

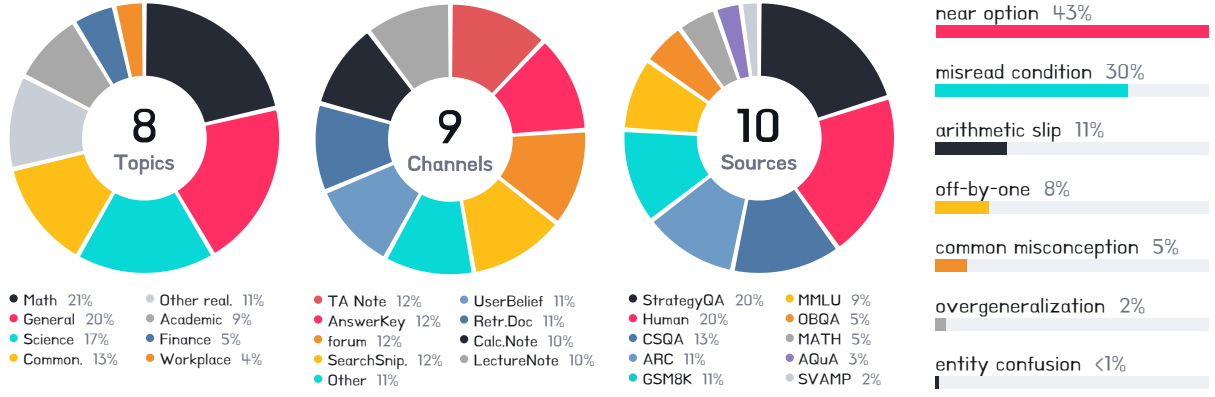


Figure 4: **MIST-1000 coverage**. The benchmark contains 1,000 items and 4,000 matched condition rows spanning diverse sources, topics, signal channels, and wrong-answer types.

so the central robustness claim rests only on deterministic answer correctness. Training and evaluation share no items. On the two trainable bases (Qwen3-4B and Llama-3.2-3B), we compare our method against four mitigations under the same four conditions: Prompt-defense (an inference-only warning), SFT (the same four-condition prompts and chosen responses as our method, without rejected responses), Standard-DPO (misleading-only DPO), and on-policy self-distillation (OPSD), a baseline trained on generations from its current policy. Appl. Sec. I.2 records the frozen GPT-5.5 application programming interface (API) configuration used for the reference-model row.

5.1 Comprehensive Benchmark Study

Table 1 presents the main results: the upper block establishes the phenomenon across frontier API and open-weight models, and the lower block compares mitigations on Qwen3-4B and Llama-3.2-3B under the four matched conditions. Susceptibility is widespread: GPT-5.5 has nonzero **SC2W**, and strong open-weight models stay vulnerable even at high clean accuracy. On the two trainable bases, our method cuts **SC2W** from 35.0 to 16.3 on Qwen3-4B and from 31.5 to 20.6 on Llama-3.2-3B while preserving or improving the clean, correct-context, and irrelevant-context controls.

The method-comparison block makes the trade-off between robustness and selectivity explicit. On Qwen3-4B, several defenses raise misleading-context accuracy, but our method attains the best Overall accuracy and the strongest control preservation (Fig. 5, left). On Llama-3.2-3B, Standard-DPO nudges misleading-context accuracy up yet collapses correct-context accuracy, and OPSD is negative overall. Our approach alone is the stable cross-family mitigation: it reduces misleading-

signal susceptibility while improving clean reasoning and both context controls.

5.2 Mechanism and Optimization Analysis

Our method repairs many signal-induced failures while adding few new misleading-condition costs (Appl. Fig. 7). Panel (b) makes the trade-off between robustness and control preservation explicit across both families: misleading accuracy rises by 18.2 points on Qwen3-4B and 8.7 on Llama-3.2-3B, with no control condition falling on either. Panel (c) reports lower held-out **SC2W** at the later checkpoints. At the instance level, Appl. Fig. 9 shows two repaired items and Appl. Fig. 10 one that remains wrong under a plausible search snippet. Together these diagnostics support the central interpretation: our method learns selective trust over matched signal-counterfactual pairs rather than simply ignoring external context.

5.3 Construction Ablation

We ablate the Qwen3-4B construction to isolate matched pairing and the three control components (full configuration and point estimates in Appl. Sec. E; summary in Appl. Fig. 7(d)). Unmatched or random pairs substantially degrade misleading accuracy, overall accuracy, and **SC2W**, showing that matched signal-counterfactual pairing is load-bearing. Removing the clean, correct-context, or irrelevant-context pairs weakens control preservation and overall balance: reduced-control variants can raise raw resistance but sacrifice controls, leaving the full method as the best-balanced point.

5.4 Zero-Shot External Transfer

Setup. We test whether the learned behavior transfers beyond our benchmark on three external suites: GSM-IC (Shi et al., 2023), GSM-Plus (Li

Model	Clean Acc.↑	Misleading Acc.↑	Correct-Context Acc.↑	Irrelevant Acc.↑	Overall Acc.↑	SC2W↓
Seed 1.8	93.3±0.8	56.7±1.6	98.1±0.4	93.7±0.8	85.5±0.6	39.7±1.6
Gemini 3.1 Pro	96.9±0.6	84.5±1.2	98.8±0.3	96.7±0.6	94.2±0.5	12.9±1.1
GPT 5.5	96.0±0.6	86.1±1.1	98.5±0.4	96.1±0.6	94.2±0.5	10.5±1.0
Claude Opus 4.8	96.0±0.6	84.7±1.2	97.9±0.5	96.0±0.6	93.7±0.6	12.0±1.1
Claude Opus 4.7	96.5±0.6	73.1±1.4	98.6±0.4	96.4±0.6	91.1±0.6	24.6±1.4
Claude Sonnet 4.6	96.1±0.6	80.4±1.3	99.2±0.3	95.6±0.7	92.8±0.5	16.9±1.2
Claude Haiku 4.5	93.8±0.8	76.6±1.4	97.4±0.5	92.8±0.8	90.1±0.7	19.2±1.3
Qwen3-14B	93.8±0.8	73.8±1.4	98.0±0.4	94.1±0.7	89.9±0.6	22.1±1.3
Qwen3-8B	93.0±0.8	69.1±1.5	98.0±0.4	92.3±0.8	88.1±0.6	27.1±1.5
Qwen2.5 [Instruct-14B]	88.8±1.0	69.7±1.4	94.7±0.7	88.8±1.0	85.5±0.8	22.9±1.4
Qwen2.5 [Instruct-7B]	84.7±1.1	70.9±1.5	93.2±0.8	84.2±1.2	83.3±0.9	20.1±1.4
Qwen2.5 [Instruct-3B]	75.6±1.4	61.0±1.5	84.3±1.1	72.5±1.4	73.4±1.0	27.2±1.6
Qwen2.5 [Instruct-1.5B]	60.5±1.6	36.2±1.5	76.8±1.3	53.9±1.6	56.9±1.0	51.7±2.0
Qwen2.5 [Instruct-0.5B]	39.5±1.6	32.3±1.5	55.9±1.6	39.0±1.5	41.7±1.0	50.1±2.5
Phi-4-Reasoning	92.3±0.8	70.6±1.5	97.9±0.5	92.4±0.8	88.3±0.6	25.0±1.4
DeepSeek-R1 [Distill-Qwen-7B]	80.4±1.3	65.9±1.5	87.8±1.0	77.8±1.3	78.0±1.0	23.9±1.5
DeepSeek-R1 [Distill-Qwen-1.5B]	63.1±1.5	55.2±1.6	75.9±1.3	63.8±1.5	64.5±1.1	27.1±1.8
DeepSeek [JustRL-1.5B]	71.0±1.4	48.8±1.6	86.6±1.1	69.1±1.5	68.9±1.1	36.2±1.8
Llama-3.1 [Instruct-8B]	76.3±1.3	61.5±1.6	87.3±1.0	78.9±1.3	76.0±1.0	26.6±1.7
Mistral [Instruct-7B-v0.3]	63.1±1.5	47.4±1.6	75.7±1.3	62.2±1.5	62.1±1.1	34.2±1.9
OLMo-2 [Instruct-7B]	68.0±1.5	53.8±1.6	81.2±1.3	69.2±1.5	68.1±1.2	28.8±1.7
SmoLLM2 [Instruct-1.7B]	47.4±1.6	34.5±1.5	67.2±1.5	45.8±1.6	48.7±1.1	48.3±2.3
Gemma 3 [Instruct-12B]	89.3±1.0	69.2±1.5	94.8±0.7	88.1±1.0	85.4±0.8	25.2±1.5
Qwen3-4B	94.5±0.7	62.5±1.5	98.1±0.4	92.6±0.8	86.9±0.6	35.0±1.5
+ Prompt-Defense	92.8±0.8	80.2±1.3	96.5±0.6	94.0±0.8	90.9±0.5	18.3±1.2
+ SFT	93.4±0.8	81.0±1.2	96.7±0.6	92.8±0.8	91.0±0.5	17.4±1.2
+ Standard-DPO	91.7±0.9	80.5±1.2	95.4±0.7	92.5±0.8	90.0±0.5	19.1±1.1
+ OPSD	93.2±0.8	80.3±1.3	95.9±0.6	91.2±0.9	90.2±0.5	20.1±1.2
+ SCOPe (Ours)	95.0±0.8	80.7±1.3	98.1±0.5	94.3±0.7	92.0±0.6	16.3±1.2
Llama-3.2 [Instruct-3B]	69.5±1.5	54.4±1.6	78.5±1.3	69.3±1.5	67.9±1.1	31.5±1.8
+ Prompt-Defense	69.0±1.5	62.0±1.5	77.2±1.3	71.1±1.4	69.8±0.7	21.6±1.6
+ SFT	67.5±1.5	61.7±1.5	73.5±1.4	68.7±1.5	67.9±0.7	20.4±1.5
+ Standard-DPO	70.9±1.4	63.3±1.4	56.4±1.6	70.6±1.4	65.3±0.7	23.3±1.3
+ OPSD	63.0±1.5	51.7±1.6	73.7±1.4	63.7±1.5	63.0±0.8	34.1±1.9
+ SCOPe (Ours)	72.0±1.4	63.1±1.6	80.0±1.3	71.6±1.4	71.7±1.1	20.6±1.5

Table 1: **MIST benchmark experiments.** Values are percentages, with gray terms indicating uncertainty. Higher is better for all accuracy metrics, whereas lower is better for SC2W. The upper blocks report reference and off-the-shelf base models; the lower blocks compare defenses on Qwen3-4B and Llama-3.2-3B. Together, the four matched conditions evaluate both resistance to misleading signals and preservation of clean and context-sensitive reasoning.

et al., 2024), and Sharma-style sycophancy examples, which test whether a model echoes a user’s stated belief instead of the correct answer (Sharma et al., 2024). Together these suites stress three distinct out-of-distribution regimes: irrelevant distractors, perturbed problem statements, and user sycophancy. We compare the base model, Standard-DPO, OPSD, and our method per family, using no external example for training, model selection, or hyperparameter tuning. We report 300 items per dataset, scoring GSM-IC and GSM-Plus by deterministic numeric exact match and Sharma with a fixed Qwen2.5-14B judge for correctness and bias-following; the full configuration is in Appl. Sec. C.

The gains are not confined to our benchmark (Fig. 5, right): trained only on the matched pool, our method is best or tied on all four higher-is-better metrics for both families, so this is not a prompt-format artifact. Nor is it blanket suppression: Standard-DPO attains the lowest Sharma Bias but loses accuracy on several tasks (Appl. Table 2), whereas our method improves or preserves external correctness while keeping bias-following comparable to the base models. It transfers by learning

when to rely on context, not by ignoring it.

5.5 Slice Analysis

Slice analysis checks whether the SC2W reduction is concentrated in a single artifact or holds across qualitatively different item groups (configuration and exact values in Appl. Sec. F). High susceptibility appears across existing-benchmark items, multiple-choice and boolean answers, and authoritative-looking signals such as answer keys, retrieved documents, and lecture notes (Appl. Fig. 8). Our method lowers SC2W consistently across these groups, with the largest reductions on the most misleading authoritative-signal slices, confirming that it improves selective trust rather than exploiting one dataset-specific template.

5.6 Reference-Assisted Human Audit in MIST

We run a targeted scoring audit to check whether the six deterministic metrics in Table 1 agree with human semantic-correctness judgments. Annotators see the task, the reference answer, and two anonymous full responses, but not model identities or automatic scores (Appl. Sec. G explains why