DialPO: Dialectical Prompt Optimisation

Gokulakrishnan B[*]
gokulbalaji2408@gmail.com

Vijay Kumar Neeluru
Amazon, India
zvkumarn@amazon.com

Vinayak Puranik
Amazon, India
puranikv@amazon.com

Abstract

Existing textual-gradient prompt optimisers rely on a single Large Language Model (LLM) to diagnose failures, propose edits, and judge the results, conflating critique and refinement into one model's judgement. We introduce **DialPO** (Dialectical Prompt Optimisation), a multi-agent framework that replaces the single-model feedback loop with a structured debate: an Advocate defends what the prompt does well, a Critic attacks failures, cross-examination filters their arguments adversarially, a data-isolated Synthesis model drafts the revised prompt from surviving arguments alone, and a regression gate rejects edits that degrade batch accuracy. We provide a theoretical framework identifying three preconditions under which multi-agent debate improves the feedback-reasoning bottleneck of textual-gradient methods, and show that DialPO instantiates them by construction. Across 12 benchmarks spanning classification, reasoning, summarisation, and question answering, DialPO achieves the highest scores across various Target models, winning or tying a majority of datasets in head-to-head comparisons against each baseline while using 8–18× fewer LLM calls than beam-search optimisers. Component ablations across five Target models reveal that DialPO's advantage scales inversely with Target-model strength, clarifying when dialectical debate helps prompt optimisation.

1 Introduction

Prompt engineering has become the dominant interface for adapting large language models to downstream tasks, but small variations in wording can swing task performance by over 45% (Sclar et al., 2024). Automated prompt optimisers address this sensitivity by generating and scoring candidate prompts via LLM-in-the-loop search. Methods such as ProTeGi (Pryzant et al., 2023), OPRO (Yang et al., 2024), and GEPA (Agrawal et al., 2026) consistently improve over manual prompts, but share a common architectural choice: they use a single LLM to diagnose failures and propose edits, collapsing critique and refinement into a single model's judgement.

This design has two consequences. First, *unverified feedback*: a hallucinated diagnosis ("the prompt fails because it lacks context" when the real issue is noisy labelling) propagates unchallenged into the next candidate. Second, *destructive editing*: failure-driven rewrites have no mechanism to preserve components that are already working, so a fix for sentiment handling may silently remove a clause that was helping with sarcasm detection. Both are structural, not algorithmic: no amount of hyperparameter tuning inside a single-agent loop addresses them.

We present **DialPO** (Dialectical Prompt Optimisation), a multi-agent framework that replaces the single-model feedback loop with a structured debate between three LLMs with non-overlapping responsibilities. An *Advocate* defends what the current prompt is doing well; a *Critic* attacks failures and proposes targeted modifications; *cross-examination* filters their arguments adversarially; a *Synthesis* model produces a revised prompt from the surviving arguments alone, without access to raw training data; and a *regression gate* rejects edits that degrade accuracy on the sampled batch. This design instantiates three preconditions under which multi-agent debate is known to improve reasoning (Du et al., 2024; Irving et al., 2018; Wang et al., 2024): viewpoint diversity, structured challenge, and bounded-rational adjudication.

Contributions. This paper makes four contributions.

- **DialPO framework.** A multi-agent prompt optimiser that, to our knowledge, is the first to apply dialectical debate directly to prompt refinement

[*] Work done during an internship at Amazon.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 42922–42938
October 24-29, 2026 ©2026 Association for Computational Linguistics

rather than to inference-time reasoning (Abdali et al., 2025).

- **Theoretical grounding.** We articulate three preconditions $\mathcal{A}$ under which debate is known to improve reasoning, and show DialPO's components are designed to instantiate them (§3).
- **Empirical evaluation.** Across 12 benchmarks spanning classification, reasoning, summarisation, and question answering, DialPO achieves the highest scores on both open-weight and closed-weight Target models. DialPO wins or ties in head-to-head comparisons against each baseline while using $8-18\times$ fewer LLM calls than beam-search optimisers.
- **Component ablations across five Target models.** We find that the marginal value of DialPO's safety components scales inversely with Target-model strength: Full DialPO dominates on weaker Targets and is matched or exceeded by simpler ablations on stronger ones. This Target-dependent pattern suggests a capability-aware adoption rule: use the full debate pipeline for weaker Targets and simpler variants as Target capability increases.

2 Related Work

Iterative prompt optimisation. Automatic prompt optimisation has progressed from generation-based paradigms such as APE (Zhou et al., 2023) and RLPrompt (Deng et al., 2022) to iterative feedback-driven methods. ProTeGi (Pryzant et al., 2023) introduced the textual-gradient framework, where a single LLM produces natural-language critiques of the current prompt that guide subsequent edits under beam search. OPRO (Yang et al., 2024) maintains a scored optimisation trajectory in the LLM's context window. SCULPT (Kumar et al., 2025) decomposes feedback into structured dimensions and treats prompts as hierarchical trees. GEPA (Agrawal et al., 2026) reframes the problem as Pareto-evolutionary search over multiple objectives. Self-Refine (Madaan et al., 2023) and Reflexion (Shinn et al., 2023) generalise the view of verbal feedback as a learning signal. All of these methods share a single-agent architecture for the critique and edit steps, which is precisely the bottleneck DialPO addresses.

Multi-agent debate. Structured debate between LLM agents has been shown to improve factuality and reasoning (Du et al., 2024; Liang et al., 2024), help weaker judges evaluate stronger arguments (Khan et al., 2024), and admit formal guarantees under bounded-rational adjudication (Irving et al., 2018). Gains are not universal: compute-matched single-agent baselines can match debate (Smit et al., 2024), and self-critique loops can degrade reasoning (Huang et al., 2024). Wang et al. (2024) identify genuine viewpoint diversity as the critical precondition. Prior work applies debate to inference-time reasoning or evaluation (Chan et al., 2024); DialPO is, to our knowledge, the first to apply it to prompt optimisation itself.

3 Theoretical Motivation

Existing textual-gradient prompt optimisers, ProTeGi (Pryzant et al., 2023), OPRO (Yang et al., 2024), GEPA (Agrawal et al., 2026), and SCULPT (Kumar et al., 2025), share two architectural limitations: **unverified feedback** (a single LLM's failure diagnosis is trusted even when hallucinated) and **destructive editing** (failure-driven rewrites have no mechanism to preserve already-working components). DialPO addresses both with a structured multi-agent debate, motivated in three steps.

Reasoning is the bottleneck. Textual-gradient optimisers cast prompt refinement as a reasoning task: the LLM diagnoses why a prompt fails, and that diagnosis determines the edit. ProTeGi, OPRO, GEPA, and SCULPT each report ablations where the critique component dominates the search strategy; Self-Refine (Madaan et al., 2023) and Reflexion (Shinn et al., 2023) report the same when LLMs refine their own outputs. We therefore adopt the premise that per-iteration prompt improvement increases with the probability that feedback correctly identifies the failure mode, its *feedback reasoning quality*.

Debate helps reasoning under conditions $\mathcal{A}$. Multi-agent debate improves LLM reasoning when three conditions hold: agents argue toward opposing objectives (*viewpoint asymmetry*), a protocol forces each claim to survive rebuttal (*structured challenge*), and a bounded-rational adjudicator settles disputes. Under these conditions, debate beats single agents on factual and mathematical reasoning (Du et al., 2024), mitigates single-agent degeneration of thought (Liang et al., 2024), helps weaker judges verify stronger arguments (Khan

et al., 2024), and admits a formal guarantee (Irving et al., 2018). When these conditions are absent, compute-matched single agents match debate (Smit et al., 2024) and self-critique can degrade reasoning (Huang et al., 2024); Wang et al. (2024) identify viewpoint diversity as the critical precondition. We denote the three collectively as conditions $\mathcal{A}$.

DialPO instantiates $\mathcal{A}$. Advocate–Critic role asymmetry (§4.1) enforces *viewpoint asymmetry*; cross-examination enforces *structured challenge*; data-isolated Synthesis plus the Target-model regression gate enforce *bounded-rational adjudication*. Under $\mathcal{A}$, debate raises feedback reasoning quality above a single agent, predicting higher per-iteration improvement for DialPO than for single-agent baselines. We state this as a motivating conjecture, not a theorem: the reasoning-to-prompt-quality link lacks closed-form derivation, and $\mathcal{A}$ can fail when a Target's single-pass reasoning already saturates. Our ablations (§7) confirm the predicted pattern: full debate dominates on weaker Targets and is matched by simpler variants on stronger ones.

4 Methodology

We present DialPO (Dialectical Prompt Optimisation), a multi-agent framework that replaces the single-model feedback loop of prior textual-gradient optimisers with a structured debate grounded in actual execution results. This section describes the six-stage pipeline (§4.1) and the resulting computational profile (§4.2).

4.1 Architecture and Pipeline

DialPO uses three separate LLM instances with non-overlapping responsibilities. This separation is a deliberate design choice: it prevents the confirmation bias that arises when the same model both proposes an edit and judges whether it worked.

Each iteration proceeds through six stages, illustrated in Figure 1. Grounding comes first so that everything downstream is tethered to real model behaviour; the debate happens in the middle where it is most productive; and regression checking comes last as a safety net.

Let P_r denote the prompt at iteration r, $\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{n}$ the training set, and $\mathcal{S}_r \subset \mathcal{X}$ a randomly sampled subset used for grounding. We write $\hat{y}_i = \text{Target}(P_r, x_i)$ for the Target model's prediction on example x_i.

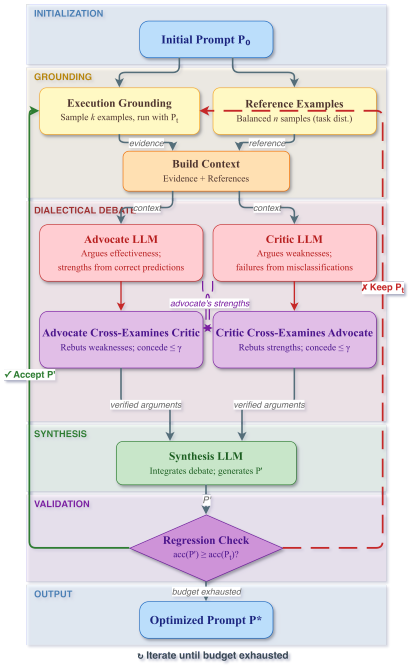


Figure 1: The DialPO pipeline. Each iteration grounds the current prompt on sampled examples, runs an Advocate–Critic debate with cross-examination, synthesises an improved candidate P', and accepts it only if it passes a regression check.

Stage 1: Execution grounding. The Target model runs P_r on $\mathcal{S}_r$ and partitions the results into correct and incorrect sets:

$$\mathcal{S}_r^+ = \{(x_i, y_i) \in \mathcal{S}_r : \hat{y}_i = y_i\}, \quad \mathcal{S}_r^- = \mathcal{S}_r \setminus \mathcal{S}_r^+. \tag{1}$$

A set of reference examples representing the broader task distribution is also supplied, so the debate has context beyond the sampled batch.

Stage 2: Advocate. The Teacher, in its Advocate role, examines $\mathcal{S}_r^+$ and builds a case for what P_r is doing well, identifying components (instructions, formatting choices, constraints) that contribute to correct predictions. The Advocate's mandate is preservation: anything working should be protected from unnecessary revision. Let A_r denote the resulting argument set.

Stage 3: Critic. The Teacher switches to the Critic role and examines $\mathcal{S}_r^-$ from the opposite direction, identifying failure clusters, edge cases, and prompt components contributing to errors. The Critic argues for targeted modifications, not wholesale rewriting. Let K_r denote the resulting argument set. Role asymmetry between Advocate and Critic instantiates the viewpoint-diversity precondition of $\mathcal{A}$.

Stage 4: Cross-examination. This is the core quality-filtering step. The Advocate challenges the Critic's proposed changes, and the Critic challenges the Advocate's preservation claims; each side either concedes or rebuts, and only claims that survive adversarial scrutiny are retained. A configurable concession ratio κ (default 0.4) requires each side to reject at least a fraction κ of the opposing side's arguments, preventing trivial mutual agreement. Formally, cross-examination produces a filtered argument set

$$\mathcal{F}_r = \mathrm{CrossExam}(A_r, K_r) \subseteq A_r \cup K_r, \quad (2)$$

with $A_r \setminus \mathcal{F}_r$ denoting the Advocate arguments rejected during the exchange (and symmetrically for K_r), subject to the per-side rejection constraints

$$\frac{|A_r \setminus \mathcal{F}_r|}{|A_r|} \geq \kappa, \quad \frac{|K_r \setminus \mathcal{F}_r|}{|K_r|} \geq \kappa. \quad (3)$$

This step addresses the unverified-feedback problem: every proposed change must withstand challenge before influencing the prompt, realising the structured-challenge condition of $\mathcal{A}$.

Stage 5: Synthesis. The Synthesis model receives $\mathcal{F}_r$, the surviving arguments from both sides, and nothing else. No raw examples, no execution logs, just the debate output. It drafts a revised prompt P'_r:

$$P'_r = \mathrm{Synthesis}(\mathcal{F}_r, P_r). \quad (4)$$

Because Synthesis cannot see training data, its edits are necessarily abstract ("add explicit handling of negation" rather than "add the phrase 'not happy' to the prompt"), which mitigates overfitting. This data isolation, combined with the regression gate in Stage 6, jointly implements the bounded-rational adjudication condition of $\mathcal{A}$: Synthesis reasons only over the filtered debate, while final acceptance is delegated to the Target model's measurable accuracy on the sampled batch.

Stage 6: Regression protection. The Target model re-evaluates P'_r on the full sampled batch $\mathcal{S}_r$ and compares the resulting accuracy against the accuracy achieved by the previous prompt P_r. Let δ be the configurable regression tolerance. The acceptance criterion is:

$$P_{r+1} = \begin{cases} P'_r & \text{if } \mathrm{Acc}(P'_r, \mathcal{S}_r) \geq \mathrm{Acc}(P_r, \mathcal{S}_r) - \delta, \\ P_r & \text{otherwise,} \end{cases} \quad (5)$$

where $\mathrm{Acc}(P, \mathcal{S})$ is the accuracy of prompt P on example set $\mathcal{S}$ as evaluated by the Target model. If the revised prompt regresses on the sampled batch beyond δ, the edit is rejected outright and P_r is retained. This is the safety net against destructive editing: regardless of how strong the debate consensus was, the numbers have the final say. A complete worked example illustrating DialPO's output at each phase, initial prompt, Advocate analysis, Critic rebuttal, Synthesis draft, and Target evaluation, is provided in Appendix I.

4.2 Computational Complexity

DialPO's per-iteration cost is additive across its six stages and independent of any beam or paraphrase hyperparameters. Let n denote the number of examples sampled for grounding per iteration ($n = |\mathcal{S}_r|$). Each grounding example triggers one Target-model call in Stage 1, each debate stage (Advocate, Critic, Cross-examination, Synthesis) contributes a constant number of calls c independent of n, and the regression gate in Stage 6 triggers n further Target-model evaluations on the same sampled batch. The per-iteration LLM-call cost is therefore:

$$C^{\text{iter}}_{\text{DialPO}} = 2n + c, \quad (6)$$

where c is the fixed debate overhead ($c = 5$ in our implementation: one call each for the Advocate, Critic, and Synthesis and two calls for cross-examination). Over R iterations, the total cost scales linearly in both n and R:

$$C^{\text{total}}_{\text{DialPO}} = R \cdot (2n + c) = \mathcal{O}(R \cdot n). \quad (7)$$

Since only n and R appear in Eq. 7, production budgeting reduces to setting two interpretable quantities.

Empirical validation. Table 2 (§6.2) reports the measured per-dataset call, wall-clock, and token budgets on mid tier closed weight LLM across the 12 benchmarks (averaged over 5 seeds). DialPO's per-dataset budget is flat (standard deviation below 5% across datasets) and predictable across tasks.

5 Experimental Setup

Datasets. We evaluate on 12 benchmarks spanning four task families: binary and multi-class classification (SST-2, TweetEval Hate, AG News, Emotion, Liar, Banking77), reasoning (BBH Logical Deduction, StrategyQA), summarisation (CNN/DailyMail, SAMSum, XSum), and question

answering (HotpotQA). Classification tasks use macro-F1; generative tasks use ROUGE-L. Each test set is capped at 200 examples (seed 42), held constant across all algorithms and iterations. On non-deterministic Target models we report mean and standard deviation across 5 re-evaluation runs. Dataset details are in Appendix A.

Models. We refer to all Target and Teacher models by capability-based notation throughout the paper; Table 6 summarises this notation and the capability tier each label denotes.[1] All closed-weight models are accessed via commercial APIs; open-weight models are publicly available. Capability tiers (Small, Mid, High, Frontier) reflect relative performance on standard benchmarks and are ordered ascending. We evaluate DialPO on two Target models: **Target-Closed-Mid**, a closed-weight commercial model; and **Target-Open-Mid**, an open-weight model. Temperature is set to 0 for both the models. Teacher-Frontier serves as the Teacher and Synthesis model throughout (temperatures 0.7 and 0.2 respectively). For component ablations we additionally use Target-Closed-High, Target-Frontier, and Target-Open-Small as Targets, and vary the Teacher across Teacher-Open-High and Teacher-Open-Mid (Appendix F). We report mean and standard deviation across 5 re-evaluation runs.

Baselines. We compare DialPO against four strong prompt optimisers: OPRO (Yang et al., 2024), ProTeGi (Pryzant et al., 2023), GEPA (Agrawal et al., 2026), and SCULPT (Kumar et al., 2025). All algorithms start from the same task-generic initial prompt and run for 8 iterations. Per-algorithm hyperparameters are listed in Appendix B. DialPO defaults: $n = 10$ samples per iteration, concession ratio $\kappa = 0.4$, regression tolerance $\delta = 0.2$.

Evaluation protocol. Optimisation cost is tracked separately from evaluation cost. Final scores are measured on the 200-example held-out test set after all 8 iterations. For the cost–accuracy analysis (§6.2), we additionally evaluate every algorithm on the test set after each iteration, yielding 540 algorithm $\times$ dataset $\times$ checkpoint evaluations.

[1]We anonymize model identities per institutional publication policy.

6 Results

6.1 Main Results

Table 1 reports the final test score of each algorithm across the 12 benchmarks and average across task types on Target-Closed-Mid, with standard deviations over 5 re-evaluation runs, aggregated by task family. DialPO leads on reasoning (0.746), summarisation (0.242), and question answering (0.712), and is competitive on classification (0.625, within noise of the strongest baseline). The same ordering holds on Target-Open-Mid (full results in Appendix C). Head-to-head, DialPO wins $9/12$ datasets over OPRO, $8/12$ over GEPA, $9/12$ over SCULPT, and ties at $6/6$ against ProTeGi on Target-Closed-Mid; the corresponding numbers on Target-Open-Mid are $10/12$, $7/12$, $11/12$, and $7/12$ over ProTeGi. DialPO is particularly strong on tasks where the initial prompt is far from optimal: HotpotQA ($0.088 \rightarrow 0.712$, $+0.624$), TweetEval Hate ($0.291 \rightarrow 0.680$, $+0.389$), and SAMSum ($0.214 \rightarrow 0.324$, $+0.110$) show the largest absolute gains on Target-Closed-Mid.

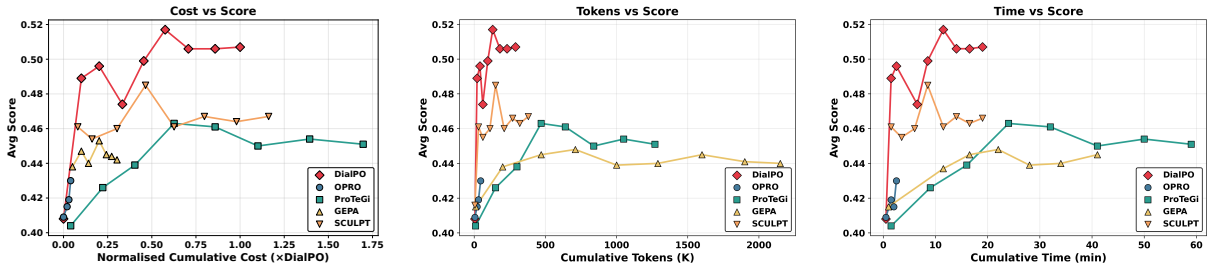
6.2 Cost–Accuracy Analysis

DialPO incurs a fixed 200 LLM calls per dataset (25 per iteration $\times$ 8 iterations), compared to ProTeGi's 1,544 calls and GEPA's 3,593 calls. OPRO uses fewer calls (98) but achieves a substantially lower final F1 (Table 2). We report overall optimization cost normalised to DialPO ($= 1.00$): OPRO is the cheapest ($0.03\times$), while ProTeGi ($1.64\times$) and SCULPT ($1.15\times$) cost more than DialPO for lower final F1. 96% of DialPO's calls are Teacher-side (Teacher-Frontier) calls; the 160 Target-Closed-Mid Target calls make a negligible contribution, keeping DialPO's per-dataset budget flat (standard deviation below 5%) at the price of linear Teacher-quality dependence (§6.3).

Figure 2 shows per-iteration budget–accuracy trajectories across three cumulative budget metrics: normalised cost, token count, and wall-clock time. DialPO rises steeply in the first few iterations and stabilises once the regression gate begins rejecting further edits. ProTeGi continues to spend substantial additional budget (over 60 minutes of wall-clock time) without closing the F1 gap. GEPA accumulates the highest token volume owing to its Pareto-evolutionary candidate pool, yet plateaus earlier than DialPO. OPRO uses the fewest calls at every point but converges to a lower final F1. At final budget (Appendix G), DialPO, GEPA, and

Dataset	Init.	DialPO	OPRO	ProTeGi	GEPA	SCULPT
SST-2	0.915	**0.925**±.000	0.907±.008	0.931±.004	0.913±.011	0.905±.000
TweetEval Hate	0.291	**0.680**±.005	0.301±.000	0.626±.004	0.661±.012	0.645±.002
AG News	0.858	0.850±.010	0.906±.007	0.885±.003	**0.908**±.000	0.884±.002
Emotion	0.521	0.493±.002	0.465±.018	0.515±.010	**0.517**±.002	0.516±.000
Liar	0.161	0.173±.014	0.096±.014	**0.209**±.024	0.151±.008	0.176±.003
Banking77	0.622	0.628±.006	0.615±.009	**0.633**±.006	0.611±.008	0.568±.001
Classification Avg.	0.561	0.625	0.548	**0.633**	0.627	0.616
BBH Logical Deduction	0.770	0.800±.066	0.811±.053	0.804±.052	**0.812**±.039	0.787±.038
StrategyQA	0.610	**0.692**±.016	0.612±.009	0.621±.002	0.632±.004	0.625±.008
Reasoning Avg.	0.690	**0.746**	0.712	0.713	0.722	0.706
CNN/DailyMail	0.150	**0.202**±.002	0.152±.002	0.105±.001	0.097±.001	0.164±.001
SAMSum	0.214	**0.324**±.001	0.216±.001	0.119±.001	0.198±.001	0.166±.002
XSum	0.175	0.199±.002	0.176±.001	0.110±.001	**0.218**±.002	0.176±.001
Summarisation Avg.	0.180	**0.242**	0.181	0.111	0.171	0.169
HotpotQA	0.088	**0.712**±.002	0.055±.001	0.042±.001	0.036±.001	0.037±.001
Question Answering Avg.	0.088	**0.712**	0.055	0.042	0.036	0.037

Table 1: Main results on 12 benchmarks with Target-Closed-Mid as the Target model. Values are final test scores (macro-F1 or ROUGE-L) after 8 iterations, mean ± standard deviation over 5 re-evaluation runs. **Bold** indicates best per row. Target-Open-Mid results in Appendix C.



(a) Normalised cumulative cost (×DialPO) vs. Average score (b) Cumulative tokens vs. Average score (c) Cumulative wall-clock time vs. Average score

Figure 2: Per-iteration budget–accuracy trajectories averaged across the 12 benchmarks on Target-Closed-Mid. DialPO reaches its peak score early under every budget axis and stabilises; beam-search baselines (ProTeGi, GEPA) continue consuming budget with smaller marginal gains. Cost is normalised relative to DialPO (= 1.00). All algorithms start from the same initial prompt.

Algorithm	Norm. Cost (×DialPO)	Time (min)	Calls
DialPO	1.00	27.6	200
OPRO	0.03	10.5	98
ProTeGi	1.64	69.7	1,544
GEPA	0.78	85.9	3,593
SCULPT	1.15	26.7	330

Table 2: Average optimisation budget per dataset on Target-Closed-Mid, averaged across the 12 benchmarks. Cost is normalised relative to DialPO (= 1.00). Time and LLM calls are reported as measured.

OPRO lie on the Pareto frontier under the reported budget metrics; ProTeGi and SCULPT are dominated.

6.3 Model Variation

DialPO's advantage depends on Target model strength. Across five Target models (Table 3), DialPO's improvement Δ over the initial prompt correlates negatively with Target strength ($r \approx -0.7$): DialPO dominates on Target-Open-Small ($\Delta = +0.119$, rank 1/8 among variants and baselines), Target-Closed-Mid ($\Delta = +0.185$, rank 2/8), and Target-Open-Mid ($\Delta = +0.126$, rank 2/8), and loses to ProTeGi on the strongest Target, Target-Frontier, where its Δ is slightly negative (-0.029 vs. ProTeGi's $+0.017$).

Across three Teacher models (Table 4, full ablation in Appendix F), Teacher choice substantially affects which DialPO components dominate. Teacher-Frontier yields the largest debate-value gap (the difference between Full DialPO and the merged Advocate+Critic ablation, $+0.160$), while the open-weight Teachers in our study show smaller or negligible gaps. This heterogeneity is consistent with the viewpoint-diversity precondition in

Target	Init.	DialPO Δ	ProTeGi Δ	Rank
Target-Open-Small	0.384	**+0.119**	+0.007	1/8
Target-Closed-Mid	0.412	**+0.185**	+0.041	2/8
Target-Open-Mid	0.470	**+0.126**	−0.053	2/8
Target-Closed-High	0.403	+0.094	+0.109	4/8
Target-Frontier	0.503	−0.029	+0.017	4/8

Table 3: Target-model variation. Rank is among 7 DialPO variants plus ProTeGi on 6 core datasets. DialPO's improvement decreases monotonically as Target capability increases.

Teacher	Full DialPO Δ	Debate gap
Teacher-Frontier	**+0.185**	**+0.160**
Teacher-Open-Mid	+0.148	+0.033
Teacher-Open-High	+0.118	−0.012

Table 4: Teacher-model variation on Target-Closed-Mid Target, 6 core datasets. "Debate gap" is the difference $\Delta_{\text{Full}} - \Delta_{\text{Merged A+C}}$, measuring the marginal value of role asymmetry. Only Teacher-Frontier yields a substantial positive gap.

$\mathcal{A}$ (Wang et al., 2024): structured debate benefits the most when the Teacher can produce genuinely opposing arguments across the Advocate and Critic roles.

7 Component Ablation

We evaluate seven variants of DialPO:

- **Full DialPO.** The complete pipeline with Advocate, Critic, cross-examination, data-isolated Synthesis, and regression gate.
- **No CrossExam.** Advocate and Critic arguments pass directly to Synthesis without adversarial filtering.
- **No RegGate.** Every synthesised prompt is accepted unconditionally.
- **No DataIsol.** Synthesis sees the raw training batch $\mathcal{S}_r$ alongside the filtered debate.
- **No Advocate.** Only the Critic analyses failures; no preservation case is built.
- **No Critic.** Only the Advocate analyses successes; no failure diagnosis is performed.
- **Merged A+C.** A single LLM call produces both preservation and modification arguments simultaneously.

Each variant is evaluated on 6 core datasets (TweetEval Hate, Emotion, BBH Logical Deduction, StrategyQA, XSum, HotpotQA) across five Target models (Table 5).

Full DialPO is the most consistent variant. It is the only variant with positive Δ on four of five Targets. No other variant is positive on more than three. Specific ablations beat it on individual Targets, e.g., No DataIsol on Target-Closed-Mid (+0.199 vs +0.185) and Merged A+C on Target-Open-Mid (+0.131 vs +0.126), making Full DialPO the safest default rather than the absolute best.

HotpotQA drives most of the variance. On Target-Closed-Mid, Full DialPO gains +0.624 on HotpotQA while Merged A+C gains only +0.003; on tasks where the initial prompt already works (XSum: Full DialPO +0.024), the debate structure adds little. This suggests that role-asymmetric debate is most valuable when the initial prompt is far from optimal and the Target has substantial reasoning headroom.

Adversarial separation matters. Merged A+C is the worst variant on Target-Closed-Mid (+0.025) and Target-Closed-High (+0.013), confirming that role asymmetry enforces the viewpoint-diversity precondition in $\mathcal{A}$. A single model instructed to "argue both sides" produces convergent rather than genuinely opposing arguments.

Stronger Targets prefer simpler pipelines. On Target-Frontier and Target-Closed-High, the top-ranked variants are No CrossExam and No RegGate; on Target-Open-Small, the top-ranked variant is Full DialPO. This inversion is consistent with the prediction in §3 that conditions $\mathcal{A}$ fail when single-pass reasoning saturates. The implications are discussed in §8.

7.1 Hyperparameter Sensitivity

DialPO is robust across reasonable hyperparameter ranges. Performance varies under 15% across a 4× range of samples per iteration ($n = 3$ to 10) and a 2× range of concession ratio ($\kappa = 0.3$ to 0.7). Optimal defaults are $n = 10$ samples, 20 reference examples, and regression tolerance $\delta = 0.2$. Performance degrades only at extreme values: $n = 20$ (insufficient grounding diversity) and $\kappa = 0.5$ (over-pruning of the argument set). Full sensitivity tables are in Appendix E.

8 Discussion

The empirical patterns in §6 and §7 support and refine the theoretical framing of §3 in three sub-

Variant	Target-Closed-Mid	Target-Closed-High	Target-Frontier	Target-Open-Small	Target-Open-Mid
Full DialPO ⋆	+0.185	+0.094	−0.029	**+0.119**	+0.126
No CrossExam	+0.089	+0.101	**+0.045**	+0.053	+0.070
No RegGate	+0.028	**+0.191**	−0.055	+0.006	+0.005
No DataIsol	**+0.199**	+0.071	−0.016	+0.069	+0.120
No Advocate	+0.171	+0.092	−0.080	+0.049	+0.075
No Critic	+0.128	+0.109	+0.005	+0.070	+0.022
Merged A+C	+0.025	+0.013	−0.064	+0.070	**+0.131**

Table 5: Component ablation: average Δ across 6 core datasets per Target. Bold indicates the highest Δ per Target. Full DialPO is rank 1 on Target-Open-Small and rank 2 on Target-Closed-Mid and Target-Open-Mid; on stronger Targets simpler ablations outperform.

stantive ways.

The Target-dependent pattern is the empirical signature of $\mathcal{A}$. Our framework predicts that gains shrink when the Target's single-pass reasoning saturates. The observed inverse correlation between Target strength and DialPO's Δ ($r \approx -0.7$) is exactly this signature: weaker Targets have more reasoning headroom; stronger Targets are closer to ceiling. The Teacher-variation results reinforce this reading, since Teacher-Frontier yields a $+0.160$ debate-value gap while open-weight Teachers in our study show smaller or negligible gaps, consistent with the viewpoint-diversity precondition $\mathcal{A}$ identifies as critical (Wang et al., 2024). We interpret this as evidence that prompt-optimisation gains under multi-agent debate depend on the joint capability of Target and Teacher rather than on either model in isolation.

Component value is conditional, not universal. Standard ablation methodology reports a single ranking of components by importance. Our results suggest this view is incomplete for debate-based optimisers: the ranking inverts across Target capability tiers. The regression gate, for example, contributes positively in aggregate on Target-Closed-Mid, Target-Open-Small, and Target-Open-Mid but ranks below No RegGate on Target-Closed-High, where the gate's safety margin appears to block beneficial edits more often than destructive ones. Cross-examination shows the same inversion: removing it roughly halves DialPO's gain on Target-Closed-Mid (+0.185 → +0.089) but improves performance on Target-Frontier. We read this as indicating that debate-based methods admit a richer design space than single-pass optimisers: the safety apparatus can be tuned to match the Target's failure-mode profile rather than treated as universally beneficial. This motivates a capability-aware adoption rule: deploy Full DialPO for Target models below the Target-Frontier capability tier; for stronger Targets, simpler variants such as No CrossExam or No RegGate match or exceed full debate.

Connection to multi-agent debate literature. Prior work applies multi-agent debate to inference-time reasoning (Du et al., 2024; Liang et al., 2024) and to evaluation (Chan et al., 2024), where the debate's role is to settle a single decision. DialPO repurposes the same primitives, role-asymmetric agents, structured rebuttal, and bounded-rational adjudication, to optimise a textual artefact (the prompt) that is itself reused across many decisions. The fixed-cost amortisation this enables, $R \cdot (2n+c)$ Teacher calls during optimisation to improve all subsequent Target inferences, is, to our knowledge, a novel use of debate primitives in the prompt-optimisation literature.

Limits of the conjecture framing. Our theoretical motivation explicitly stops short of a closed-form derivation: we state conditions $\mathcal{A}$ as a synthesis of prior empirical findings, not as a theorem. The reverse-direction failures observed in the ablations would be hard to predict from a strict theorem but follow naturally from the conjecture's qualifying clause about saturation. Closing this gap, formalising the relationship between single-pass reasoning capacity and debate's marginal return, is a direction we leave to future work.

9 Conclusion

We presented DialPO, a multi-agent prompt optimiser that replaces single-model feedback loops with structured dialectical debate. Across 12 benchmarks and two Target models, DialPO achieves the highest average score against four strong baselines while using $8-18\times$ fewer LLM calls than beam-search alternatives. Our theoretical framework, conditions $\mathcal{A}$, predicts that the marginal value of debate-based optimisation depends on the joint ca-

pability of Target and Teacher models, a prediction confirmed by ablations across five Target and three Teacher models. The Target-dependent component inversion we observe motivates a capability-aware adoption rule and suggests that future work on multi-agent prompt methods should explicitly characterise the model-capability regime in which gains are expected.

Limitations

Strong Target models do not benefit. Full DialPO shows negative improvement on Target-Frontier ($\Delta = -0.029$), the strongest Target in our study. The debate structure appears to over-edit prompts that are already near-optimal, a pattern consistent with saturation of single-pass reasoning quality. DialPO is best applied to Target models below this capability tier; for stronger Targets, simpler pipelines match or exceed full debate.

Teacher quality dependence. 96% of DialPO's cost is Teacher-side, and the marginal value of role-asymmetric debate varies substantially by Teacher. In our experiments, only Teacher-Frontier yielded a substantive debate-value gap ($+0.160$ between Full DialPO and merged Advocate+Critic on Target-Closed-Mid); the open-weight Teachers we tested showed smaller or negligible gaps despite comparable parameter counts. Settings constrained to cheap Teachers are not currently well supported, and we have not evaluated whether fine-tuning a smaller Teacher on debate traces could close the gap.

Task coverage. Our 12 benchmarks span classification, reasoning, summarisation, and question answering, but exclude tasks with hard output-format constraints (e.g., JSON schema validation, exact-match set selection, structured parseable output). DialPO's debate-driven rewrites may drift from strict formats when optimisation signal is dominated by semantic quality rather than format conformance; we leave systematic evaluation on format-strict benchmarks to future work.

Limited Target sample size. Our model-variation findings ($r \approx -0.7$ between Target capability and DialPO improvement) are based on five Target models. While the inverse-with-capability pattern is consistent across the five, the small sample limits the strength of generalisation claims; a larger Target panel could refine or qualify the relationship.

Theoretical guarantees. Our theoretical motivation (§3) is a motivating conjecture rather than a proof: the reasoning-to-prompt-quality link lacks closed-form justification, and conditions $\mathcal{A}$ fail for some deployments. The ablations test these predictions but do not prove them.

References

Sara Abdali and 1 others. 2025. Self-reflecting large language models: A hegelian dialectical approach. *arXiv preprint arXiv:2501.14917*.

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alex Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. 2026. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *The Fourteenth International Conference on Learning Representations*.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650. Association for Computational Linguistics.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. 2020. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 38–45, Online. Association for Computational Linguistics.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better LLM-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations*.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3369–3391, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the*

Association for Computational Linguistics, 9:346–361.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

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

Geoffrey Irving, Paul Christiano, and Dario Amodei. 2018. AI safety via debate. *arXiv preprint arXiv:1805.00899*.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive llms leads to more truthful answers. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Shanu Kumar, Akhila Yesantarao Venkata, Shubhanshu Khandelwal, Bishal Santra, Parag Agrawal, and Manish Gupta. 2025. SCULPT: Systematic tuning of long prompts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14996–15029, Vienna, Austria. Association for Computational Linguistics.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

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

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with "gradient descent" and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7957–7968, Singapore. Association for Computational Linguistics.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

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

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik R Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Andries Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. 2024. Should we be going mad? a look at multi-agent debate strategies for llms. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key?

In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6106–6131, Bangkok, Thailand. Association for Computational Linguistics.

William Yang Wang. 2017. "Liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V Le, Denny Zhou, and Xinyun Chen. 2024. Large language models as optimizers. In *The Twelfth International Conference on Learning Representations*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

A Dataset Details

Table 7 lists the 12 benchmarks used throughout the paper. We deliberately selected datasets spanning four task families, binary classification, multi-class classification, reasoning, and generative tasks (summarisation and question answering), so that performance gains cannot be attributed to a single task format. The benchmarks comprise classification datasets (SST-2 (Socher et al., 2013), AG News (Zhang et al., 2015), Emotion (Saravia et al., 2018), TweetEval Hate (Barbieri et al., 2020), Banking77 (Casanueva et al., 2020), Liar (Wang, 2017)), reasoning datasets (BBH Logical Deduction (Suzgun et al., 2023), StrategyQA (Geva et al., 2021)), summarisation datasets (CNN/DailyMail (Nallapati et al., 2016), SAMSum (Gliwa et al., 2019), XSum (Narayan et al., 2018)), and HotpotQA (Yang et al., 2018).

Where an official test split exists, we sample 200 examples from it; where only a train split is available, we hold out the last 200 examples before any optimisation begins. No training example appears in any test set, and the test subset is identical across all algorithms and re-evaluation seeds to ensure direct comparability.

B Per-Algorithm Hyperparameters

Table 8 reports the hyperparameter settings used for each algorithm in the main evaluation. All algorithms run for the same 8 iterations on the same initial prompt per dataset to isolate algorithmic differences from budget differences. DialPO's own parameters were selected via the sensitivity analysis reported in Appendix E.

C Target-Open-Mid Main Results

Table 9 reports the full per-dataset scores with standard deviations over 5 re-evaluation runs on Target-Open-Mid as the Target model, aggregated by task family. DialPO leads on summarisation (0.239) and question answering (0.637) and is competitive on classification and reasoning, winning 10/12 datasets against OPRO, 7/12 against both GEPA and ProTeGi, and 11/12 against SCULPT. The qualitative pattern matches Target-Closed-Mid (Table 1): DialPO dominates on tasks where the initial prompt is far from optimal (e.g., HotpotQA, $0.357 \rightarrow 0.637$), and baselines are more competitive on tasks where the initial prompt is already strong (e.g., SST-2, Banking77).

D Optimisation Cost Details

Table 10 reports the average optimisation budget per dataset averaged across the 12 benchmarks on Target-Closed-Mid. DialPO's call count is fixed at 200 per dataset and its budget is flat across tasks (standard deviation below 5%), properties the multiplicative-budget baselines do not share. The cost column is normalised to DialPO ($= 1.00$).

Cost breakdown for DialPO. Of the 200 calls per dataset, 40 are Teacher/Synthesis calls on Teacher-Frontier (8 Advocate + 8 Critic + 16 cross-examination + 8 Synthesis), and 160 are Target calls on Target-Closed-Mid (10 grounding + 10 regression-gate evaluations × 8 iterations). The 40 Teacher-Frontier calls account for approximately 96% of both the total token volume and the normalised cost; the 160 Target-Closed-Mid calls make a negligible contribution. Standard deviation of the total cost across the 12 datasets is below 5%, a flatness not observed in any multiplicative-budget baseline.

E Hyperparameter Sensitivity

We sweep each DialPO hyperparameter independently on 6 core datasets with Target-Closed-Mid as the Target. The remaining hyperparameters are held at their defaults ($n = 10$ samples, 20 reference examples, concession ratio $\kappa = 0.4$, regression tolerance $\delta = 0.2$).

Across all four sweeps, the best-performing setting beats the worst by under 15% of Δ, indicating that DialPO is robust to hyperparameter choice within reasonable ranges. Performance degrades only at extreme values ($n = 20$, $\kappa = 0.5$).

F Teacher-Model Variation

Table 15 reports DialPO Δ across three Teacher models and the seven DialPO variants from the component ablation (§7). Teacher-Frontier is the only Teacher where Full DialPO yields a substantively positive debate-value gap; the open-weight Teachers in our study show smaller or negligible gaps.

Debate-value gap. Defining the debate-value gap as $\Delta_{\text{Full}} - \Delta_{\text{Merged A+C}}$, Teacher-Frontier shows $+0.160$, Teacher-Open-Mid shows $+0.033$, and Teacher-Open-High shows -0.012. The size and sign of this gap indicate whether debate's role separation contributes additional signal beyond what

Model Notation	Description
Target models (optimised for and evaluated on; capability ascending)	
Target-Open-Small	A 4B-parameter open-weight model
Target-Closed-Mid	A mid-tier closed-weight commercial model
Target-Open-Mid	A 12B-parameter open-weight model
Target-Closed-High	A high-tier closed-weight commercial model
Target-Frontier	A frontier-class closed-weight commercial model
Teacher models (run the dialectical optimisation)	
Teacher-Frontier	A frontier-class closed-weight model; also serves as the Synthesis model
Teacher-Open-High	A 120B-parameter open-weight mixture-of-experts model (∼5B active parameters)
Teacher-Open-Mid	An 80B-parameter open-weight dense model

Table 6: Capability-based notation used for all Target and Teacher models throughout the paper. Target models (capability ascending) are optimised for and evaluated on; Teacher models run the dialectical optimisation.

#	Dataset	Task type	Classes	Metric	Test size
1	SST-2	Binary classification	2	macro-F1	200
2	TweetEval Hate	Binary classification	2	macro-F1	200
3	AG News	Multi-class classification	4	macro-F1	200
4	Emotion	Multi-class classification	6	macro-F1	200
5	Liar	Multi-class classification	6	macro-F1	200
6	Banking77	Multi-class classification	77	macro-F1	200
7	BBH Logical Deduction	Reasoning	5	macro-F1	200
8	StrategyQA	Reasoning	2	macro-F1	200
9	CNN/DailyMail	Summarisation	–	ROUGE-L	200
10	SAMSum	Summarisation	–	ROUGE-L	200
11	XSum	Summarisation	–	ROUGE-L	200
12	HotpotQA	QA (multi-hop)	–	ROUGE-L	200

Table 7: The 12 benchmarks spanning four task families. Each test set is fixed at 200 examples sampled with seed 42 from the official test split and held constant across all algorithms, seeds, and iterations.

a single model argues on its own. Only Teacher-Frontier provides a substantively positive gap, suggesting that the marginal value of structured debate depends on the Teacher's ability to produce genuinely opposing arguments across the Advocate and Critic roles, not merely on the Teacher's raw parameter count.

G Pareto Frontier Analysis

Figure 3 shows the Pareto frontier of final F1 score against three complementary budget metrics, averaged across the 12 benchmarks on Target-Open-Mid. Three algorithms occupy the frontier: OPRO (cheapest, $0.04\times$ DialPO cost, F1 $= 0.448$), GEPA (mid-range, $0.78\times$, F1 $= 0.506$), and DialPO (highest-quality, $1.00\times$, F1 $= 0.563$). ProTeGi ($1.92\times$, F1 $= 0.453$) and SCULPT ($1.22\times$, F1 $= 0.482$) are Pareto-dominated: both cost more than GEPA yet achieve a lower final F1 than DialPO. Cost is reported normalised to DialPO ($= 1.00$).

The plots together support the efficiency claim for DialPO from two directions. First, DialPO sits on the Pareto frontier under every reported budget metric, meaning no baseline achieves higher F1 at lower budget on cost, calls, or time. Second, DialPO's dominance gap over the next frontier algorithm (GEPA) is larger on F1 ($+0.057$) than the corresponding gap on calls (3,343 fewer; 200 vs 3,543) and time (47.7 minutes faster; 28.4 min vs 76.1 min), indicating that the additional budget DialPO uses is spent efficiently on genuine accuracy gains rather than diminishing-return search.

H Statistical Significance

To quantify the reliability of the differences in Table 1, we run paired bootstrap significance tests (10,000 resamples, $\alpha = 0.05$) between DialPO and each baseline on every dataset. Tables 16 and 17 summarise the per-baseline win/loss/tie counts across the 12 datasets, together with the number of wins and losses that are statistically significant.

On Target-Closed-Mid (Table 16), DialPO sig-

Parameter	DialPO	OPRO	ProTeGi	GEPA	SCULPT
Iterations	8	8	8	8	8
Samples / iter	10	10	10	3 (minibatch)	10
Beam / population	–	top-20 history	4	grows	2
Gradients / iter	–	–	4×4	–	–
D_{pareto}	–	–	–	300	–
Concession ratio κ	0.4	–	–	–	–
Regression tolerance δ	0.2	–	–	–	–
Teacher / optimiser LLM	Teacher-Frontier (temp 0.7)				
Synthesis LLM	Teacher-Frontier (temp 0.2)	–	–	–	–
Target LLM	Target-Closed-Mid / Target-Open-Mid (temp 0.0)				

Table 8: Hyperparameters used in the main evaluation.

Dataset	Init.	DialPO	OPRO	ProTeGi	GEPA	SCULPT
SST-2	0.932	**0.945**±.000	0.933±.006	0.931±.002	0.931±.002	0.916±.002
TweetEval Hate	0.502	**0.666**±.015	0.501±.008	0.379±.004	0.658±.000	0.661±.000
AG News	0.861	0.830±.003	0.849±.003	**0.878**±.000	0.866±.000	0.821±.003
Emotion	0.444	**0.540**±.001	0.438±.003	0.456±.004	0.509±.008	0.493±.003
Liar	0.095	0.236±.006	0.115±.021	**0.262**±.001	0.248±.000	0.138±.002
Banking77	0.649	0.635±.001	0.648±.001	**0.701**±.008	0.700±.000	0.663±.002
Classification Avg.	0.581	0.642	0.581	0.601	**0.652**	0.615
BBH Logical Deduction	0.641	0.796±.035	0.629±.021	0.803±.010	**0.851**±.009	0.697±.041
StrategyQA	0.700	0.749±.014	0.320±.000	**0.767**±.009	0.688±.000	0.704±.012
Reasoning Avg.	0.671	0.773	0.475	**0.785**	0.770	0.701
CNN/DailyMail	0.163	**0.190**±.001	0.163±.001	0.068±.001	0.137±.000	0.096±.000
SAMSum	0.244	**0.342**±.001	0.245±.001	0.088±.000	0.217±.001	0.286±.001
XSum	0.174	0.185±.001	0.175±.001	0.073±.000	**0.191**±.000	0.177±.001
Summarisation Avg.	0.194	**0.239**	0.194	0.076	0.182	0.186
HotpotQA	0.357	**0.637**±.000	0.358±.000	0.023±.000	0.075±.000	0.128±.000
Question Answering Avg.	0.357	**0.637**	0.358	0.023	0.075	0.128

Table 9: Main results on 12 benchmarks with Target-Open-Mid as the Target model, mean ± standard deviation over 5 re-evaluation runs. **Bold** indicates best per row.

nificantly outperforms all baselines on reasoning (BBH +24–58%, StrategyQA +7–12%) and generative tasks (HotpotQA +67%, SAMSum +8–17%, CNN/DailyMail +3–8%), with an overall win rate of 69% (33/48 comparisons). The main weakness is the Emotion dataset, where DialPO significantly underperforms all baselines (-15% to -20%); on the remaining classification tasks (Banking77, AG News, SST-2) it is competitive or dominant.

On Target-Open-Small (Table 17), DialPO shows strong significant gains on generative tasks (HotpotQA +16–55%, CNN/DailyMail +5–14%, and SAMSum), but overall classification performance is mixed: ProTeGi, GEPA, and SCULPT significantly outperform DialPO on StrategyQA and TweetEval Hate. The win/loss ratio is roughly balanced ($\approx 50\%$), indicating that DialPO's dialectical approach is less dominant on this weaker Target than on Target-Closed-Mid, particularly on classification tasks where simpler optimisers suffice.

I Worked Example: DialPO Iteration Trace

We illustrate one complete DialPO iteration on the HotpotQA dataset (Target-Closed-Mid Target, iteration 2). The prompt enters the iteration at 60% accuracy on the sampled batch and exits at 100% after the regression gate.

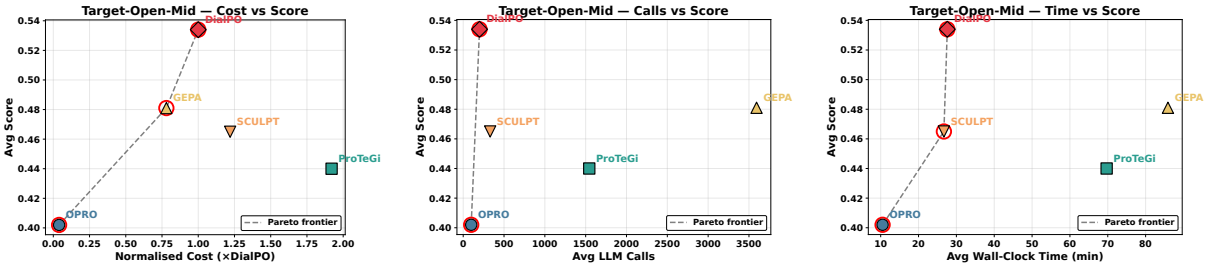
Prompt before iteration.

```
Answer the following question.
Context: {context}
Question: {input}
Answer: For yes/no questions, respond with only "yes" or "no" without additional explanation. For other questions, provide the specific information requested.
```

Stage 1: Grounding. Ten examples sampled. The Target model classifies six correctly and four incorrectly. Three of the four failures share an identical pattern: the model responds "yes" or "no"

Algorithm	Norm. Cost (×DialPO)	Avg wall-clock (min)	Avg LLM calls	Avg tokens
DialPO	1.00	27.6	**200**	270K
OPRO	0.03	10.5	98	39K
ProTeGi	1.64	69.7	1,544	1,270K
GEPA	0.78	85.9	3,593	3,457K
SCULPT	1.15	26.7	330	376K

Table 10: Per-dataset optimisation budget averaged across the 12 benchmarks on Target-Closed-Mid. Cost is normalised relative to DialPO (= 1.00). Wall-clock time, LLM calls, and tokens are reported as measured. DialPO's LLM call count is fixed at 200 per dataset.



(a) Normalised cost (×DialPO) vs. Average Score

(b) LLM calls vs. Average Score

(c) Wall-clock time (min) vs. Average Score

Figure 3: Pareto frontier analysis on Target-Open-Mid, averaged across 12 benchmarks, for three budget metrics. Each point is one algorithm; the connecting line marks the Pareto-efficient frontier. Cost is normalised relative to DialPO (= 1.00).

Samples per iteration n	Avg Δ	Wins
3	+0.162	4/6
5	**+0.191**	**6/6**
10	+0.170	4/6
20	+0.026	4/6

Table 11: Samples per iteration. Five samples dominate.

Reference examples	Avg Δ	Wins
5	+0.163	5/6
10	+0.102	3/6
20	**+0.190**	**5/6**
30	+0.095	5/6

Table 12: Number of reference examples shown to the debate.

Concession ratio κ	Avg Δ	Wins
0.3	+0.163	5/6
0.5	+0.049	6/6
0.7	+0.152	5/6

Table 13: Cross-examination concession ratio. The dip at $\kappa = 0.5$ suggests over-pruning of the argument set.

Regression tolerance δ	Avg Δ	Wins
0	+0.138	5/6
0.05	+0.146	4/6
0.1	+0.151	4/6
0.2	**+0.160**	**4/6**

Table 14: Regression tolerance δ. Higher tolerance allows the synthesised prompt more room to improve without being rejected.

to factual wh-questions that require extractive answers.

Stage 2: Advocate (key arguments). The Teacher, in its Advocate role, examines the six correct predictions and produces preservation arguments. Key points:

- **Direct answer instruction without preamble** produces clean phrase-level outputs (Examples 3, 6, 7).
- **Context-then-question structure** ensures the model processes context before answering (Example 8 synthesises GCHQ location details from multiple context elements).
- **"Provide the specific information requested"** allows adaptive response length: minimal for simple questions, elaborated for complex ones.
- **No length constraints** let response length match question complexity (2–3 words to 1–2 sentences).

Stage 3: Critic (key arguments). The Teacher in its Critic role examines the four incorrect predic-

Teacher	Full DialPO	No CrossExam	No RegGate	No DataIsol	No Advocate	No Critic	Merged A+C
Teacher-Frontier	**+0.185**	+0.089	+0.028	+0.199	+0.171	+0.128	+0.025
Teacher-Open-High	+0.118	+0.132	+0.110	+0.130	+0.106	+0.129	+0.130
Teacher-Open-Mid	+0.148	+0.108	−0.020	+0.093	+0.123	+0.177	+0.115

Table 15: Teacher-model variation on 6 core datasets (Target-Closed-Mid Target). Average Δ over the initial prompt.

Baseline	Wins	Losses	Ties	Sig. Wins	Sig. Losses
OPRO	7	4	1	6	1
ProTeGi	9	3	0	6	3
GEPA	7	5	0	5	2
SCULPT	8	3	1	6	1

Table 16: Paired bootstrap significance results (10,000 resamples, $\alpha = 0.05$) for DialPO vs. each baseline across the 12 datasets on Target-Closed-Mid.

Baseline	Wins	Losses	Ties	Sig. Wins	Sig. Losses
OPRO	6	6	0	4	4
ProTeGi	6	6	0	4	2
GEPA	5	7	0	4	4
SCULPT	6	6	0	3	4

Table 17: Paired bootstrap significance results (10,000 resamples, $\alpha = 0.05$) for DialPO vs. each baseline across the 12 datasets on Target-Open-Small.

tions:

- **Question-type misclassification** (75% of failures): the model responds "yes/no" to factual wh-questions because the yes/no instruction is too vague.
- **Under-specification of definitional context** (1 failure): for type/category questions, the model returns just the label without explanatory framing.
- **Missing question-type exemplars**: the prompt has no in-context examples distinguishing yes/no from informational questions.

Stage 4: Cross-examination (summary). Of seven Advocate arguments, four are conceded and three survive. Of four Critic arguments, two are conceded and two survive. Both sides exceed the $\kappa = 0.4$ minimum concession ratio. The Critic's argument that the yes/no instruction directly causes 75% of failures is unanimously conceded by the Advocate. The Critic's argument for adding extensive type/category guidance is conceded because it addresses only 1 of 4 failures.

Surviving arguments passed to Synthesis. *From Preservation:* context-then-question structure, "specific information requested" instruction, no length constraints. *From Modification:* question-type clarification, minimal type exemplars.

Stage 5: Synthesis. The Synthesis model receives only the surviving arguments and the current prompt and produces:

> *"Based on the surviving arguments, I'm addressing the primary failure mode (75% of errors) by disambiguating question types with concise clarification and minimal exemplars. I'm preserving the context-then-question structure, absence of length constraints, and the 'specific information requested' instruction which work in successful cases."*

Prompt after iteration.

```
Answer the following question based on
the provided context.

Context: {context}

Question: {input}

Answer: Respond based on the question
type:
- For yes/no questions (asking whether
something is true/false): respond with
only "yes" or "no" without additional
explanation
- For informational questions (asking
who/what/which/when/where/how): provide
the specific information requested from
the context

Examples of question types:
- Yes/no: "Is X true?" "Does Y happen?"
"Can Z occur?"
- Informational: "Who did X?" "What is
Y?" "Which Z was involved?"
```

Stage 6: Regression gate. The Target re-evaluates the revised prompt on all ten sampled examples. Accuracy improves from 60% to 100%; the edit is accepted. The question-type disambiguation resolves the misclassification issue: the model now correctly identifies wh-questions as informational and provides extractive answers instead of defaulting to yes/no.

Result	Input	Gold	Predicted
Incorrect	In which British comedy series did Toby Williams appear (Sky Living)?	Trying Again	yes
Incorrect	Who had more fame: Jacoby Shaddix or John Waite?	John Charles Waite	no
Correct	What is another name for the blood disorder Emmanuel Ntima Weyi supports?	Sickle-cell disease	Sickle-cell disease
Incorrect	What did the Irish special operations force have access to?	a wide variety of specialist weapons	yes
Correct	Where is the Peppers Manor House located?	Australia	Sutton Forest, NSW, Australia
Correct	What NFL team does the 2010 Washington Huskies head coach work for?	Atlanta Falcons	Atlanta Falcons
Correct	What country of origin do Patrick Brice and Creep share?	American	American
Correct	Where is GCHQ based?	Cheltenham	In the suburbs of Cheltenham.
Correct	Where did Queen Mab originate?	William Shakespeare	Shakespeare's play "Romeo and Juliet."
Incorrect	What type of soprano was Emma Matthews famous for?	lyric soprano (operatic voice type)	lyric coloratura soprano

Table 18: Grounding results: 6 correct, 4 incorrect (60% accuracy).