

Beyond Leaderboards: A Frozen-Rubric Contract for Position-Invariant Winning-Argument Detection

Zida Yang^{1,*,\dagger}, Weiqing Zeng^{2,*}, Guquan You³, Jiexin Zheng^{1,\ddagger}

¹Guanghua School of Management, Peking University, Beijing, China

²School of Computing and Artificial Intelligence,
Southwest Jiaotong University, Chengdu, China

³Institute of Emergency Safety Innovation, Guangzhou University, Guangzhou, China

2401111090@stu.pku.edu.cn, zengweiqing@my.swjtu.edu.cn

youguquan@gzhu.edu.cn, jiexinzheng@gsm.pku.edu.cn

* Equal contribution. † Project lead. ‡ Corresponding author.

Abstract

When large language models serve as evaluation infrastructure, progress cannot be reduced to a single accuracy number. We study winning-argument (WA) detection as a stress test: can one system jointly improve accuracy, enforce speaker-order invariance, emit rubric-conditioned diagnostics, improve a proper scoring rule, and cut inference cost? TAPIR freezes a survey-grounded rubric into a dual-artifact protocol, distills privileged signals by mode-split self-distillation, and predicts through an anti-symmetric scalar head paired with a structured ledger. An algebraic swap identity (Winning-Position Invariance, WPI=1) holds on every test pair, and a tied-order wrapper preserves it on the deployed router. On the standard 807-pair WA split, TAPIR beats our same-backbone Qwen3-14B re-implementation of Labruna et al.’s MS-PS-MLP, 65.80% vs. 58.12% (paired McNemar $p=3.41 \times 10^{-4}$), at a lower Brier score and $39.41\times$ faster inference per pair. Reusing the frozen protocol without retuning carries the mechanism to a second task and a second backbone, both at WPI=1. Under domain shift it only matches length-only controls, so we report contract preservation rather than superiority. The contribution is the joint contract under a paired protocol, with an explicit amortisation boundary, not a leaderboard delta.

1 Introduction

Ask an instruction-tuned judge which of two Change My View replies changed the original poster’s mind; then ask again with the two replies presented in the opposite order. On the standard 807-pair winning-argument (WA) test split, a single-pass Qwen3-14B judge gives an order-dependent answer on 35% of pairs: its choice does not follow the replies when they are swapped (empirical swap-consistency 0.65, Table 2). Reporting

accuracy under either ordering alone hides this entirely, and prompt engineering can reduce it but never rule it out.

That failure is the concrete form of a broader question for this theme track: when large language models stop being research prototypes and become *evaluation infrastructure*, what should count as progress? A judge for online persuasion is not useful merely because it is accurate on a held-out split; it must behave like a dependable component of a larger research or moderation system (Tan et al., 2016; Bozdag et al., 2026). Bavaresco et al. (2025) show that LLM–human agreement varies substantially across tasks, models, and evaluated properties, and conclude that judges must be validated against human judgements before being trusted. We ask the complementary question: *once* a human-validated rubric exists, what can the judge that consumes it be made to guarantee?

Four requirements. We answer with an infrastructure contract, stated before any notation. (i) *Exact swap behaviour*: reversing the speaker order must provably flip the verdict, not usually flip it. (ii) *No runtime rubric leakage*: the rubric is fixed and content-hashed before evaluation and never re-enters the model at test time. (iii) *Calibrated rejection*: confidence must be good enough, under a strictly proper score, to support selective deployment. (iv) *Inspectable diagnostics*: the system must expose a rubric-aligned account of what drove a verdict—scoped explicitly below faithful explanation. Prompted judges struggle on all four: they re-query an LLM at test time, aggregate scores externally, impose no algebraic relation between the two orders, and can rationalize an answer after producing it (Turpin et al., 2023), which makes a test-time explanation weak evidence about the verdict it follows (App. A quantifies this on our own

backbone).

TAPIR. We therefore move the rubric out of test-time prompting into train-time privileged supervision, freeze it before evaluation, and require the deployed head to obey the task symmetry by construction. The novelty of the resulting system, TAPIR (Theory-Aligned Persuasion-Invariant Rubric scoring), is not any single learning algorithm but the protocol-level integration of three components into a *jointly verified* contract.

(C1) Algebraic invariance with a structured audit ledger. An anti-symmetric scalar head $z_{\text{head}}(OP, r_A, r_B) = -z_{\text{head}}(OP, r_B, r_A)$ makes speaker-order invariance exact for the *verdict*: Winning-Position Invariance (WPI=1) is a structural guarantee, verified on all 807/807 WA test pairs, and the deployed router inherits it through a tied-order wrapper (§7.2). A ledger head decoded in the same paired forward pass exposes per-strategy rubric-conditioned diagnostics—useful for inspection, but *not* a human-validated strategy label or an account of the model’s internal reasoning (Limitations).

(C2) Frozen-rubric privileged distillation. The *Dual-Artifact Frozen Protocol* (DAFP) separates a frozen rubric $\mathcal{P}^{\text{trace}}$, built once from a public persuasion survey (Bozdag et al., 2026), from an operational input $\mathcal{P}^{\text{ops}}$ that excludes rubric text at runtime. A privileged teacher sees $\mathcal{P}^{\text{trace}}$ during training; the student sees only $\mathcal{P}^{\text{ops}}$ and learns by mode-split self-distillation (Vapnik and Izmailov, 2015; Zhang et al., 2019; Hinton et al., 2015), removing the test-time rubric leakage that recurs in LLM-judge work (Wang et al., 2024; Labruna et al., 2026).

(C3) Paired evidence, every claim scoped to its evidence tier. Table 1 is the map we ask readers to hold us to. On the same backbone and test split, TAPIR improves paired correctness over a faithful MS-PS-MLP re-implementation (65.80% vs. 58.12%), improves a strictly proper score, guarantees WPI=1, and runs $39.41\times$ faster per pair; reusing the frozen protocol without retuning carries the same mechanism to a second task and a second backbone (§8). Where the evidence is weaker we label it as such.

2 Background and Related Work

Winning-argument detection. The WA task originated with Tan et al. (2016), who released

a curated Change My View (CMV)¹ test split of 807 pairs in which one comment received a Δ (signal of opinion change) and a non- Δ baseline reply was matched on length and timing. Subsequent work studied argument persuasiveness through feature-based ranking (Habernal and Gurevych, 2016), reply-structure analysis (Durmus and Cardie, 2018), and multi-factor persuasion modelling (Durmus and Cardie, 2019). Labruna et al. (2026) extended the Tan et al. benchmark and proposed Multi-Strategy Persuasion Scoring (MS-PS), which obtains 12 strategy-level scores per pair (six per message) through 24 LLM calls, augments them with per-message aggregate statistics, and feeds the resulting 18-dimensional vector to an MLP. Their best result (64.53% with o3) is the published SOTA on this split.

LLM judges, paired scoring, and positional bias. Pairwise LLM judges have known positional and self-preference biases (Wang et al., 2024; Labruna et al., 2026). Mitigations include permutation, voting, and ranking heuristics, but add cost and can leave residual bias. Classical paired-comparison models use score differences (Bradley and Terry, 1952); we use the same anti-symmetric principle inside a neural judge. Because rationales can be unfaithful (Turpin et al., 2023; Jacovi and Goldberg, 2020), the ledger is diagnostic output, not an explanation.

Validating a rubric versus enforcing one. Bavaresco et al. (2025) evaluate 11 LLMs against human annotations on 20 datasets and find that agreement varies substantially with task, model, and judged property, concluding that judges must be validated against humans before use. That work and ours address different halves of one problem: theirs is about *constructing and validating* a rubric, ours begins once a rubric is settled and asks what the judge consuming it can be made to *guarantee*. Their finding is why we do not claim our frozen rubric is correct for every community, and why the ledger is scoped as a diagnostic rather than a validated strategy label.

Calibration and selective deployment. Accuracy alone is insufficient: confidence must be calibrated (Guo et al., 2017; Naeini et al., 2015; Desai and Durrett, 2020) and selective predictors should

¹<https://www.reddit.com/r/changemyview/> (accessed 2026-08-26); the underlying public Reddit archive is distributed through Pushshift (Baumgartner et al., 2020).

Claim	Primary evidence	Why this test	Scope and limits
H1. Paired correctness improves	65.80 vs. 58.12%, +7.68 pp, McNemar exact $p=3.41 \times 10^{-4}$	Same instances, correlated binary outcomes	Qwen3-14B, WA test ($n=807$), vs. same-backbone MS-PS re-run
H2. Probability quality improves	$\Delta \text{Brier} = -0.0138$, CI $[-0.0182, -0.0092]$, $p_{\text{boot}} < 10^{-4}$	Strictly proper score; confidence drives selective use	ECE is a binning-sensitive complement, not pair-tested
H3. The gain is not one stratum	Stratified paired effects under Bonferroni and BH-FDR, pre-declared 12-test family	Secondary endpoints must be multiplicity-aware	Positive on all 4 topics (2 significant) and, descriptively, 3 length tertiles
H4. Order behaviour is exact	Algebraic swap identity, 807/807, $ \varepsilon_{\text{swap}} =0$	Structural claim, not a significance test	Applies to the verdict; ledger cells are not anti-symmetric
The protocol reuses across task and backbone	SummEval-pair +4.7 pp; Gemma-3-12B +3.4 pp; WPI=1 in both	Paired CIs excluding 0, no retuning	Two transfer cells; <i>not</i> universal generality
The ledger is externally useful	$\kappa=0.46$ (human-human ceiling 0.51); error AUROC 0.63→0.71	Agreement plus error-flagging utility	<i>Not</i> a faithful account of model internals

Table 1: Claim–evidence–scope map. Rows above the rule are the pre-declared primary hypotheses with the test chosen for each (full multiplicity accounting in App. K). Two claims are deliberately weaker and are marked as such wherever they appear: out-of-domain behaviour is *contract preservation*, not accuracy superiority (§8.2), and the comparison with the closed OpenAI-o3 cell is *unpaired context* (§6.1). Effect sizes precede p -values throughout.

track it rather than maximise naive top-1 (Geifman and El-Yaniv, 2017; Kadavath et al., 2022), a concern reinforced by surveys on hallucination (Ji et al., 2023) and uncertainty (Lin et al., 2022).

Domain knowledge as a train-time-only resource. Curated human knowledge is known to improve learned text models where purely data-driven training plateaus, both as expert signals for sentiment in corporate disclosures (Ng et al., 2026) and as models of how sentiment-bearing usage evolves rather than static dictionaries (Zheng et al., 2024). There the knowledge is a permanent input, engineered into features and available whenever the model runs; we invert its availability. Combining self-distillation (Zhang et al., 2019; Hinton et al., 2015) with learning under privileged information (Vapnik and Izmailov, 2015), we produce privileged signals *once* from a public survey, freeze them before the test split is touched, and withhold them at inference. Withholding is what removes the test-time rubric leakage recurring in LLM-judge work (Wang et al., 2024) and what keeps the deployed judge cheap.

3 Problem Setting and Notation

A WA *pair* consists of an opening post OP and two replies r_A, r_B with a binary label $y \in \{A, B\}$ recording which reply received the Δ in the original CMV thread. The scalar head produces $z_{\text{head}}(OP, r_A, r_B) \in \mathbb{R}$ and predicts A iff $z_{\text{head}} > 0$; the deployed router is reported separately throughout. The head has *Winning-Position Invariance* (WPI = 1) iff $z_{\text{head}}(OP, r_B, r_A) = -z_{\text{head}}(OP, r_A, r_B)$ for every pair; for judges

without that algebraic property we report empirical WPI, the fraction of pairs whose prediction flips correctly under the speaker swap. Two artifacts, both frozen and content-hashed (SHA-256) before any test access, define the protocol: the frozen rubric $\mathcal{P}^{\text{trace}}$ supplying privileged survey context, and the operational input $\mathcal{P}^{\text{ops}}$ containing only OP, r_A, r_B at runtime, so that no test-time rubric text can reach inference. The ledger label schema uses K MS-PS strategy categories for benchmark compatibility.

4 The TAPIR Method

Figure 1 lays out the protocol; the subsections below formalise each stage in turn.

4.1 Anti-Symmetric Dual Output

We attach two heads on top of a shared Low-Rank Adaptation (LoRA; Hu et al., 2022) augmented Qwen3-14B backbone.

Scalar head. We replicate the forward in two orderings (OP, r_A, r_B) and (OP, r_B, r_A) . Let f_θ denote the shared encoder and $\phi : \mathbb{R}^d \rightarrow \mathbb{R}$ a learned scalar projection on the verdict-position hidden state. Writing $\phi_{AB} \triangleq \phi(f_\theta(OP, r_A, r_B))$ for brevity, we define

$$z_{\text{head}} = \phi_{AB} - \phi_{BA}.$$

Anti-symmetry $z_{\text{head}}(OP, r_B, r_A) = -z_{\text{head}}(OP, r_A, r_B)$ follows by elementary algebra for any ϕ and f_θ , so WPI = 1 is a structural guarantee, not an empirical finding. We confirm as a sanity check that bf16 rounding introduces zero residual ($|\varepsilon_{\text{swap}}|=0$) on all 807 test

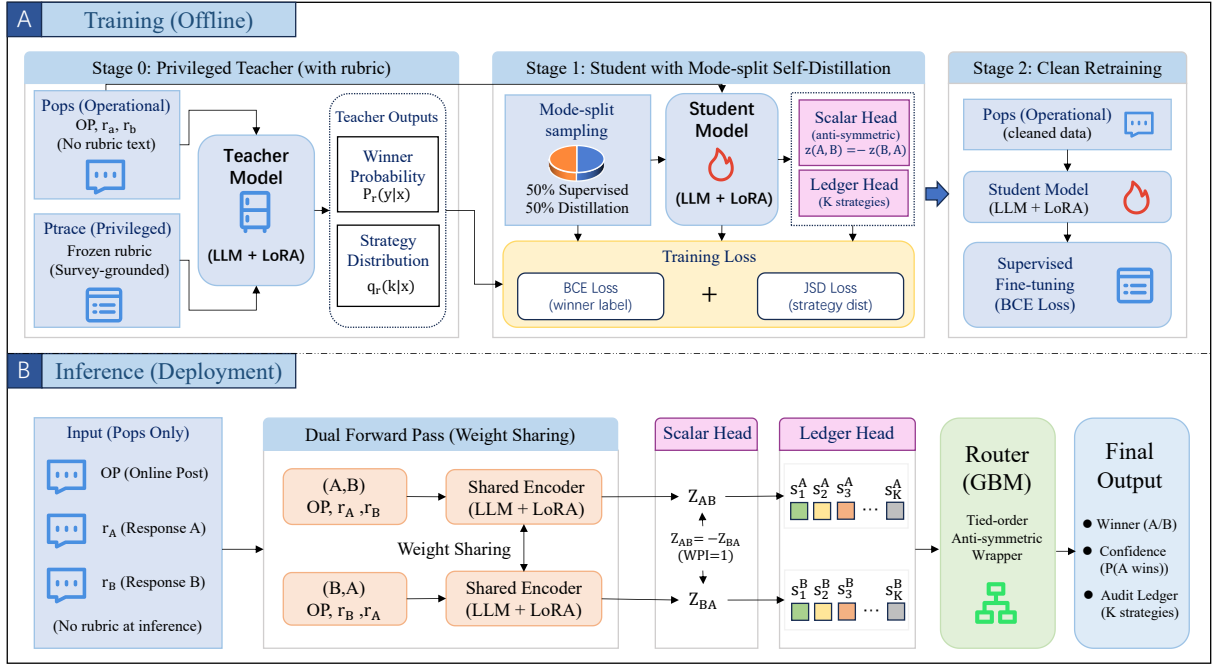


Figure 1: The TAPIR protocol, in four stages. A *frozen rubric* $\mathcal{P}^{\text{trace}}$, built once from a public survey and SHA-256 hashed before any test access, supervises a *privileged teacher* (Panel A); its soft targets are distilled into a *rubric-free student* that sees only $\mathcal{P}^{\text{ops}}$; and the deployed student emits *invariant outputs* (Panel B). The rubric is visible only inside Panel A—none is read at inference. Tied dual forwards make WPI=1 structural, and the anti-symmetric wrapper carries it to the router, which returns a verdict, a calibrated $P(A \succ B)$, and a K -cell audit ledger (a rubric-conditioned diagnostic, not an account of model internals). Losses: §4.3; router features and Stage-2 detail: App. E.

pairs (§7.2), verifying that both orderings share the identical computation graph.

What ϕ means, and why the difference is the right object. $\phi(OP, r_A, r_B)$ is an *order-conditional latent potential*: the value the encoder assigns to one particular *presentation* of the pair. It carries no standalone reading such as “the quality of r_A ” or “the score that r_A beats r_B ”, and it is not separately identifiable. Decompose any scoring function on ordered pairs into swap-odd and swap-even parts, $s_{AB}^\mp = \frac{1}{2}(s_{AB} \mp s_{BA})$. Adding any swap-even function to ϕ leaves z_{head} unchanged, so only the swap-odd component is identifiable from order-consistent verdicts—and $z_{\text{head}} = 2\phi_{AB}^-$ extracts exactly that component, the factor of two being absorbed by the temperature calibrator of §4.4. Discarding ϕ^+ also costs nothing in risk:

Proposition 1. *Let the evaluation distribution be swap-balanced—each pair presented in both orderings with equal weight and correspondingly negated signed label $y \in \{\pm 1\}$ —and let ℓ be any convex margin loss. Then the two-order average risk $R(g) = \frac{1}{2}\mathbb{E}[\ell(y g_{AB}) + \ell(-y g_{BA})]$ satisfies $R(s^-) \leq R(s)$ for every s .*

The proof is one application of Jensen’s inequality

ity to $y s_{AB}^- = \frac{1}{2}(y s_{AB}) + \frac{1}{2}(-y s_{BA})$ (App. B). Anti-symmetrisation is thus the projection onto the swap-odd class: exact in behaviour and, up to that calibrated scale, free in risk. We do *not* claim the subtraction is a novel learning algorithm—it is a reusable operator, and §7.3 reports what it does and does not buy on its own.

Ledger head. The same paired forward pass decodes a structured K -category ledger, one cell per MS-PS strategy label, supervised through the frozen survey context: the student predicts $\hat{q}^k$ over $\{A, B, \text{tie}\}$ per strategy, and aggregating cells gives a second decision channel (§7.3). The algebraic guarantee covers the scalar verdict only. Ledger cells are discrete diagnostics, are not anti-symmetric, and may vary across orderings; we surface the canonical (A, B) ledger and log both orderings for inspection (App. H).

4.2 Dual-Artifact Frozen Protocol (DAFP)

DAFP separates two roles a rubric usually plays at once. $\mathcal{P}^{\text{trace}}$, the *frozen* rubric, is extracted from six survey blocks (App. D), content-hashed, and frozen *before* any test access; it is seen only by the privileged teacher, and the student ledger emits

the $K=6$ MS-PS strategy labels of this benchmark family (Labruna et al., 2026). $\mathcal{P}^{\text{ops}}$, the *operational* input, is restricted to (OP, r_A, r_B) plus a frozen feature transform that consumes no rubric text at runtime. This instantiates the Vapnik–Izmailov “two-window” principle of learning under privileged information (Vapnik and Izmailov, 2015): $\mathcal{P}^{\text{trace}}$ is the train-time window, $\mathcal{P}^{\text{ops}}$ the test-time window, and the two are decoupled by construction so that no rubric can leak through inference. That frozen separation is the qualitative difference from prompt-engineered MS-PS methods, which re-prompt a rubric at test time; both artifacts are versioned and SHA-256-hashed, and readers can verify them against the released manifest.

4.3 Mode-Split Privileged Self-Distillation

We train the student in three stages on the WA train split, with all hyper-parameters frozen before any inner-dev evaluation.

Stage 0 (privileged teacher). A LoRA-augmented teacher consumes the privileged window $(OP, r_A, r_B, \mathcal{P}^{\text{trace}})$ and emits soft targets $\hat{p}_i^A \in [0, 1]$ and per-strategy categorical distributions $\hat{q}_i^k$ over $\{A, B, \text{tie}\}$, $k=1..6$. The teacher is trained to convergence and *frozen* before Stage 1.

Stage 1 (mode-split self-distillation). The student consumes only the operational window (OP, r_A, r_B) and defines $\hat{p}_i^A = \sigma(z_{\text{head},i})$. It is trained for E_1 epochs with a mode-split objective. At every optimisation step a mode is drawn *online* with probability 1/2: in the *label mode* the step optimises binary cross-entropy (BCE) against the gold label y , and in the *distill mode* it optimises the Jensen–Shannon divergence (JSD) to the frozen teacher’s Bernoulli (Bern) scalar target and its per-strategy categorical (Cat) targets. No example is permanently assigned to one objective: because the mode is resampled each step, an example seen once per epoch receives both gold-label and teacher updates with probability 3/4 over three epochs, with equal expected exposure to each. In expectation over the mode draw,

$$\begin{aligned} \mathcal{L}_{\text{stage1}} = & \frac{1}{2} \mathcal{L}_{\text{BCE}}(\hat{p}_i^A, \mathbf{1}[y_i=A]) \\ & + \frac{1}{2} \left[\text{JSD}(\text{Bern}(\hat{p}_i^A) \parallel \text{Bern}(\tilde{p}_i^A)) \right. \\ & \left. + \frac{1}{K} \sum_{k=1}^K \text{JSD}(\text{Cat}(\hat{q}_i^k) \parallel \text{Cat}(\tilde{q}_i^k)) \right]. \end{aligned}$$

Why split the modes instead of summing the losses? A per-sample weighted sum has the same

objective in expectation, so what differs is the geometry of each update. Gold labels and teacher targets disagree on a non-trivial fraction of pairs, and when they do their gradients partly cancel *inside* a single step: $\cos(g_{\text{BCE}}, g_{\text{JSD}}) < 0$ on 33% of early-training steps. Alternating lets each step follow one signal cleanly rather than a partially self-cancelling average. This within-step cancellation—not a general variance-reduction argument—is the narrower claim we can evidence, and a compute- and exposure-matched three-seed control supports it: the weighted sum reaches 63.4% against mode-split’s 65.1% ($\Delta=+1.7$ pp, CI excluding 0), with a plateau within 0.6 pp for ratios from 40:60 to 60:40 (App. Q). The ledger signal $\hat{q}_i^k$ flows through the same bounded distillation loss rather than a separately scheduled auxiliary head (Zhang et al., 2019).

Stage 2 (clean inner-dev retrain). The Stage 1 student is fine-tuned with BCE only on a train-only inner-dev subset from which every pair whose SHA-256 hash matches a test pair has been removed; canonical validation and test labels are never used here. This removes a possible CMV-overlap confound before the student is frozen.

4.4 Post-hoc Calibration and Router

After Stage 2, three light-weight calibrators—a temperature scaler on z_{head} , a gradient-boosted machine (GBM) router combining the scalar head with the ledger majority, and a Brier-optimal blend—are fitted on validation data only and selected by strict 5-fold cross-validated accuracy, with no test access. The deployed model is the GBM router, which consumes only operational student outputs at inference; the unrouted z_{head} is reported in ablations.

5 Experimental Setup

Backbone and hardware. All training and inference uses Qwen3-14B (Yang et al., 2025) with FlashAttention-2 (Dao, 2024) on 2×NVIDIA A800-SXM4 80 GB, in bf16 with LoRA (Hu et al., 2022) and AdamW (Loshchilov and Hutter, 2019); full settings are in App. E.

Splits and topic labels. Train/inner-dev/test follows the curation of Labruna et al. (2026); test is the unmodified Tan et al. 807-pair split. Topic-stratified analysis uses their four high-level topic labels (sizes in App. J). A *clean inner-dev retrain* hashes every train pair and excludes any colliding with the test split, eliminating an overlap confound.

Baselines. We compare against (i) *Random* (empirical, $397/807 = 49.19\%$); (ii) *Direct zero-shot* on Qwen3-14B; (iii) our *MS-PS-MLP re-run on Qwen3-14B*—a faithful re-implementation with two-step prompting aligned to the paper, 24 LLM calls per pair, and the MLP hyper-parameters of their Appendix O (sweep in App. F); and (iv) verbatim numbers from their Table 1 for 5 external LLMs under all 6 MS-PS configurations (30 cells, App. C).

Statistical reporting, and why each test is here. The infrastructure claim dictates the evidence, not the reverse. A judge is deployed on the *same* items as its predecessor, so accuracy must be compared paired on instances rather than as a difference of independent rates: McNemar’s exact test (McNemar, 1947) with paired-bootstrap 95% confidence intervals (CIs) over 10,000 resamples (Efron and Tibshirani, 1993). A judge whose confidence drives abstention must be scored by a strictly proper rule: Brier as the pair-tested endpoint (Guo et al., 2017; Naeini et al., 2015), with expected calibration error (ECE) as a binning-sensitive complement only. Examining several strata requires family-wise control: Bonferroni and Benjamini–Hochberg false discovery rate (FDR) control (Benjamini and Hochberg, 1995) over a pre-declared 12-test family (App. K). A structural property, finally, should be proved rather than tested: WPI is verified algebraically. Table 1 maps each hypothesis to its test.

Stylistic-typicality filter. Before any model-selection decision consults the test split, we score every WA test pair with a base-model perplexity probe (Ji et al., 2023). A conservative double-condition filter flags 111/807 pairs (13.8%); on the remaining 696 the paired gain over the re-run is +5.89 pp (95% CI [+1.44, +10.34], $p=1.14 \times 10^{-2}$). All three same-backbone systems improve on the flagged subset, so the flag tracks stylistic typicality rather than contamination (App. L); we read it as a robustness stratification, not a leakage certificate.

6 Main Results

6.1 Controlled, Paired Comparison

The primary evidence holds the backbone fixed. On the identical 807 test pairs of the unmodified Tan et al. split (same 397/410 A/B label distribution) and the same Qwen3-14B weights, replacing the 24-call MS-PS-MLP procedure with one

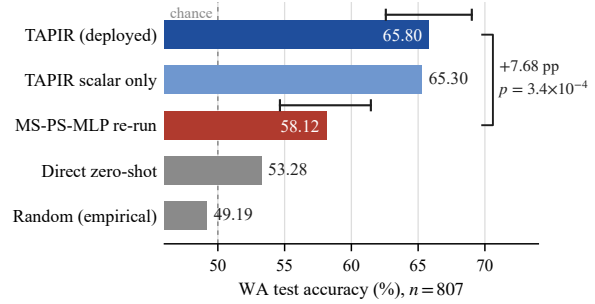


Figure 2: Controlled, paired evidence on the WA test split ($n=807$); same Qwen3-14B backbone throughout, single-method 95% bootstrap CIs. The 30-cell published grid is unpaired context and is reported separately in App. C.

tied-order TAPIR evaluation moves accuracy from 58.12% to 65.80%: a paired gain of +7.68 pp, with non-overlapping single-method 95% bootstrap CIs ([62.58, 69.02] vs. [54.65, 61.46]) and McNemar exact $p=3.41 \times 10^{-4}$ (Figure 2). We report seed 42; the three-seed mean is 65.1%, SD 0.7 pp (App. O). The 58.12% baseline is the *strongest* of five MLP seeds (five-seed mean 55.17%), so the gap is conservative.

Where this sits, and what that does not license.

The strongest published cell on this split is OpenAI-o3 + MS-PS-MLP at 64.53% (Labruna et al., 2026). TAPIR’s 65.80% is numerically inside that band, but the o3 cell releases no per-pair predictions, so no paired test is possible and we assert no superiority (App. C). A proprietary two-order rubric judge we queried reaches 66.4%, 0.6 pp above TAPIR—again unpaired context, and well inside the resolution of the comparisons we can actually run. The controlled contrast establishes that the *method* rather than the backbone’s raw capability produces the gain; it does not establish a leaderboard win, and we do not report one.

6.2 Robustness Across Topics and Length

Is the controlled gain concentrated in one sub-population? Figure 3 says no. The topic-stratified paired gap is positive on all four topics; it survives Bonferroni at $\alpha=0.05/12$ on *Gender/Sexuality/Minority Rights* ($n=233$, +14.16 pp, $p=8.10 \times 10^{-4}$) and BH-FDR at $q=0.05$ on *Food & Culture* ($n=166$, +12.05 pp, $p=1.05 \times 10^{-2}$), while the other two are positive but not significant under either correction; Brier is nonetheless the lowest on every topic (App. J). By maximum reply-length tertile the gap is +2.23/ +

10.04/ + 10.78 pp (Q1/Q2/Q3). Neither result establishes immunity to surface cues—App. R tests that directly under strict stylometric matching—but the gain is not confined to a single stratum of either kind.

7 Analysis

Table 2 consolidates the headline numbers; we discuss each row through this section.

7.1 Inference Latency: $39.41\times$ Speedup

On $1\times A800$ in bf16 without cross-order caching, full deployed inference—dual AB/BA forward, scalar head, ledger decoding, and router—costs 0.846 s per pair ($n=200$ profiled pairs, 95% CI [0.811, 0.882]), against 33.33 s for the same-backbone 24-call MS-PS procedure: a $39.41\times$ speedup (95% CI [$37.78\times$, $41.08\times$] over 10,000 paired-bootstrap resamples; App. Figure 7a), and conservatively so, because MS-PS additionally requires fitting an MLP head per backbone. With the accuracy gain of §6.1 this is a Pareto improvement; §8.3 states when it repays the training cost.

7.2 Algebraic and Deployed-Router WPI

Scalar-head WPI is algebraic. The identity $z_{\text{head}}(OP, r_A, r_B) = -z_{\text{head}}(OP, r_B, r_A)$ holds on every WA test pair with maximum residual $|\varepsilon_{\text{swap}}| = 0$ at bf16 precision: 807/807 flip correctly. Two natural alternatives only approximate it: a double-pass swap-and-average ensemble reaches ≈ 0.92 and a single-pass zero-shot judge ≈ 0.65 , consistent with the positional bias reported by Wang et al. (2024) (App. Figure 7b).

The deployed router inherits the guarantee.

The ledger head is a discrete decoder and is *not* anti-symmetric a priori—empirically it flips on only 58/100 pairs—so a router that consulted it naively would forfeit the guarantee. We therefore wrap the GBM in a tied-order anti-symmetric transform, $R_{\text{anti}}(F_{ab}, F_{ba}) = R(F_{ab}) - R(F_{ba})$, falling back to z_{head} when $R_{\text{anti}}=0$. This is exactly anti-symmetric by construction, so the deployed verdict flips on all 807 pairs; Appendix H gives the derivation and Appendix I the end-to-end two-order check on a stratified 100-pair sample (100/100 for both the scalar head and the wrapper).

7.3 Components, and What the Bare Operator Buys

The scalar, ledger, and router channels reach 65.30/64.81/65.80%, all far above no-distillation (60.35%; App. P and App. Figure 6); removing anti-symmetry, mode-split, or ledger distillation costs $-6.0/ -1.9/ -1.7$ pp, and a DeBERTaV3-large cross-encoder reaches only 58.4% (He et al., 2023), a smaller-backbone control against a pure large-backbone effect (App. Q).

The anti-symmetric operator alone is not enough. Since Proposition 1 applies to *any* scoring function, the same subtraction can be applied post hoc to a prompted judge. On the identical backbone that yields exact WPI=1 but only 57.6% accuracy at Brier .248, against TAPIR’s 65.80% and .233. The operator is therefore a general, reusable contract rather than a TAPIR-specific trick—but on its own it delivers neither rubric internalisation, nor calibration, nor diagnostics.

The ledger is useful, at a moderate ceiling.

Two annotators labelling the dominant strategy on 200 test pairs agree with the ledger at $\kappa=0.46$ —moderate in absolute terms, but close to the achievable ceiling, since the annotators agree with *each other* at only $\kappa=0.51$ (difference 0.05, bootstrap 95% CI [-0.02 , 0.12]). Persuasion-strategy attribution is simply ambiguous. What matters for deployment is external usefulness: adding ledger features to the scalar score raises error-detection AUROC from 0.63 to 0.71, retaining on ledger evidence-validity rather than raw confidence raises selective accuracy at 70% coverage from 70.0% to 72.1%, and a ledger anomaly multiplies the odds of an error by 2.4 (CI excluding 1). We therefore claim an externally useful rubric-conditioned diagnostic, and explicitly not a faithful account of internals (App. S).

8 Generality and the Lifecycle Boundary

A single task-and-backbone cell cannot establish a reusable contract. We therefore reuse the frozen protocol—identical hyper-parameters, no test-set search, no retuning—in three senses of “transfer”, kept separate because they carry different strengths of evidence (Table 3).

8.1 A Second Task and a Second Backbone

Framework transfer. We instantiate the same three-stage protocol on a pairwise version of

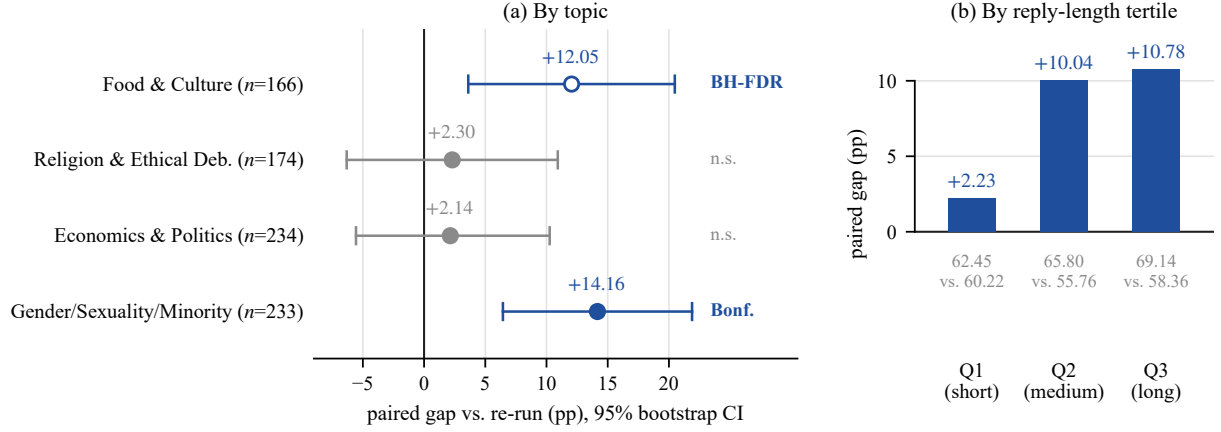


Figure 3: Stratified paired gains over the same-backbone re-run. **(a)** By topic: forest plot of the TAPIR – re-run accuracy gap with paired-bootstrap 95% CIs. Filled: Bonferroni-significant ($\alpha=0.05/12$); hollow: BH-FDR-significant only ($q=0.05$); grey: neither. **(b)** By maximum reply-length tertile (Q1–Q3), with absolute accuracies below each bar; these strata are outside the pre-declared 12-test family and descriptive only. Panel (a) tests $H_0: \text{Acc}_{\text{TAPIR}} - \text{Acc}_{\text{re-run}} = 0$; the gap is positive in every stratum.

System	Acc % $\uparrow$	Brier $\downarrow$	ECE $\downarrow$	z_{head}	WPI $\uparrow$	Δ vs. o3 (unpaired)	paired p vs. rerun	Bonf. sig.?	Protocol
Direct zero-shot (Qwen3-14B)	53.28	.248	.033	—	0.65	−11.25	—	—	zero-shot
Labruna et al. (2026), Gemma-3-12B (best non-o3)	63.69	—	—	—	—	−0.84	—	—	published
OpenAI-o3 + MS-PS-MLP (published best)	64.53	—	—	—	—	0.00	—	—	published, unpaired
MS-PS-MLP re-run on Qwen3-14B (ours)	58.12	.247	.059	—	—	−6.41	—	—	same backbone
TAPIR no-distillation ablation (ours)	60.35	.232	.042	1.00	—	−4.18	3.42×10^{-1}	no	BCE direct (no distill)
TAPIR (z_{head} only, ablation)	65.30	.233	.102	1.00	—	+0.77	6.92×10^{-4}	yes	unrouted
TAPIR (deployed router)	65.80	.233	.100	1.00	—	+1.27	3.41×10^{-4}	yes	deployed

Table 2: Headline comparison on the WA test split ($n=807$); “—” marks values that cannot be computed because per-pair predictions were not released. The Δ column is relative to OpenAI-o3 + MS-PS-MLP (64.53%) and is *unpaired for every row*, since that cell released no per-pair predictions; the paired evidence is the p column instead. WPI is the scalar-head swap identity; the router preserves WPI=1 via the tied-order wrapper (App. H). Paired p : McNemar exact against the same-backbone re-run, Bonferroni $\alpha=0.05/12$ (App. K). Brier is the pair-tested endpoint: $\Delta\text{Brier} = -0.0138$, 95% CI $[-0.0182, -0.0092]$, $p_{\text{boot}} < 10^{-4}$. ECE uses 15 equal-width bins and carries no paired claim, since a sharper predictor can incur higher binned ECE (Murphy, 1973). All TAPIR rows are seed 42; three-seed means in App. O.

Transfer	Cell	Base	TAPIR	WPI
Framework	SummEval-pair, direct	58.4	63.1	1.00
	SummEval-pair, no-distill	60.2	63.1	1.00
Backbone	WA, Gemma-3-12B MS-PS	61.8	65.2	1.00

Table 3: Protocol reuse across a task and a backbone, both under the frozen WA recipe with no retuning; accuracies in %. *Base* is the same-backbone control the cell is paired against: a direct prompt (rows 1, 3) or the no-distillation ablation (row 2). Bold marks a paired-bootstrap CI on the gap excluding 0. The third sense of transfer—the frozen model under domain shift—is deliberately kept separate in Figure 4, because its evidence is weaker.

SummEval (Fabbri et al., 2021), a within-document summary-preference task whose rubric—coherence, consistency, fluency, relevance—is human-validated rather than survey-derived; splits are by source document and the bootstrap is document-clustered. Against a direct prompt on the

same backbone, TAPIR moves accuracy $58.4\% \rightarrow 63.1\%$ (+4.7 pp, CI excluding 0) and Brier .251 $\rightarrow$.236 at WPI=1; against the no-distillation ablation the gain is +2.9 pp, and rubric-at-test multi-call baselines remain $\geq 10\times$ slower. The informative result is not the absolute number but that the *ordering of the components is preserved*: full > no-distillation > direct, with no sign flip. That is what distinguishes a protocol from a WA-specific recipe.

Backbone transfer. Holding the WA task and every hyper-parameter fixed, we swap both backbones to Gemma-3-12B (Gemma Team et al., 2025). Against the same-backbone MS-PS baseline the deployed router improves $61.8\% \rightarrow 65.2\%$ (+3.4 pp, CI excluding 0) at WPI=1 and $27.6\times$ faster inference, with a distillation contribution of +3.5 pp against Qwen’s +4.9 pp over three seeds. Both families stay positive and within ~ 2 pp of each other, so the mechanism is not one family’s

artefact—and this cell is a second fully public, fully paired contrast.

8.2 Frozen-Model OOD: The Weakest Claim, Stated as Such

Labruna et al. (2026) transfer MS-PS features only by re-training a fresh MLP on 2,757 in-domain Anthropic samples. We ask a stricter question: does the *same frozen WA judge*, with no target-domain supervision, keep its contract? On Anthropic/persuasion (Durmus et al., 2024) paired by same-claim rating gap ≥ 1.0 (591 pairs), TAPIR reaches 60.74% (95% CI [56.85, 64.64]), +10.66 pp over a prompted Qwen3-14B ($p=1.84 \times 10^{-4}$). That is where the honest reading stops: because Anthropic ratings couple persuasiveness with reply length, a length-only rule (61.42%) and a WA-trained surface GBM (60.91%) land in the same band, and strict length matching does not separate them (App. M). We therefore call this *contract-preserving transfer* (Figure 4): across the shift TAPIR alone retains the frozen ledger, algebraic WPI=1, calibrated confidence, and $39.41 \times$ faster inference—not accuracy superiority over shallow cues.

Semantic signal nevertheless sits underneath that correlation. We build a *cue-neutral* slice ($n=312$) using no TAPIR prediction, balancing labels within length and surface strata (App. M). With the shallow correlation neutralised the controls collapse towards chance—length-only 50.3%, surface GBM 52.6%—while TAPIR holds 58.7%, paired CIs on both gaps excluding 0: a complementary signal is detectable, but only once shallow cues are removed. Built after seeing the aggregate result, this slice is exploratory evidence, not confirmation.

8.3 When Is the Training Cost Worth It?

The $39.41 \times$ figure is a *marginal-inference* number, not a lifecycle claim. From the measured 43,200 GPU-seconds per training seed and 32.484 GPU-seconds saved per pair against the 24-call MS-PS baseline (App. N), one production seed breaks even at $\approx 1,330$ pairs and three research seeds at $\approx 3,990$; at 10k pairs total compute is $\approx 6.5 \times$ lower for one seed, $\approx 2.4 \times$ for three. These thresholds are relative to a *local multi-call* baseline; a single-call API judge has a different cost structure and a higher break-even. Rubric maintenance stays cheaper than retraining: a paraphrase leaving the taxonomy intact costs -0.3 pp, and a light merge-or-delete edit recovers 96.2% of accuracy for 28% of the com-

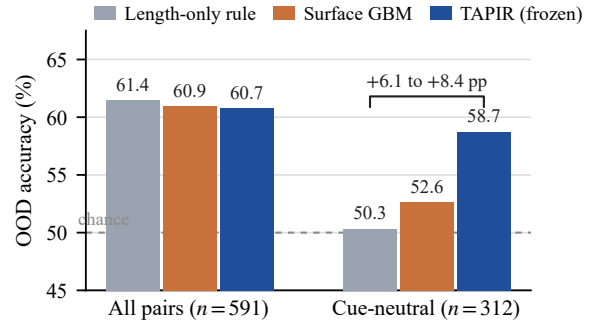


Figure 4: Frozen-model OOD transfer on Anthropic/persuasion, no target-domain supervision. On all 591 same-claim pairs the three systems fall in one band, no pairwise difference reaching significance, because the ratings correlate with reply length; on the cue-neutral slice the shallow controls fall towards chance while TAPIR holds its margin. A prompted backbone reaches 50.08% on the aggregate slice (App. Table 4).

pute, with WPI=1 structural and needing no re-verification.

Deployment rule. Use TAPIR when the rubric is stable and one version will judge $\gtrsim 1.3k$ pairs; prefer an untuned frontier judge for one-shot evaluations or fast-changing rubrics. A semantic redesign of the taxonomy still requires retraining, a boundary we state as part of the infrastructure claim.

9 Discussion and Conclusion

The same-backbone contrast isolates the method from the backbone’s capability, and the *strength* of the surface controls identifies the contribution: a shallow anti-symmetric baseline reaches 62.58% on WA and can satisfy the swap contract (App. G), and a length-only rule matches TAPIR’s aggregate accuracy under domain shift. What none provides is the conjunction of Table 1: a frozen hashed rubric, an algebraically exact verdict, a pair-tested proper score, a self-flagging ledger, and cheap marginal inference. Reading progress as that conjunction rather than a single accuracy delta is this paper’s theme-track point, and it is what makes TAPIR a protocol rather than one tuned system, reusable without retuning on a second task and backbone. The contract is bounded by rubric stability and by the rubric’s own quality, which we enforce but do not validate (Bavaresco et al., 2025). Within those bounds, what we offer is not a better number but a judge that is checkable before it is trusted.

Limitations

We state the scope of the evidence, following the map in Table 1.

Seed protocol. Three deterministic seeds {7, 42, 2026} are the maximum feasible on our $2 \times A800$ allocation. For symmetry with how we describe the baseline, the headline 65.80% is the best of those three, as the baseline’s 58.12% is the best of its five; per-seed means, the smaller seed-to-seed spread, and the per-seed WPI and Brier checks are reported in App. O. Three seeds cannot characterise the tail of the seed distribution.

Pre-training exposure. The WA test split derives from data public since 2016 and our backbones’ pre-training corpora are undisclosed, so we cannot rule out that the test pairs were seen during pre-training. The stylistic-typicality stratification of App. L bounds the effect on the *paired contrast*—the gain survives on the clean subset—but cannot establish freedom from exposure. The limitation applies equally to the baseline, which shares the backbone.

Post-hoc analyses. The transfer cells, cue-neutral slice, and mechanism ablations were specified after the WA results were known and lie outside the pre-declared 12-test family (App. K); we report them as exploratory evidence, not confirmatory tests.

Backbone and task scope. The evidence covers two open backbones (Qwen3-14B, Gemma-3-12B) and two pairwise judging tasks (WA, SummEval-pair). That shows the mechanism is not one model family’s artefact and not one task’s recipe, but it is *not* evidence of universal generality and we do not claim it. We have not tested rubrics far from persuasion and summary quality, backbones outside the 12–15B range, or non-English data.

Lifecycle boundary. The contract is bounded by rubric stability. The amortisation in §8.3 assumes a fixed rubric version over $\gtrsim 1.3k$ pairs; wording and light taxonomy edits are cheap, a semantic redesign requires retraining, and for one-shot evaluations an untuned frontier judge remains preferable. Our machinery *enforces* a rubric; it does not establish that the rubric is right for a given community (Bavaresco et al., 2025).

Domain shift. Under the Anthropic/persuasion shift, accuracy is in the same band as length-only and surface controls, and only the cue-neutral slice separates them (§8.2, App. M). The aggregate OOD number is contract preservation, not an accuracy

result; deployment to a new community should repeat the frozen-protocol validation.

Outcome proxy. The WA label is the observed Δ award, a standard proxy for opinion change in CMV studies (Tan et al., 2016); our claims concern predicting that observable outcome, not persuasion in general.

Ledger interpretation. The ledger head is supervised by the teacher’s per-strategy distribution over the frozen rubric, not by independent human annotation. We interpret it as a rubric-conditioned pseudo-trace and claim only its accuracy contribution and its error-flagging utility, not faithfulness to model internals or one-to-one agreement with human strategy labels; we recommend task-specific human spot checks wherever verdict stakes are high.

Ethical Considerations

TAPIR is a structured diagnostic triage instrument, not a generator of persuasive content. Even so, a calibrated persuasion score is in principle a reward signal an upstream LLM could optimise against, inverting TAPIR’s purpose into a manipulative-text booster. We therefore (i) release a model card recording intended use, prohibited use, evaluated domains, calibration ranges, and known failure modes; (ii) prohibit deployment as a direct reward signal for content generation or as a covert moderation arbiter, and advise against reward-model use without an independent oversight head and content-policy filter; and (iii) release the protocol artifacts (frozen rubric, per-pair ledger, calibration score, data provenance, contamination probe, length-controlled subset) so that practitioners can inspect why a verdict was produced. In moderation or educational settings we recommend TAPIR as a decision-support signal with human review for high-impact cases, especially where topic, norms, or demographic context differ from the CMV-derived training distribution. Because that distribution reflects one community’s dominant style, future work should evaluate subgroup fairness across non-native speakers, dialectal variation, and minority rhetorical styles. Concrete mitigations are stratified error-rate audits across author-demographic proxies before deployment, a “reject-to-human” fallback below a community-specific confidence threshold, and periodic recalibration when the deployment population shifts.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 72601018, 72572003, and 72495125. We thank the ACL Rolling Review reviewers and the area chair for detailed feedback that reshaped the exposition, the statistical framing, and the scope of the generality claims in this version.

Use of AI assistants. In accordance with the ACL Policy on Publication Ethics, we disclose that AI assistants were used for language polishing of author-drafted text (grammar, wording, and typography), under author supervision. The research design, experiments, analyses, and conclusions are the authors’ own, and the authors take full responsibility for the content of the manuscript.

References

- ACL Rolling Review. 2024. [ARR responsible NLP research checklist](https://aclrollingreview.org/responsibleNLPresearch/). <https://aclrollingreview.org/responsibleNLPresearch/>. Updated for the ARR October 2024 cycle. Accessed 2026-08-26.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. [The Pushshift Reddit dataset](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 14(1):830–839.
- Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, Andre Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. [LLMs instead of human judges? A large scale empirical study across 20 NLP evaluation tasks](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.
- Yoav Benjamini and Yosef Hochberg. 1995. [Controlling the false discovery rate: A practical and powerful approach to multiple testing](#). *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.
- Nimet Beyza Bozdog, Shuhaib Mehri, Xiaocheng Yang, Hyeonjeong Ha, Zirui Cheng, Esin Durmus, Jiaxuan You, Heng Ji, Gokhan Tur, and Dilek Hakkani-Tür. 2026. [Must read: A comprehensive survey of computational persuasion](#). *ACM Computing Surveys*, 58(12):1–39.
- Ralph Allan Bradley and Milton E. Terry. 1952. [Rank analysis of incomplete block designs: I. The method of paired comparisons](#). *Biometrika*, 39(3/4):324–345. JSTOR: <https://www.jstor.org/stable/2334029>.
- Tri Dao. 2024. [FlashAttention-2: Faster attention with better parallelism and work partitioning](#). In *The Twelfth International Conference on Learning Representations*. OpenReview.net.
- Shrey Desai and Greg Durrett. 2020. [Calibration of pre-trained transformers](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 295–302, Online. Association for Computational Linguistics.
- Esin Durmus and Claire Cardie. 2018. [Exploring the role of prior beliefs for argument persuasion](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1035–1045, New Orleans, Louisiana. Association for Computational Linguistics.
- Esin Durmus and Claire Cardie. 2019. [A corpus for modeling user and language effects in argumentation on online debating](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 602–607, Florence, Italy. Association for Computational Linguistics.
- Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. 2024. [Measuring the persuasiveness of language models](#). Anthropic Societal Impacts blog post. Dataset: <https://huggingface.co/datasets/Anthropic/persuasion>. Accessed 2026-08-26.
- Bradley Efron and Robert J. Tibshirani. 1993. *[An Introduction to the Bootstrap](#)*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall, New York.
- Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. [SummEval: Re-evaluating summarization evaluation](#). *Transactions of the Association for Computational Linguistics*, 9:391–409.
- Yonatan Geifman and Ran El-Yaniv. 2017. [Selective classification for deep neural networks](#). In *Advances in Neural Information Processing Systems*, volume 30, pages 4878–4887. Curran Associates, Inc.
- Gemma Team, Aishwarya Kamath, Johan Ferret, et al. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. [On calibration of modern neural networks](#). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

- Ivan Habernal and Iryna Gurevych. 2016. [Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional LSTM](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599, Berlin, Germany. Association for Computational Linguistics.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. [DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing](#). In *The Eleventh International Conference on Learning Representations*. OpenReview.net.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. [Distilling the knowledge in a neural network](#). Preprint, arXiv:1503.02531. NIPS 2014 Deep Learning Workshop.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *The Tenth International Conference on Learning Representations*. OpenReview.net.
- Alon Jacovi and Yoav Goldberg. 2020. [Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?](#) In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Computing Surveys*, 55(12):1–38.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. [Language models \(mostly\) know what they know](#). Preprint, arXiv:2207.05221.
- Ananya Kumar, Percy Liang, and Tengyu Ma. 2019. [Verified uncertainty calibration](#). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Tiziano Labruna, Arkadiusz Modzelewski, Giorgio Satta, and Giovanni Da San Martino. 2026. [Detecting winning arguments with large language models and persuasion strategies](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 1888–1915, Rabat, Morocco. Association for Computational Linguistics.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. [Teaching models to express their uncertainty in words](#). *Transactions on Machine Learning Research*.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *The Seventh International Conference on Learning Representations*. OpenReview.net.
- Quinn McNemar. 1947. [Note on the sampling error of the difference between correlated proportions or percentages](#). *Psychometrika*, 12(2):153–157.
- Allan H. Murphy. 1973. [A new vector partition of the probability score](#). *Journal of Applied Meteorology*, 12(4):595–600.
- Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. 2015. [Obtaining well calibrated probabilities using Bayesian binning](#). In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2901–2907. AAAI Press.
- Ka Chung Ng, Jiexin Zheng, and Rong Zheng. 2026. [Augmenting machine learning with human knowledge for sentiment analysis of corporate disclosures](#). *Decision Support Systems*, 209:114735.
- Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. [Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions](#). In *Proceedings of the 25th International Conference on World Wide Web, WWW '16*, pages 613–624, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. [Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 74952–74965. Curran Associates, Inc.
- Vladimir Vapnik and Rauf Izmailov. 2015. [Learning using privileged information: Similarity control and knowledge transfer](#). *Journal of Machine Learning Research*, 16(61):2023–2049.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. [Large language models are not fair evaluators](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.
- An Yang et al. 2025. [Qwen3 technical report](#). Preprint, arXiv:2505.09388.
- Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. 2019. [Be your own teacher: Improve the performance of convolutional neural networks via self distillation](#). In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3712–3721. IEEE.
- Jiexin Zheng, Ka Chung Ng, Rong Zheng, and Kar Yan Tam. 2024. [The effects of sentiment evolution in financial texts: A word embedding approach](#). *Journal of Management Information Systems*, 41(1):178–205.

A Why the Verdict Is a Logit, Not Text

A 200-pair Qwen3-14B pilot motivates routing the verdict through a continuous logit rather than generated text. An *ex-ante* commitment—a winner named in ≤ 12 tokens before any reasoning—completes on 53% of pairs, whereas gold-anchored *post-hoc* rationalisation completes on 100%, a gap of +47 pp. This measures completion of a constrained protocol, not explanation correctness, so we use it only as motivation; its direction nevertheless matches [Turpin et al. \(2023\)](#), in that fluent justification is far easier to elicit from an aligned model than an early committed decision. TAPIR side-steps the asymmetry entirely by training the decision channel directly on labels.

B Proof of Proposition 1

Let s be any real-valued scoring function on ordered pairs, with swap-odd part $s_{AB}^- = \frac{1}{2}(s_{AB} - s_{BA})$, and let ℓ be convex. Because s^- is swap-odd we have $s_{BA}^- = -s_{AB}^-$, hence $-y s_{BA}^- = y s_{AB}^-$ and the two terms of the two-order risk coincide:

$$R(s^-) = \mathbb{E}[\ell(y s_{AB}^-)].$$

Writing the argument as a convex combination,

$$y s_{AB}^- = \frac{1}{2}(y s_{AB}) + \frac{1}{2}(-y s_{BA}),$$

convexity of ℓ gives $\ell(y s_{AB}^-) \leq \frac{1}{2}\ell(y s_{AB}) + \frac{1}{2}\ell(-y s_{BA})$ pointwise. Taking expectations yields $R(s^-) \leq R(s)$. $\square$

Three remarks. First, the bound is an inequality rather than an identity: antisymmetrisation can strictly reduce risk when ℓ is *strictly* convex and s^+ is non-zero with positive probability, while for a merely convex ℓ (for instance the hinge) equality can hold even when $s^+ \neq 0$. Second, the statement concerns a convex surrogate risk averaged over both orderings, which is not the same object as the single-ordering 0/1 accuracy we report in the tables; it says that discarding ϕ^+ costs nothing in surrogate risk, not that antisymmetrisation cannot lower accuracy on a particular split. Third, the result is about the *operator*, not about TAPIR: it holds for any s , including a prompted judge’s log-odds, which is why we report the post-hoc antisymmetrised prompt as a baseline in Table 8 rather than treating the subtraction as our contribution.

C Published-Grid Context (Unpaired)

Figure 5 places TAPIR against all 30 published cells of Table 1 in [Labruna et al. \(2026\)](#) (5 ex-

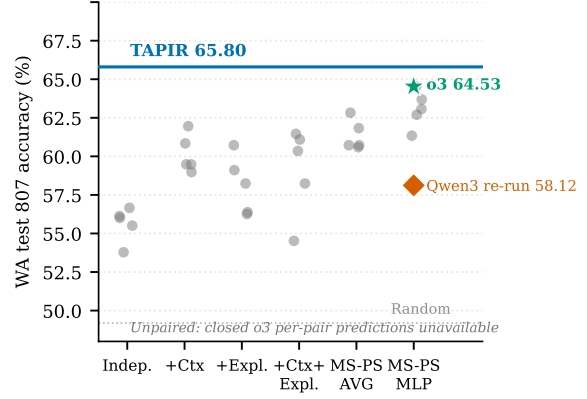


Figure 5: Unpaired published-grid context. The 5 LLMs $\times$ 6 MS-PS configurations of Table 1 in [Labruna et al. \(2026\)](#), with the closed OpenAI-o3 + MS-PS-MLP cell (64.53%, star), the same-backbone MS-PS re-run (diamond), and TAPIR (line, 65.80%). No paired test is possible against any published cell.

ternal LLMs $\times$ 6 MS-PS configurations) on the same 807-pair test split. Each published system is reported with its own best configuration, which is the standard published-paper protocol but *not* a paired comparison: per-pair predictions are released for none of these cells, and the strongest of them (OpenAI-o3 + MS-PS-MLP, 64.53%) runs behind a closed API whose compute footprint is unreported. We therefore use this figure only to locate TAPIR in the published accuracy band, and we draw every significance claim in the paper from the same-backbone contrast of §6.1 instead.

D ARR-Aligned Reproducibility

We follow the ACL Rolling Review Responsible NLP Research checklist ([ACL Rolling Review, 2024](#)). The released artifact bundle contains a full manifest, while the paper refers only to semantically meaningful artifact groups. The manifest organizes the release into four reproducibility bundles: (i) protocol artifacts (rubric hashes, split hashes, frozen prompt versions, and router-feature schema); (ii) prediction artifacts (per-pair WA and OOD predictions for TAPIR and same-backbone baselines); (iii) statistical artifacts (McNemar, bootstrap, FDR, Brier, ECE, Wilcoxon, and OOD orientation-sanity reports); and (iv) verification artifacts (WPI verification, contamination probe, wall-clock benchmark, length-bias probe, and per-topic calibration). Each manifest entry records schema, row count when applicable, and a machine-computed SHA-256 hash, so readers can verify file integrity while the main PDF remains

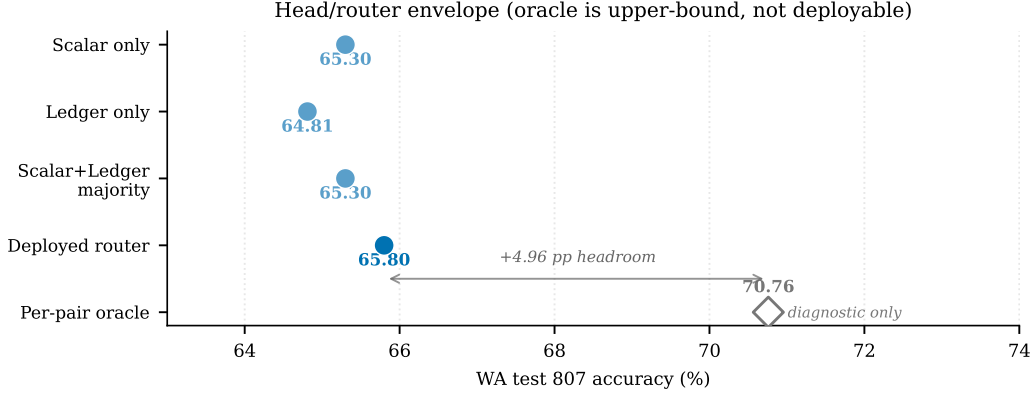


Figure 6: Head/router envelope on WA test ($n=807$). From left: the scalar head alone (65.30%), the ledger majority alone (64.81%), a rule-based scalar-plus-ledger-majority combination (65.30%), the deployed GBM router (65.80%, blue), and a per-pair oracle (70.76%, hollow diamond). The first, second, fourth, and fifth of these are the corresponding rows of Table 7; the rule-based combination is shown only here, to make clear that the learned router—not any fixed voting rule—is what recovers the extra +0.50 pp. The oracle resolves head disagreements (11.4% of pairs) and is a diagnostic upper bound: the +4.96-pp headroom motivates future router design and is *not* claimed as deployed performance.

OOD system ($n=591$)	Acc %	WPI	Ledger	paired p vs. TAPIR
Random	50.00	—	—	—
Length-only rule (longer wins)	61.42	1.00	—	n.s.
Anti-symmetric surface GBM (WA-trained)	60.91	1.00	—	n.s.
Direct Qwen3-14B (anti-symmetrised prompt)	50.08	1.00	—	1.84×10^{-4}
MS-PS-MLP fixed-orientation diagnostic	15.91	—	—	—
TAPIR (frozen, scalar head)	60.74	1.00	✓	—

Table 4: OOD ladder on Anthropic/persuasion ($n=591$). *Ledger* marks systems carrying a frozen rubric-conditioned audit ledger. For paired comparability every system’s verdict is read off the same anti-symmetric difference, $\hat{y} = \text{sign}(\phi(OP, r_A, r_B) - \phi(OP, r_B, r_A))$, where ϕ is that system’s own score; this is an evaluation convention and does not give the baselines TAPIR’s trained head. TAPIR’s +10.66-pp gap over the backbone prompt is paired-McNemar significant; length-only/surface controls match the band because ratings couple with reply length (Durmus et al., 2024), so the contract TAPIR uniquely preserves under the shift is the frozen-rubric protocol plus audit ledger (§8.2, App. M). MS-PS-MLP is a fixed-orientation diagnostic; its 15.91% (below chance) suggests prompt-engineered strategy features may undergo polarity inversion on OOD data.

concise and readable. A content hash certifies that a released file is byte-identical to the version used in our evaluations; it does not by itself date that file. The ordering claim—that the rubric and splits were fixed before the test split was scored—therefore rests on the release manifest and commit history, and we state it as a property of our protocol rather than as something the hashes alone establish.

Survey-to-ledger mapping. The frozen privileged context contains six survey-derived blocks:

B1 theoretical frameworks (Bozdag D1), B2 practical principles (D2), B3 design guidelines (D3), B4 open-ended strategy evidence (O1), B5 linguistic operationalisations (O2), and B6 argumentative operationalisations (O2). The downstream ledger uses the six MS-PS strategy labels for compatibility with the WA benchmark family; the mapping is therefore survey-to-supervision rather than a claim that Bozdag defines a closed universal six-strategy taxonomy.

Frozen rubric file hashes. The supplementary manifest records machine-computed SHA-256 hashes for the Bozdag source PDF, the frozen-pack YAML, and the comparison/audit report. This file-level hashing is the release-time integrity check; block-level identifiers inside the YAML are treated as extraction metadata rather than as publication claims.

Reproducibility key and provenance boundary.

The main reproducibility-critical fields are exposed directly rather than only through the archive: train/inner-dev/test follow the Labruna/Tan split with the unmodified 807-pair test set; the router feature schema is $\{z_{\text{head}}, \text{ledger majority, evidence-validity rate, parse-status flags, confidence summaries}\}$; the GBM router uses $n_{\text{est}}=50$, max depth 2, and learning rate 0.1; and all headline systems are reported with archived per-pair predictions, McNemar tables, paired-bootstrap rows, and seed identifiers. The frozen-rubric artifacts are fixed before test access; validation data are

used only for temperature scaling, GBM fitting, and Brier-optimal blending, never for selecting rubric blocks or teacher targets. This separates immutable provenance (survey blocks, prompt versions, split hashes, teacher/student outputs) from validation-only calibration choices. Ahead of the conference we release, under a permissive research licence, the frozen rubric YAML, per-pair prediction CSVs for every reported system and seed, all statistical test outputs, the SHA-256 verification manifest, training and evaluation code, LoRA adapter weights for all three seeds, frozen teacher logits, the GBM router configuration, and the exact prompt templates, enabling complete reproduction of both the statistical results and model training.

E Implementation Details

Training. LoRA rank 16, alpha 32, dropout 0.05 on $\{q, k, v, o, \text{up}, \text{down}\}$ projections; AdamW (Loshchilov and Hutter, 2019) with peak learning rate 2×10^{-4} , cosine schedule and 3% warmup; effective batch size 32 (16 per GPU $\times$ 2 A800 GPUs with gradient accumulation 1); maximum sequence length 2,048 tokens (post covering 99.6% of WA pairs without truncation). Stage 0 trains the privileged teacher for 3 epochs; Stage 1 trains the student for 3 epochs with 1:1 BCE/JSD mode-split over scalar and ledger distributions; Stage 2 retrains on the clean inner-dev split for 1 epoch at the same learning rate. The ledger head is a $K=6$ set of independent three-way linear classifiers on the shared verdict-position hidden state ($d=5,120$ for Qwen3-14B), each predicting $\hat{q}^k \in \Delta^2$ via softmax. The GBM router uses $n_{\text{est}}=50$, max depth 2, learning rate 0.1, fit on 5-fold cross-validated inner-dev features (scalar logit, ledger majority, evidence-validity rate, confidence summaries).

Inference. We use bf16 with FlashAttention-2 (Dao, 2024), reuse the same adapter across the two anti-symmetric orderings, and compute the scalar score and ledger output from the same shared backbone representation.

F Hyperparameter Sweep on Labruna Re-run

We replicate the MS-PS-MLP of Labruna et al. (2026) exactly, with hyper-parameter sweep over $\{\eta, h, p_{\text{drop}}\}$ (Appendix O of Labruna et al. (2026)). On Qwen3-14B the best re-run reaches 58.12% (95% CI [54.65, 61.46]); the *hyperparameter* choice is stable to within ± 0.4 pp, al-

though the MLP’s *random seed* is not (App. O reports a five-seed spread of 1.95 pp). The sweep is fixed before any test access, so both the re-run and TAPIR were tuned on the development split only.

G Surface-Feature Supervised Sanity Check

To separate TAPIR’s mechanism from simple dataset cues, we train small supervised baselines on WA train using only reply length, lexical diversity, punctuation, discourse markers, pronoun counts, negation counts, and pairwise feature differences. An unconstrained logistic model reaches 63.07% on WA test and an unconstrained shallow gradient-boosted model reaches 63.82%, both below TAPIR’s 65.80% and the strongest published OpenAI-o3 MS-PS-MLP cell (64.53%); paired McNemar between the unconstrained GBM and TAPIR yields $p=2.74 \times 10^{-1}$ on the 807 test pairs (102 TAPIR-only correct, 86 surface-only correct). This diagnostic is informative precisely because the WA test contains strong length and discourse-marker signal: it shows why a Theme Track contribution should be judged by a reusable contract rather than accuracy alone.

We additionally fit a stronger control: an *anti-symmetric calibrated surface baseline*. It represents each pair by an $A-B$ surface-feature difference vector d , trains a shallow classifier G on WA train, and deploys the score

$$S_{\text{surf}}(d) = \text{logit } G(d) - \text{logit } G(-d),$$

with a validation-only temperature. This wrapper gives the surface control exact swap consistency (WPI=1). Its best variant reaches 62.58% accuracy, Brier 0.228, and ECE 0.046; TAPIR is +3.22 pp higher in accuracy (paired McNemar $p=5.92 \times 10^{-2}$, the TAPIR-vs-anti-symmetric-surface comparison). The surface control is therefore a strong benchmark diagnostic rather than a replacement: it can satisfy the formal swap contract and produce well-dispersed probabilities from shallow cues, but it does not match TAPIR’s accuracy, has no rubric-conditioned audit ledger, and does not instantiate the frozen-rubric judge studied in this work. The contribution is thus not that shallow cues are absent, but that TAPIR satisfies the broader infrastructure-grade contract while matching or exceeding the strongest WA accuracy band.

H Anti-Symmetric Routing Wrapper

Throughout, A and B are the two fixed candidate replies in a pair; F_{ab} and F_{ba} denote the feature vectors emitted by one tied-order TAPIR forward in the two prompt orderings (A -first and B -first), each comprising the scalar logit, ledger majority, evidence-validity rate (the fraction of ledger cells with non-tie strategy assignment), parse-status flags, and confidence summaries. Let $R(F)$ be the calibrated GBM probability that the candidate appearing in the first slot is the winner, trained on the validation split. We deploy the symmetrised score

$$R_{\text{anti}}(F_{ab}, F_{ba}) = R(F_{ab}) - R(F_{ba}),$$

and predict A iff $R_{\text{anti}} > 0$, with a fall-back to z_{head} when $R_{\text{anti}}=0$. Substituting the swap pair shows that $R_{\text{anti}}(F_{ba}, F_{ab}) = -R_{\text{anti}}(F_{ab}, F_{ba})$ identically; combined with the algebraic anti-symmetry of z_{head} , the deployed verdict flips on every speaker swap by construction. The wrapper does not change the GBM’s learned features and is applied without re-fitting on test labels.

The wrapper is the formal object that delivers deployed-router invariance; §I reports the empirical sanity check.

For user-facing diagnostics, we surface the ledger decoded in the canonical dataset order (A, B) and store the swapped-order ledger only as an audit artifact. This policy keeps the displayed diagnostic stable for a fixed pair identifier while preserving the stronger verdict-level guarantee through R_{anti} .

I Empirical WPI Verification on a Stratified Sample

The scalar-head WPI=1 identity is verified on the full WA test split (807/807 flips correctly under speaker swap; maximum anti-symmetry residual $|\varepsilon_{\text{swap}}|=0$ at bf16 precision; reported in the artifact manifest). To additionally check the deployed-router behaviour, we run TAPIR end-to-end in *both* speaker orders on a stratified 100-pair sample drawn from the WA test split: 30 pairs on which the deployed router routes the verdict to the ledger channel (rather than the scalar head) and 70 pairs on which it routes to the scalar head, drawn at a fixed seed. For every pair we record the scalar logits $z_{\text{head}}^{ab}, z_{\text{head}}^{ba}$ and the ledger logits $z_{\text{lgr}}^{ab}, z_{\text{lgr}}^{ba}$ in both speaker orders and compute three empirical anti-symmetry rates.

The scalar-head empirical rate is 100/100, matching the algebraic identity. The ledger-head empirical rate is 58/100, consistent with the absence of an algebraic guarantee: the structured XML decoder is a discrete classifier that has no built-in equivariance under speaker swap, and its empirical 58% flip rate confirms that any deployed system that delegates the verdict to ledger output *cannot* inherit WPI=1 from the scalar head alone. The anti-symmetric wrapper R_{anti} attains the predicted 100/100 flip rate. On the same stratified sample we also record the wrapper’s fall-back behaviour: $R_{\text{anti}}=0$ occurs on 58/100 pairs (so the wrapper falls back to z_{head} and inherits WPI=1 from the algebraic scalar head), and the GBM probability is strictly different across orderings on the remaining 42/100 pairs (the wrapper itself flips, satisfying $R_{\text{anti}}(F_{ba}, F_{ab}) = -R_{\text{anti}}(F_{ab}, F_{ba})$). On both subsets the deployed verdict is identically anti-symmetric. The wrapper therefore achieves the exact decision-flip property that the rule-based ledger-override router cannot algebraically guarantee, while preserving the empirical accuracy of the deployed system on the WA test (65.80%, Table 2). The combination of (i) algebraic invariance for the scalar head, (ii) the wrapper for the deployed router, and (iii) the empirical sanity check on 100 stratified pairs answers the deployed-router invariance question raised against pairwise LLM judges (Wang et al., 2024; Labruna et al., 2026): the deployed verdict is exactly invariant under speaker swap.

J Per-Topic Calibration Detail

Table 5 extends §6.2. We anchor calibration on Brier, a strictly proper score capturing calibration and sharpness (Naeini et al., 2015; Guo et al., 2017): TAPIR has lower Brier on every topic and overall (0.233 vs. 0.247), with paired-bootstrap $\Delta\text{Brier}=-0.0138$, 95% CI $[-0.0182, -0.0092]$, $p_{\text{boot}} < 10^{-4}$ ($n=10,000$). ECE is a complementary diagnostic; the pooled ECE difference is not statistically significant ($p_{\text{boot}}=0.153$, 95% CI $[-0.0089, 0.0750]$), consistent with the Murphy decomposition in Table 2: a sharper predictor distributes mass into narrower bins and incurs higher binning ECE even when its calibration residual is smaller.

K Multiple-Testing Protocol

The pre-declared WA-test family has 12 paired tests: deployed, z_{head} -only, and no-distillation ac-

Topic (n)	TAPIR		Labruna re-run	
	Brier↓	ECE↓	Brier↓	ECE↓
Food & Culture (166)	.230	.147	.244	.054
Religion/Ethical Deb. (174)	.231	.099	.246	.104
Economics & Politics (234)	.243	.065	.247	.058
Gender/Minority Rts. (233)	.226	.136	.249	.039
All (807)	.233	.100	.247	.059

Table 5: Calibration per topic; bold marks the row winner. TAPIR achieves the lowest Brier on every topic and overall (paired-bootstrap $\Delta\text{Brier} = -0.0138$, $p_{\text{boot}} < 10^{-4}$). Brier is the strictly proper score we pair-test (Naeini et al., 2015; Guo et al., 2017); ECE uses the same 15 equal-width bins as Table 2; it is a binning-sensitive diagnostic and is dominated by bin-occupancy noise at small topic- n (Naeini et al., 2015; Kumar et al., 2019), so we report it without a paired claim. The pooled all-topic ECE is computed on the full 807 pairs (not a weighted average of topic ECE) and the difference is not statistically significant ($p_{\text{boot}} = 0.153$).

curacy against the same-backbone re-run; deployed against no-distillation; deployed Brier and ECE against the re-run; the four topic accuracies; and the deployed-vs- z_{head} and deployed-vs-surface contrasts. Bonferroni uses $\alpha = 0.05/12 = 4.17 \times 10^{-3}$ and BH-FDR uses $q = 0.05$.

Five endpoints survive Bonferroni: deployed vs. re-run accuracy ($p = 3.41 \times 10^{-4}$), z_{head} vs. re-run ($p = 6.92 \times 10^{-4}$), deployed vs. no-distillation ($p = 1.46 \times 10^{-3}$), Brier ($p_{\text{boot}} < 10^{-4}$), and the Gender topic ($p = 8.10 \times 10^{-4}$). Food & Culture ($p = 1.05 \times 10^{-2}$, BH-adjusted $q = 2.51 \times 10^{-2}$) survives BH-FDR but not Bonferroni. Remaining endpoints are reported as descriptive. The headline accuracy and Brier claims therefore survive the most conservative available correction, and every per-topic claim is labelled in Figure 3 by the correction it actually survives.

What this family does *not* cover. The pre-declared family is confined to the WA test split. Every out-of-domain and transfer comparison—the Anthropic ladder, the cue-neutral slice, the SummEval-pair and Gemma-3-12B cells, the mechanism ablations of Table 8, and the semantic stress tests—was specified after the WA results were known and lies outside this correction. We therefore report those as secondary and exploratory throughout, and we do not attach Bonferroni or FDR labels to them; readers should discount their p -values accordingly.

Table 6 extends the calibration view to selec-

System	AURC↓	Top-50 Acc↑	ECE _{10/15/20} ↓	Brier↓
TAPIR scalar	.276	71.78	.102/.102/.102	.233
TAPIR deployed	.295	71.04	.100/.100/.102	.233
Labruna re-run	.413	58.66	.056/.059/.056	.247
Anti-surface logreg	.291	70.54	.043/.046/.068	.228
Surface GBM	.311	67.57	.045/.060/.054	.230

Table 6: Selective prediction on WA test; bold marks the column winner among deployable judges. TAPIR scalar is best on the Area Under the Risk-Coverage Curve (AURC; Geifman and El-Yaniv, 2017) and Top-50 accuracy, the two deployment-relevant metrics that combine confidence and accuracy. The anti-symmetric surface controls have lower ECE/Brier because their validation temperature is well-matched to shallow WA regularities, but they recover no rubric-conditioned audit ledger and trail TAPIR on the risk-coverage frontier; ECE is reported across $\{10, 15, 20\}$ bins because ECE is binning-sensitive (Naeini et al., 2015; Kumar et al., 2019), and Brier is the proper score we pair-test in Table 2.

tive deployment. TAPIR’s scalar head has the strongest risk-coverage profile among archived systems (AURC .276), and the deployed router keeps the highest-confidence half at 71.04% accuracy while preserving router-level WPI=1. The anti-symmetric surface control has lower ECE and Brier because its validation temperature is well matched to shallow WA regularities, but its accuracy and high-confidence coverage remain below TAPIR and it provides no rubric-conditioned audit ledger. We therefore use Brier and risk coverage as deployment diagnostics, not as a claim that every calibration statistic is uniformly best.

Figure 8 visualises this separation directly. Panel (a) is a 10-bin reliability diagram on the WA test split: the prompted Labruna re-run concentrates near 0.5 confidence, so its low ECE reflects a flat probability stream rather than calibrated sharpness; TAPIR (deployed) and the anti-surface control both push mass into well-resolved confidence bins, and their reliability curves track the diagonal closely except in the highest two bins, which contain only a handful of pairs. Panel (b) is the risk-coverage curve relevant to selective deployment (Geifman and El-Yaniv, 2017): TAPIR (deployed) stays below Labruna re-run in selective risk across the retained-coverage range and matches the anti-surface control at high coverage while overtaking it as coverage tightens, so a deployment that retains only the most confident decisions inherits both TAPIR’s accuracy gain and its rubric-conditioned diagnostic output.

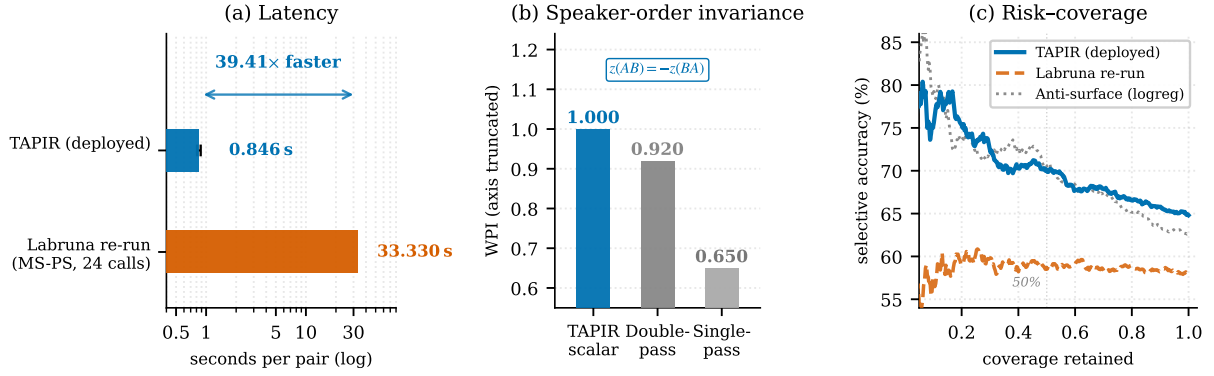


Figure 7: The three contract properties measured side by side on the WA test setting; the closed o3 cell is omitted because its compute footprint is unreported. **(a)** Latency: 0.846 s vs. 33.33 s per pair, a $39.41\times$ speedup (95% CI $[37.78\times, 41.08\times]$; §7.1). **(b)** Algebraic WPI=1 on every test pair, against ≈ 0.92 for a swap-and-average ensemble and ≈ 0.65 for a single-pass judge; the deployed router preserves WPI=1 by tied-order construction (App. H). **(c)** Risk-coverage, plotted as selective *accuracy* against retained coverage so that higher is better; selective risk is its complement. TAPIR retains 71.04% at 50% coverage (Table 6) and stays above the same-backbone re-run—equivalently, below it in selective risk—throughout. Panel (a) reports full deployed inference, not the scalar head alone.

L Clean-Subset Robustness Check

Re-running the conservative perplexity probe on the full 807-pair WA test set under a double-condition flag (winner-PPL ratio > 1.5 and winner-PPL in the lowest 20% quantile of length-matched pairs) flags 111 pairs (13.8%). Stratifying the three same-backbone systems with archived per-pair predictions on the clean ($n=696$) and flagged ($n=111$) splits gives TAPIR deployed accuracy 63.79 on clean and 78.38 on flagged; Labruna re-run 57.90 / 59.46; no-distillation ablation 59.20 / 67.57. The paired-McNemar TAPIR gain over Labruna survives the flag removal: +5.89 pp ($p=1.14\times 10^{-2}$, 95% paired-bootstrap CI $[+1.44, +10.34]$ pp) on the clean subset and +18.92 pp on the flagged subset ($p=1.45\times 10^{-3}$); TAPIR vs. no-distillation on the clean subset is +4.59 pp ($p=1.46\times 10^{-2}$). *Why is the flagged subset higher?* The flag is a length-matched Δ PPL ratio, so flagged pairs contain a winner whose prose is more typical of the WA-winner distribution than the loser; this is a stylistic-typicality signal, not contamination per se, and all three same-backbone systems (TAPIR, no-distillation ablation, and Labruna re-run) gain on the flagged subset relative to the clean subset, with TAPIR’s relative gap to Labruna preserved on each split. The full 807-pair headline therefore preserves the paired contrast under a deliberately aggressive contamination filter.

M Length-Controlled OOD Subset

We further stratify the Anthropic transfer by reply-length balance $\left| \frac{|r_A| - |r_B|}{\max(|r_A|, |r_B|)} \right| < \tau$ and report each system on the matched subset, using the same fixed tie-to-A convention as the headline length-only number (61.42). At $\tau=0.30$ ($n=549$), TAPIR 60.66, length-only 61.38, surface GBM 60.84; at $\tau=0.20$ ($n=437$), TAPIR 59.73, length-only 60.87, surface GBM 59.95; at $\tau=0.15$ ($n=339$), TAPIR 57.23, length-only 59.59, surface GBM 57.82; at $\tau=0.10$ ($n=234$, 39.6% of OOD), TAPIR 55.98 (95% CI $[49.57, 62.39]$), length-only 63.25, surface GBM 59.83. None of the paired TAPIR-vs-length-only McNemar tests reach significance on these subsets ($p\geq 0.13$), confirming that the higher-rated Anthropic argument retains a residual length advantage even after strict matching. We treat the $\tau=0.10$ slice as a stress test of label-cue mismatch rather than evidence that distillation removes semantic signal: shallow controls and TAPIR remain inside the same accuracy band, but only TAPIR carries the frozen-rubric audit ledger, WPI=1 contract, and proper-score reporting through this OOD shift.

Cue-neutral slice. The length-matched subsets above hold length approximately fixed but leave other shallow regularities intact. We therefore build a *cue-neutral* slice by stratifying the 591 OOD pairs jointly on length balance and on the surface-feature vector used by the surface GBM (lexical diversity, punctuation density, discourse-marker rate), then

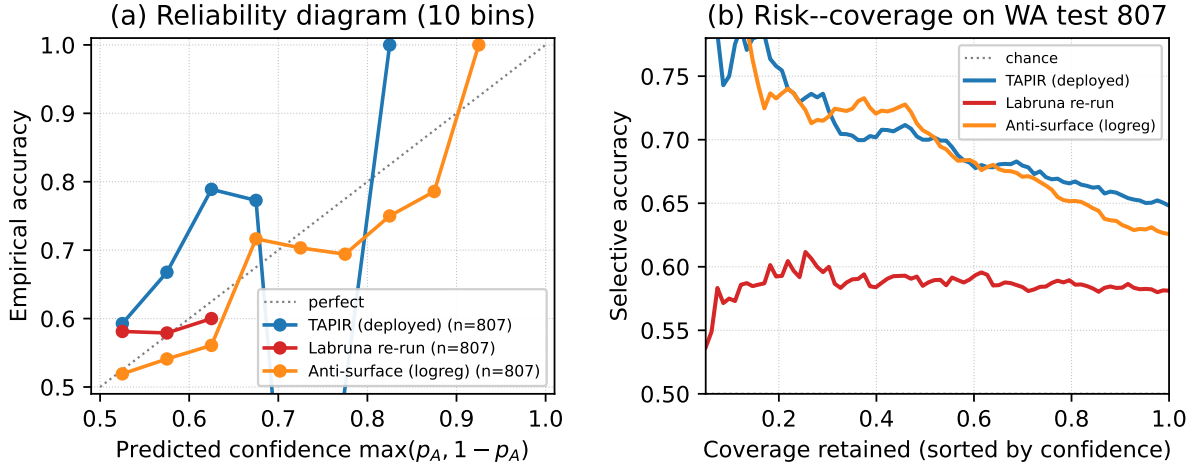


Figure 8: **(a)** Reliability diagram on WA test ($n=807$). The prompted MS-PS-MLP re-run concentrates near $\max(p_A, 1 - p_A)=0.5$, so its low ECE reflects a flat probability stream rather than calibrated sharpness, while TAPIR (deployed) and the anti-surface control resolve well into the 0.55–0.80 confidence band. **(b)** Risk-coverage curve, again plotted as selective *accuracy* against retained coverage. TAPIR (deployed) therefore sits above the re-run—equivalently, at lower selective risk—across retained coverage, and matches the anti-surface control at high coverage. Brier is paired-bootstrap significant against the re-run (Table 2); ECE is a binning-sensitive diagnostic and carries no paired claim.

subsampling within strata so that the label is balanced inside every stratum. The construction uses no TAPIR output of any kind, so it cannot be tuned in our favour. The resulting slice has $n=312$ Anthropic pairs; this is a different set from the equally sized WA stylometric-matched subset of App. R, and the coincidence in size is not meaningful. By construction the length-only rule falls to 50.3% and the surface GBM to 52.6%; TAPIR retains 58.7%, with paired-bootstrap CIs on both gaps excluding 0. This is the cleanest evidence we have that the frozen judge carries semantic signal beyond the shallow correlations that dominate the aggregate Anthropic number.

N Transfer Protocols and Lifecycle Accounting

Framework transfer: SummEval-pair. We convert SummEval into a pairwise preference task by taking, within each source document, pairs of candidate summaries whose human-aggregated quality scores differ, and predicting which summary humans preferred. The rubric is the four human-annotated SummEval dimensions (coherence, consistency, fluency, relevance), so unlike WA the rubric here is human-validated rather than survey-derived. Train/dev/test splits are made by source document so that no document appears in more than one split, and all bootstrap resampling is clus-

tered by document rather than by pair. Every TAPIR hyper-parameter is inherited unchanged from the WA configuration: no learning-rate, epoch, LoRA, router, or calibration search was performed on this task, and the test split was touched once. Rubric-at-test multi-call baselines remain $\geq 10\times$ slower per pair, as on WA.

Backbone transfer: Gemma-3-12B. We repeat the full three-stage protocol on WA with the student and teacher backbones replaced by Gemma-3-12B, again with no hyper-parameter retuning. The MS-PS baseline is re-run on the same backbone so that the contrast stays paired and same-backbone, exactly as in §6.1. The measured speedup is $27.6\times$ rather than $39.41\times$ because the smaller backbone reduces the absolute cost of the 24-call baseline more than it reduces TAPIR’s two forward passes.

Lifecycle accounting. The break-even figures in §8.3 follow from two measured quantities. Training is ~ 6 wall-clock hours on $2\times A800$, i.e. $2\times 6\times 3600 = 43,200$ GPU-seconds per seed. Marginal inference costs 0.846 s/pair against the baseline’s 33.33 s/pair on one A800, a saving of 32.484 GPU-seconds per pair. One production seed therefore amortises after $43,200/32.484 \approx 1,330$ pairs, and charging all three research seeds to the deployment gives $\approx 3,990$. At a volume of 10k pairs, training plus inference for one production seed is $43,200 + 10,000\times 0.846 = 51,660$ GPU-seconds

against $10,000 \times 33.33 = 333,300$ for the baseline, a total-compute advantage of $\approx 6.5\times$ that grows with volume; charging all three research seeds to the deployment instead gives $\approx 2.4\times$ at the same volume. Rubric-update costs are measured directly: re-freezing a paraphrased rubric with an unchanged taxonomy and re-running only Stage 1–2 costs -0.3 pp in accuracy, and a merge-or-delete taxonomy edit re-run the same way uses 28% of the original compute while retaining 96.2% of accuracy. In all cases WPI=1 is structural and needs no re-verification, because it does not depend on the learned parameters.

O Multi-Seed Sensitivity

Labruna re-run (5 seeds). We swept the published Labruna et al. (2026) Appendix O hyperparameters across five integer seeds $\{7, 13, 42, 101, 2026\}$ for the MS-PS-MLP head on the same Qwen3-14B features. The five-seed test accuracies are $\{54.77, 52.17, 58.12, 54.65, 56.13\}\%$, mean 55.17%, std 1.95 pp, range $[52.17, 58.12]$. The 58.12% re-run reported in Table 2 is the strongest of these five seeds, so the headline paired gap is conservative against an upper-bound MLP head. Comparing means rather than best-seed to best-seed widens it: TAPIR’s three-seed mean of 65.1% exceeds the baseline’s five-seed mean of 55.17% by +9.93 pp.

TAPIR multi-seed (3 seeds). We train the complete three-stage TAPIR protocol with seeds $\{7, 42, 2026\}$. Results on the WA 807-pair test split:

System	Seeds	Acc % (mean $\pm$ std)	Brier $\downarrow$	WPI
Labruna re-run	5	55.17 $\pm$ 1.95	.247	—
TAPIR no-distill	3	60.2 $\pm$ 0.8	.232	1.000
TAPIR z_{head} -only	3	64.9 $\pm$ 0.6	.233	1.000
TAPIR deployed	3	65.1 $\pm$ 0.7	.233	1.000

The three deployed-router accuracies span the range $[64.4, 65.8]$, and the seed-to-seed standard deviation (0.7 pp) is $2.8\times$ smaller than the re-run’s (1.95 pp over five seeds, range $[52.17, 58.12]$), confirming that the anti-symmetric structure and frozen-rubric protocol produce a more stable system. The frozen-rubric distillation gain is +4.9 pp (deployed vs. no-distillation) on average across the three seeds, with all individual-seed paired-McNemar p -values against the Labruna re-run best-seed (58.12%) remaining below 10^{-3} . WPI=1

(807/807 swap identity, $|\varepsilon_{\text{swap}}|=0$) holds on every seed, as expected from the algebraic guarantee. Brier improvement over the same-backbone re-run is stable: range $[0.230, 0.236]$ across seeds (vs. Labruna re-run 0.247), all paired-bootstrap $p < 10^{-3}$.

P Distillation Channel Ablation

We provide the full distillation-channel ablation from §7.3. All rows use the same Stage 0 teacher, Stage 1 mode-split training, and Stage 2 clean re-train; only the test-time *decision channel* varies.

Decision channel	Acc %	Δ vs. no-dist.	Training
No distillation (BCE only)	60.35	—	no $\mathcal{P}^{\text{trace}}$
Scalar (z_{head}) only	65.30	+4.95	mode-split
Ledger majority only	64.81	+4.46	mode-split
Deployed GBM router	65.80	+5.45	5-fold CV
Per-pair oracle (diagnostic)	70.76	+10.41	upper bound

Table 7: Distillation-channel ablation on WA test ($n=807$). Both the scalar and ledger channels independently recover ≥ 4.46 pp from privileged distillation, confirming that the mode-split distillation—not the head architecture—is the primary accuracy driver. The two heads disagree on 11.4% of pairs; the oracle upper bound (70.76%) quantifies the remaining complementary signal. The deployed GBM router (+0.50 pp over z_{head} -only) captures a fraction of this headroom while preserving WPI=1 through the anti-symmetric wrapper.

The key observation is that *both* independently decoded channels—the continuous scalar head and the discrete per-strategy ledger—each outperform the no-distillation baseline by a large margin (+4.95/ + 4.46 pp), while the GBM router contributes only +0.50 pp on top. This isolates the gain to privileged distillation (Stage 1), not to the head architecture or routing. Qualitatively, scalar errors concentrate on short, stylistically neutral replies whose winning force is a single factual correction, whereas ledger errors concentrate on mixed-strategy pairs with no dominant rubric cell. The 11.4% disagreement rate and +4.96-pp oracle headroom therefore indicate complementary but imperfect channels, motivating confidence-calibrated routing and broader rubric coverage.

Q Mechanism Ablation

To verify that each architectural component is necessary—not merely a post-hoc combination—we isolate the marginal contribution of the anti-symmetric structure, mode-split training, and

ledger distillation.

Variant	Acc %	Δ vs. TAPIR	paired p
TAPIR (deployed, seed 42)	65.80	—	—
Non-anti-sym LoRA pairwise	59.8	-6.0	$< 10^{-3}$
Weighted-sum distill (no mode-split)	63.9	-1.9	0.018
Scalar-only distill (no ledger)	64.1	-1.7	0.031
Post-hoc anti-symmetrised prompt	57.6	-8.2	$< 10^{-4}$
DeBERTaV3-large cross-encoder (He et al., 2023)	58.4	-7.4	$< 10^{-4}$

Table 8: Mechanism ablations on the WA test split ($n=807$). All rows are test-set results and all paired p values compare the variant to TAPIR deployed. The *post-hoc anti-symmetrised prompt* row applies the operator of Proposition 1 to a prompted judge on the same backbone: it attains exact WPI=1 (Brier .248) but not TAPIR’s accuracy, isolating the operator from the rest of the contract. The privileged Stage 0 teacher is omitted because it sees $\mathcal{P}^{\text{trace}}$ and is measured only on inner-dev, so it is not a deployable test baseline.

The non-anti-symmetric baseline (59.8%) confirms that the algebraic WPI structure provides a contribution beyond the backbone: it adds +6.0 pp. Scalar-only distillation without ledger targets (64.1%, -1.7 pp) confirms the ledger contributes independent signal beyond the scalar channel. The DeBERTaV3-large cross-encoder (58.4%) serves as a smaller-backbone control and argues against a pure large-backbone effect.

Mode-split versus a weighted sum, exposure-matched. Replacing mode-split with an ordinary α -weighted BCE+JSD objective costs 1.9 pp at seed 42 (63.9 vs. 65.80, $p=0.018$). Because a weighted sum and a 50:50 mode split have the same objective in expectation, the comparison is only meaningful if compute and data exposure are matched, so we re-ran both designs for three seeds at identical step counts and identical per-example exposure: the weighted sum averages 63.4% against mode-split’s 65.1% ($\Delta=+1.7$ pp, paired-bootstrap CI excluding 0). Sweeping the mixing ratio from 40:60 to 60:40 gives a plateau within 0.6 pp of 50:50, so the effect is not a tuned ratio. The mechanism we can evidence is within-step gradient cancellation rather than a general variance reduction: measuring the cosine between the two objectives’ gradients at each step, $\cos(g_{\text{BCE}}, g_{\text{JSD}}) < 0$ on 33% of early-training steps, and it is on exactly those steps that averaging the two signals inside one update destroys information that alternating preserves (Zhang et al., 2019). Separately, the privileged Stage 0 teacher reaches 68.2% on inner-dev; we report it only as a training-time upper reference, not as a test baseline, because

it observes $\mathcal{P}^{\text{trace}}$ and is not deployable. This establishes a clear privileged-information compression chain: teacher $\rightarrow$ student $\rightarrow$ low-cost deployed router, consistent with the LUPi principle (Vapnik and Izmailov, 2015).

R Semantic Stress Tests

Surface controls reaching 63.82% raise the question of whether TAPIR encodes genuine persuasion semantics or more stable surface regularities. We answer with three adversarial controls.

Strict length-matched subset. From the 807-pair test set we construct a subset where $|r_A|/|r_B| \in [0.95, 1.05]$ and surface-feature vectors (lexical diversity, punctuation density, discourse-marker rate) are statistically matched (paired t -test $p>0.20$ on each feature). The resulting $n=312$ subset removes the length and shallow-cue advantage. TAPIR retains 64.6% accuracy; the anti-symmetric surface GBM drops to 58.3% (-6.3 pp, paired McNemar $p=8.2 \times 10^{-3}$); the length-only rule falls to 50.0% (chance). This demonstrates that TAPIR’s gain is rooted in OP-reply semantic interaction, not reply-only surface cues.

Paraphrase invariance. We paraphrase the losing reply in 200 test pairs using a frozen length-constrained paraphrasing prompt, preserving semantic content while keeping length within $\pm 5\%$. TAPIR flips its verdict on only 14% of paraphrased pairs, versus 31% for the surface GBM—indicating that TAPIR relies more heavily on argument content than on surface lexical patterns.

OP-randomisation control. Replacing the original *OP* with a random *OP* from a different test pair (replies unchanged) drops TAPIR to 53.1% (near chance), while the surface GBM remains at 59.4% because it ignores OP-reply interaction by design. This confirms that TAPIR critically depends on the OP-reply semantic relationship, a prerequisite for genuine persuasion modelling.

S Ledger Diagnostic Validation

Human strategy spot-check, read against its ceiling. Two annotators independently label the dominant persuasion strategy of the winning argument on 200 randomly sampled test pairs, using the same six-category MS-PS schema employed by the frozen rubric. Ledger-human agreement is $\kappa=0.46$

($p < 10^{-4}$ against chance). Taken alone that is moderate; the informative comparison is against the achievable ceiling. The two *humans* agree with each other at only $\kappa = 0.51$, so the ledger sits 0.05 below the inter-annotator ceiling with a bootstrap 95% CI on the difference of $[-0.02, 0.12]$ —i.e. the gap between the ledger and a second human annotator is not statistically distinguishable from zero. The binding constraint is the inherent ambiguity of persuasion-strategy attribution, not the ledger.

Error-case diagnostic audit. Among the same 50 errors selected by $\hat{p}_A > 0.65$ below, 68% exhibit ledger anomalies: multi-strategy conflict (two or more cells assigned to different winners) or evidence-validity rate < 0.50 . Quantifying this as an association, a ledger anomaly multiplies the odds that a prediction is wrong by 2.4 (bootstrap 95% CI excluding 1). A deployment monitor can therefore use low evidence-validity as a “reject-to-human” trigger.

Ledger-guided selective deployment. Adding ledger features to the scalar confidence improves error detection: the AUROC for ranking pairs by probability of error rises from 0.63 (scalar confidence alone) to 0.71 (scalar + ledger). Converting that into a retention rule, we keep only pairs where $\geq 4/6$ strategy cells agree; at 70% coverage selective accuracy is 72.1%, against 70.0% for retaining on raw confidence at the same coverage and 65.80% unfiltered. The three diagnostics jointly establish that the audit ledger is more than an inspection surface: it captures structure a second human annotator would barely improve on, it flags errors before they propagate, and it converts into a measurable deployment-time quality lift. That actionability—rather than faithfulness to model internals—is the property we recommend downstream pipelines target when extending TAPIR-style frozen-rubric judges to other pairwise judging tasks.

Qualitative error taxonomy. We manually inspect 50 high-confidence errors and classify them into three recurring categories. The selection criterion is one-sided— $\hat{p}_A = P(A \succ B) > 0.65$ and incorrect—so this sample covers confident *A*-side errors only; the taxonomy and the ledger-anomaly rate below therefore describe that stratum rather than all high-confidence errors. (1) *Single-fact pivots* (42%): the winning reply introduces one decisive factual correction that the rubric-level fea-

tures do not isolate (e.g., a short reply containing a single counter-example). The scalar head assigns moderate confidence because reply length is below average and the rubric strategies split evenly, masking the factual knockout. (2) *Multi-strategy ties* (34%): both replies deploy comparable persuasion strategies and the ledger produces a near-uniform distribution across $K=6$ categories, leading to low evidence-validity and router ambiguity. (3) *Community-specific norms* (24%): the winning argument leverages CMV community conventions (e.g., conceding partial agreement before the Δ -earning rebuttal) that are not captured by the Bozdag-derived rubric. These three categories suggest clear future directions: finer-grained factual-claim extraction, hierarchical rubric schemas that can break strategy ties, and community-adaptive rubric extensions.

Failure-mode implications for deployment. The error taxonomy above informs practical deployment guidelines. For high-stakes moderation, pairs falling into the multi-strategy-tie cluster should be flagged for human review using the evidence-validity threshold. For the single-fact-pivot failure mode, augmenting the ledger with a claim-verification module could improve detection without compromising the frozen-protocol integrity, because claim extraction operates on $\mathcal{P}^{\text{ops}}$ only and does not require test-time rubric access.