, data authoring, hyperparameter search, evaluation, and unsuccessful research runs. For extra training hours H and teacher/student seconds t_T, t_S on the same device, $N_{\text{break}} = \lceil 3600H/(t_T - t_S) \rceil$ when $t_T > t_S$; no finite saving is claimed otherwise. This assumes the workload represents future traffic and does not price answer quality. PDD-1's 2.69× ratio contrasts with only 1.04× for the prompt teacher, which needs about 21,200 requests to amortize training under the displayed assumptions. The break-even count is a projection based on the reported timing and training cost. Direct prompting remains attractive at low volume and has higher context JCS. These workload-specific ratios do not establish production throughput or hardware-independent gains.

Reproduction boundary. The clean package contains manuscript sources and aggregate summaries for rebuilding the tables and PDF. Evaluation replay additionally requires complete inputs, responses, scoring records, and an immutable judge snapshot; retraining requires executable code, checkpoint and data identities, conditional targets, and environment details. The mutable judge alias limits exact replay, and aggregate tables cannot replace token distributions or serving logs. Package contents and revision details are documented in the accompanying notes.

E Theory and Scope

E.1 Fixed-Target Sampling Design and KL Decomposition

Fix a profile pair, conversation, collection round, and cached teacher targets. Let $C \in \{0, 1\}$ index the queried endpoint, H a retained prefix, and V a next token. The prefix measures μ_0, μ_1 include collection and prefix sampling and are held fixed. Own-condition supervision uses $P_{\text{own}}(C = j, H \in dh) = \mu_j(dh)/2$; paired supervision uses $P_{\text{pair}}(C = j, H \in dh) = \bar{\mu}(dh)/2$, where $\bar{\mu} = (\mu_0 + \mu_1)/2$. The queried condition and history are therefore independent in the paired design, conditional on this fixed pair. A history may still reveal its source condition, but that source no longer determines which condition is supervised. This is a sampling statement, not an identified human personality effect or a total effect of changing the profile before generation.

At a fixed h, write q_j, s_j for the teacher and student token distributions. Define $Q(j, v) = q_j(v)/2$, $S(j, v) = s_j(v)/2$, $\bar{q} = (q_0 + q_1)/2$, and $\bar{s} = (s_0 + s_1)/2$. Their condition posteriors are $r_q(j \mid v, h) = q_j(v)/(q_0(v) + q_1(v))$ and r_s analogously. The KL chain rule in its two orders gives

$$\mathrm{KL}(Q\|S) = \tfrac{1}{2}\sum_j \mathrm{KL}(q_j\|s_j), \tag{11}$$

$$\mathrm{KL}(Q\|S) = \mathrm{KL}(\bar{q}\|\bar{s}) + \mathbb{E}_{v\sim\bar{q}}\,\mathrm{KL}\left(r_q(\cdot \mid v, h)\|r_s(\cdot \mid v, h)\right). \tag{12}$$

Equating these expressions proves Eq. 4. Zero teacher mass contributes zero; positive teacher mass with zero student mass yields infinite KL, with the identity understood in the extended sense. The second term fits the teacher's conditional dependence; it is neither the student's mutual information nor a requirement to maximize distinguishability. If $q_0 = q_1$, perfect fitting has $s_0 = s_1 = q_0$.

E.2 Condition-Blind Risk and Occupancy

For a condition-blind student $s_0 = s_1 = s$, the same identity reduces to

$$\tfrac{1}{2}\sum_j \mathrm{KL}(q_j\|s) = \mathrm{JS}(q_0, q_1) + \mathrm{KL}(\bar{q}\|s). \tag{13}$$

Thus the minimum is $\mathrm{JS}(q_0, q_1)$, attained at $s = \bar{q}$ (Lin, 1991). Integrating pointwise gives the minimum paired blind risk $\mathbb{E}_{\bar{\mu}}\,\mathrm{JS}(q_0^h, q_1^h)$. This minimum allows an unrestricted history-dependent student; for a capacity-constrained model it is only a lower bound.

Disjoint-history counterexample. Let $H \in \{a, b\}$, $\mu_0 = \delta_a$, and $\mu_1 = \delta_b$. For distinct token distributions u, w, set $q_0^a = q_0^b = u$ and $q_1^a = q_1^b = w$. The condition-blind rule $s(\cdot \mid a) = u$, $s(\cdot \mid b) = w$ has zero own-condition loss, whereas every condition-blind rule has paired loss at least $\mathrm{JS}(u, w) > 0$. A condition-aware rule fits all four targets exactly. This constructed separation shows that own-condition loss can permit a history shortcut; it does not show that a language-model optimizer chooses it.

For fixed-target losses $\ell_j(h) = \mathrm{KL}(q_j^h\|s_j^h)$, write $\mathcal{L}_{\mathrm{own}} = \frac{1}{2}\sum_j \mathbb{E}_{\mu_j}\ell_j$ and $\mathcal{L}_{\mathrm{pair}} = \frac{1}{2}\sum_j \mathbb{E}_{\bar{\mu}}\ell_j$. Expanding the mixture and subtracting gives

$$\mathcal{L}_{\mathrm{pair}} - \mathcal{L}_{\mathrm{own}} = \tfrac{1}{4}(\mathbb{E}_{\mu_0} - \mathbb{E}_{\mu_1})[\ell_1 - \ell_0]. \tag{14}$$

The population objectives coincide when $\mu_0 = \mu_1$; this does not imply identical finite-sample optimization. Also, $\mu_j \le 2\bar{\mu}$ implies $\mathbb{E}_{\mu_j}\ell \le 2\mathbb{E}_{\bar{\mu}}\ell$ for nonnegative ℓ. This covers the two collected prefix measures, not arbitrary unseen dialogue distributions.

E.3 Partial-Overlap Construction

Let $\eta_j(h) = P_{\mathrm{own}}(C = j \mid H = h)$. Pointwise optimization gives $\inf_s \mathcal{L}_{\mathrm{own}}^{\mathrm{blind}} = \mathbb{E}_{\bar{\mu}}\,\mathrm{JS}_{\eta(h)}(q_0^h, q_1^h) \le H_{\mathrm{own}}(C \mid H)$, where $\mathrm{JS}_\eta(q_0, q_1) = \sum_j \eta_j\,\mathrm{KL}(q_j\|\sum_k \eta_k q_k)$. The bound follows because weighted JS is a conditional mutual information. History that nearly determines its source can therefore permit small blind own-condition risk. Uniform JS need not exceed every weighted JS for asymmetric teachers; the following example is explicitly symmetric.

Set $\mu_0(a) = \mu_1(b) = 1 - \rho$ and $\mu_0(b) = \mu_1(a) = \rho$, where $0 \le \rho \le 1/2$. At both histories, let $q_0(V = 1 \mid h) = 1 - \epsilon$, $q_1(V = 1 \mid h) = \epsilon$, with $0 < \epsilon < 1/2$. For binary entropy $H_b(t) = -t\log t - (1 - t)\log(1 - t)$, the optimal blind risks are

$$R_{\mathrm{own}}^{\mathrm{blind}}(\rho, \epsilon) = H_b\big(\epsilon + (1 - 2\epsilon)\rho\big) - H_b(\epsilon), \tag{15}$$

$$R_{\mathrm{pair}}^{\mathrm{blind}}(\rho, \epsilon) = \log 2 - H_b(\epsilon). \tag{16}$$

The first identity follows by optimizing the mixture at each history and using binary-entropy symmetry; the second follows from Eq. 13. A condition-aware oracle has zero risk under either objective.

E.4 Coarsening and Interpretation Limits

The decomposition and blind-risk identity also hold after a common measurable coarsening of the teacher and student distributions, including the retained tokens plus residual bucket used in training.

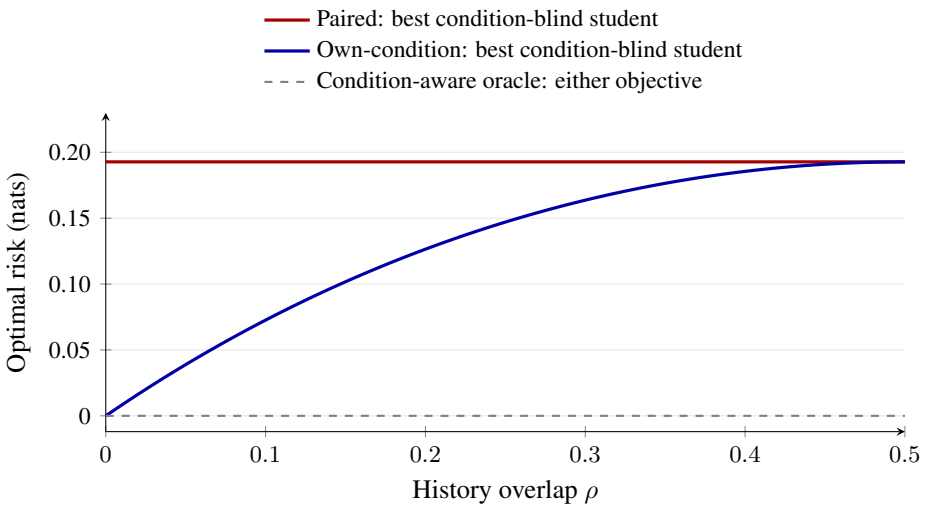


Figure 4: Exact risks in the two-history construction at $\epsilon = 0.2$. Increasing ρ makes history less informative about its source condition. At $\rho = 0$, blind own-condition risk is zero while paired risk has a positive floor; their blind optima coincide at $\rho = 1/2$. These are closed-form quantities, not model measurements.

Denote coarsened distributions by $\widetilde{q}, \widetilde{s}$. Data processing gives $\mathrm{JS}(\widetilde{q}_0, \widetilde{q}_1) \le \mathrm{JS}(q_0, q_1)$: a contrast entirely inside the residual bucket may disappear. For a fixed retained-token partition, the KL chain rule is

$$\mathrm{KL}(q\|s) = \mathrm{KL}(\widetilde{q}\|\widetilde{s}) + q(\perp)\,\mathrm{KL}\left(q(\cdot \mid \perp)\|s(\cdot \mid \perp)\right). \tag{17}$$

Thus coarse KL lower-bounds full KL; neither small coarse KL nor small residual mass supplies a full-vocabulary KL upper bound without controlling the conditional residual term.

With natural logarithms, if full-distribution $\mathrm{KL}(q_j\|s_j) \le \varepsilon_j$, Pinsker's inequality and the triangle inequality give

$$\left|\,\mathrm{TV}(s_0, s_1) - \mathrm{TV}(q_0, q_1)\right| \le \sqrt{\varepsilon_0/2} + \sqrt{\varepsilon_1/2}. \tag{18}$$

For coarse KL, the same statement holds only on the coarsened partition. It bounds contrast magnitude, not its semantic meaning or direction.

The theory excludes a zero-loss condition-blind solution when shared histories retain a positive teacher contrast. It does not prove that optimization otherwise uses a shortcut, that reduced shortcut use mediates the behavioral gains, or that paired supervision always improves generalization. Teacher error, incompatible histories, capacity limits, and occupancy changes across rounds remain outside the fixed-target argument; cross targets can also teach undesirable continuations. Future mechanism tests could separate marginal and dependence errors at identical compatible histories, vary history informativeness at comparable teacher contrast and budget, and check held-out natural contexts. Those tests have not been performed here, and aggregate behavior scores cannot reconstruct them.