

BalancedBio: Mitigating the Alignment Tax in Biomedical LLMs via Gradient Orthogonality and GRPO

Wentao Wu, Hanmeng Zhong, Linqing Chen*, Peng Zhou

PatSnap Co., LTD. Suzhou, China

{wuwentao3, zhonghanmeng, chenlinqing, tonyzhoupeng}@patsnap.com

Abstract

Aligning Large Language Models (LLMs) for specialized domains presents a fundamental optimization challenge: the “alignment tax.” In biomedicine, this manifests as a conflict between the need for rigorous, encyclopedic factual accuracy and the requirement for flexible, user-friendly instruction following. Existing methods, primarily relying on Supervised Fine-Tuning (SFT) or standard Reinforcement Learning from Human Feedback (RLHF), often fail to navigate this Pareto frontier, resulting in models that are either knowledgeable but rigid, or chatty but hallucination-prone. In this paper, we propose **BalancedBio**, a novel alignment framework that explicitly models and optimizes conflicting objectives. While utilizing standard high-quality instruction data, our contribution lies in the algorithmic innovation: (1) We introduce **Capability-Aware Group Relative Policy Optimization (GRPO)**, which eliminates the need for a value network and reduces gradient variance in high-entropy reasoning tasks; (2) We propose a **Dynamic Hybrid Reward Mechanism** that adaptively balances domain correctness, reasoning validity, and format compliance during training; (3) We provide a theoretical analysis demonstrating how our method enforces gradient orthogonality to mitigate catastrophic forgetting. Extensive experiments on BIOMED-MMLU, MedQA, and IFEval show that BalancedBio-7B achieves state-of-the-art performance, surpassing Med-PaLM-8B in domain tasks while maintaining robust general instruction-following capabilities.¹

1 Introduction

The democratization of Large Language Models (LLMs) has revolutionized information access. However, the adaptation of these general-purpose models to high-stakes vertical domains, such as

healthcare, remains fraught with difficulties (Singhal et al., 2023). A central hurdle is the phenomenon known as the *alignment tax* (Askell et al., 2021; Ouyang et al., 2022), where optimizing a model for a specific narrow objective (e.g., medical factual correctness) inadvertently degrades its performance on general capabilities (e.g., conversational fluency, formatting constraints, or safety guardrails).

In the biomedical context, this trade-off is particularly acute. Medical reasoning requires strict adherence to physiological principles and established clinical guidelines (Thirunavukarasu et al., 2023). Conversely, user interaction often demands flexibility—such as summarizing a diagnosis in a JSON format or explaining a concept to a five-year-old. Standard alignment techniques like Proximal Policy Optimization (PPO) (Schulman et al., 2017) struggle to balance these heterogeneous rewards. PPO relies on a learned Value Network (Critic) to estimate the expected return. In complex reasoning tasks with long horizons, the Critic often fails to learn an accurate value landscape, leading to high-variance gradient updates and training instability (Shao et al., 2024).

Furthermore, the gradients derived from maximizing medical accuracy often conflict with those maximizing instruction compliance. For instance, a model penalized heavily for “hallucination” may learn to be overly conservative, refusing to answer harmless formatting requests, thereby failing the instruction-following objective. This creates a multi-objective optimization problem where the Pareto front is difficult to identify using scalar reward functions.

To address these challenges, we present **BalancedBio**, a rigorous alignment framework designed to harmonize domain expertise with general instruction compliance. Unlike prior works that focus on data curation (Li et al., 2023; Wu et al., 2023), we focus on the *optimization dynam-*

*corresponding author

¹<https://github.com/patsnap/Hiro-Pharma>

ics. We hypothesize that the interference between capabilities can be minimized by enforcing a form of gradient orthogonality through group-based reinforcement learning.

Our contributions are as follows:

1. We formally define the biomedical alignment problem as a multi-objective Markov Decision Process (MDP) and identify the gradient conflict problem.
2. We propose **Capability-Aware GRPO**, a memory-efficient RL algorithm that leverages group statistics to estimate baselines, significantly reducing variance compared to PPO in reasoning tasks.
3. We introduce a **Dynamic Weight Adaptation** strategy that treats the alignment process as a closed-loop control system, dynamically adjusting reward coefficients to prevent any single objective from dominating the optimization landscape.
4. We conduct a comprehensive error analysis, revealing that our method specifically reduces "reasoning hallucinations" without compromising format adherence.

2 Related Work

2.1 Biomedical LLMs

The landscape of biomedical NLP has rapidly evolved from discriminative models like PubMedBERT (Gu et al., 2021) to generative foundation models. While BioGPT (Luo et al., 2022) demonstrated the potential of domain-specific pre-training, recent advancements have focused on multimodal capabilities and clinical reasoning. **Med-Gemini** (Saab et al., 2024) and **Med-PaLM 2** (Singhal et al., 2023) have set new benchmarks by integrating long-context understanding and multimodal processing. In the open-source domain, models like **Meditron** (Chen et al., 2023) and **BioMistral** (Labrak et al., 2024) have narrowed the gap with proprietary models via efficient adaptation techniques. However, relying primarily on Supervised Fine-Tuning (SFT) remains a bottleneck. As noted by Pal et al. (2023), SFT excels at knowledge injection but lacks the negative feedback loop required to correct hallucinations in complex diagnostic chains, often leading to models that memorize textbook patterns rather than engaging in robust clinical reasoning.

2.2 Reinforcement Learning for Alignment

RLHF remains the gold standard for aligning LLMs with human intent (Christiano et al., 2017). While PPO (Schulman et al., 2017) is widely used, it suffers from high memory costs and instability. To address this, Direct Preference Optimization (DPO) (Rafailov et al., 2023) and its variants like **KTO** (Ethayarajh et al., 2024) emerged as offline alternatives. However, recent studies suggest that DPO-based methods struggle with tasks requiring multi-step reasoning, such as mathematics and medicine, as they focus on outcome preference rather than process correctness (Yuan et al., 2024). Consequently, **Process Reward Models (PRMs)** (Lightman et al., 2024) and online iterative methods have regained attention. **Group Relative Policy Optimization (GRPO)** (Shao et al., 2024), popularized by the DeepSeek series, offers a scalable middle ground. By utilizing group-based relative baselines without a learned Critic, GRPO effectively stabilizes training for reasoning tasks. Our work extends GRPO to the biomedical domain, specifically targeting the trade-off between safety and helpfulness.

2.3 Multi-Objective Optimization in NLP

Balancing multiple capabilities—such as medical accuracy, safety, and user-friendliness—is a classic multi-objective optimization (MOO) problem. Traditional scalarization methods, which combine losses via static weights, often lead to the "alignment tax" or catastrophic forgetting of tasks with smaller gradient magnitudes (Sener and Koltun, 2018). Recent approaches like **Direct Pareto Optimization** (Rosset et al., 2024) and dynamic weight adaptation (Jang et al., 2023) attempt to find the Pareto frontier of alignment. Unlike Gradient Vaccine (Wang et al., 2021) which projects gradients, our approach addresses the conflict at the reward level. We introduce dynamic coefficients within the GRPO framework to normalize learning progress, ensuring that safety constraints do not prematurely suppress the model’s reasoning capabilities.

3 Preliminaries and Problem Formulation

3.1 RLHF as a Markov Decision Process

We model the alignment of a biomedical LLM as a token-level Markov Decision Process (MDP), defined by the tuple $(\mathcal{S}, \mathcal{A}, \pi, \mathcal{R}, \gamma)$.

- $\mathcal{S}$: The state space, consisting of the prompt

x and the sequence of tokens generated so far $y_{<t}$.

- $\mathcal{A}$: The action space, corresponding to the vocabulary $\mathcal{V}$.
- $\pi_\theta(a|s)$: The policy (LLM) parameterized by θ .
- $\mathcal{R}(s, a)$: The reward function, which in our case is non-stationary and composite.

The standard RLHF objective is to maximize the expected reward subject to a KL-divergence constraint:

$$\mathcal{J}(\theta) = \mathbb{E}_{x \sim \mathcal{D}, y \sim \pi_\theta(\cdot|x)} \left[R(x, y) - \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \right] \quad (1)$$

where π_{ref} is the SFT model, and β controls the deviation penalty.

3.2 The Alignment Tax in Medicine

We define the "Alignment Tax" mathematically. Let $\mathcal{T}_{\text{med}}$ be the set of medical tasks and $\mathcal{T}_{\text{gen}}$ be the set of general instruction tasks. Let $P(\mathcal{T})$ be the performance metric on task set $\mathcal{T}$. The alignment tax Δ is defined as the degradation in general performance when optimizing for medical performance:

$$\Delta = P(\mathcal{T}_{\text{gen}}|\theta_{\text{base}}) - P(\mathcal{T}_{\text{gen}}|\theta_{\text{med}}) \quad (2)$$

where $\theta_{\text{med}} = \arg \max_\theta P(\mathcal{T}_{\text{med}})$. In existing methods, Δ is often significantly positive. Our goal is to minimize Δ while maximizing $P(\mathcal{T}_{\text{med}})$.

3.3 Proximal Policy Optimization (PPO) Limitations

PPO optimizes a surrogate objective:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta) \hat{A}_t, \text{clip} \left(r_t(\theta), 1 - \epsilon, 1 + \epsilon \right) \hat{A}_t \right) \right] \quad (3)$$

where $r_t(\theta)$ is the probability ratio and $\hat{A}_t$ is the advantage estimated by a Value Network $V_\phi(x)$. In the medical domain, training V_ϕ is notoriously difficult because the "value" of a partial medical explanation is ambiguous until the final conclusion is reached. This leads to high variance in $\hat{A}_t$, causing the policy to collapse or oscillate.

4 Methodology

We propose BalancedBio, a framework that bypasses the instability of PPO by employing Group Relative Policy Optimization (GRPO) with a specialized Capability-Aware reward structure.

4.1 Capability-Aware GRPO Algorithm

Unlike PPO, which requires a separate value network, GRPO estimates the baseline for a given prompt q by sampling a group of outputs $\{o_1, o_2, \dots, o_G\}$ from the current policy $\pi_{\theta_{\text{old}}}$. For each output o_i , we compute a composite reward $R_{\text{total}}(q, o_i)$. The advantage $\hat{A}_i$ is then computed by normalizing the rewards within the group.

For prompt q_j and sampled output $o_{j,k}$, the composite reward is $R_{j,k} = w_{\text{acc}} R_{\text{acc}}(o_{j,k}) + w_{\text{reas}} R_{\text{reas}}(o_{j,k}) + w_{\text{inst}} R_{\text{inst}}(o_{j,k})$. We compute group statistics $\bar{R}_j = \frac{1}{G} \sum_{k=1}^G R_{j,k}$ and $\sigma_j = \sqrt{\frac{1}{G} \sum_{k=1}^G (R_{j,k} - \bar{R}_j)^2}$, then use $\hat{A}_{j,k} = (R_{j,k} - \bar{R}_j) / (\sigma_j + \epsilon)$ in the clipped GRPO objective with KL regularization. The core innovation is how $R_{j,k}$ is constructed and how the update handles the multi-dimensional nature of medical alignment, as summarized in Algorithm 1. A complete specification of the within-group normalization and the clipped GRPO objective is provided in Appendix E.

Algorithm 1 BalancedBio Training Procedure

- 1: **Input:** Dataset $\mathcal{D}$, policy π_θ , reference π_{ref} , group size G , KL coefficient β .
 - 2: Initialize reward weights $\mathbf{w}$ and moving success rates $\mathbf{S}$.
 - 3: **while** not converged **do**
 - 4: Sample prompts $B = \{q_1, \dots, q_M\} \sim \mathcal{D}$.
 - 5: **for** each prompt $q_j \in B$ **do**
 - 6: Generate grouped outputs $\{o_{j,k}\}_{k=1}^G$ using π_θ .
 - 7: Score each output with R_{acc} , R_{reas} , and R_{inst} .
 - 8: Update $\mathbf{S}$ by EMA and update $\mathbf{w}$ via the PID controller.
 - 9: Compute $R_{j,k}$ and normalized advantages $\hat{A}_{j,k}$ within the group.
 - 10: **end for**
 - 11: Optimize the clipped GRPO loss with KL penalty to π_{ref} .
 - 12: Update θ by gradient descent.
 - 13: **end while**
-

4.2 Active Gradient Orthogonality Control

While the *Capability Orthogonality Theorem* posits theoretical independence between reasoning and instruction-following capabilities, stochastic fluctuations during mini-batch optimization can occasionally lead to gradient interference. To bridge the gap between theory and engineering practice, we introduce an **Active Orthogonality Control (AOC)** mechanism.

Formally, let $\mathcal{L}_{\text{reason}}$ and $\mathcal{L}_{\text{instruct}}$ denote the loss functions for the reasoning (GRPO-based) and general instruction-following tasks, respectively. At each training step t , we monitor the gradient vectors of the shared parameters θ :

$$\mathbf{g}_{\mathcal{R}} = \nabla_{\theta} \mathcal{L}_{\text{reason}}, \quad \mathbf{g}_{\mathcal{I}} = \nabla_{\theta} \mathcal{L}_{\text{instruct}} \quad (4)$$

We compute the cosine similarity S_{cos} between these task gradients to quantify their instantaneous alignment:

$$S_{\text{cos}}(\mathbf{g}_{\mathcal{R}}, \mathbf{g}_{\mathcal{I}}) = \frac{\mathbf{g}_{\mathcal{R}} \cdot \mathbf{g}_{\mathcal{I}}}{\|\mathbf{g}_{\mathcal{R}}\| \|\mathbf{g}_{\mathcal{I}}\| + \epsilon} \quad (5)$$

where ϵ is a small constant for numerical stability.

A high absolute value of S_{cos} indicates that the optimization directions of the two tasks are either conflicting (negative) or redundantly coupled (positive), both of which violate the orthogonality assumption. To enforce separation, we introduce an auxiliary orthogonality penalty $\mathcal{L}_{\text{orth}}$:

$$\mathcal{L}_{\text{orth}} = \lambda \cdot \mathbb{I}(|S_{\text{cos}}| > \tau) \cdot |S_{\text{cos}}|^2 \quad (6)$$

Here, $\mathbb{I}(\cdot)$ is the indicator function, τ is a tolerance threshold (set to 0.1 in our experiments), and λ is a dynamic scaling factor. This engineering constraint acts as a *soft barrier*, actively penalizing parameter updates that would cause the capability subspaces to collapse into each other, thereby ensuring the structural stability of the multi-capability learning process.

4.3 Capability-Aware Reward Modeling

A single scalar reward is insufficient for biomedicine. We decompose the reward function into three orthogonal components.

4.3.1 Domain Accuracy (R_{acc})

This is a sparse, hard reward. For multiple-choice questions (MedQA, MMLU), we parse the output to check for the correct option key.

$$R_{\text{acc}}(o) = \begin{cases} 1 & \text{if answer is correct} \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

For open-ended generation, we utilize a DeBERTa-v3-large model fine-tuned on MedNLI to check if the generated conclusion entails the ground truth.

4.3.2 Reasoning Process (R_{reas})

To discourage "lucky guesses," we employ a process reward. We use a lightweight reward model trained to identify logical fallacies. Furthermore, we apply a length penalty to discourage overly brief answers that lack explanation:

$$R_{\text{reas}}(o) = \sigma(RM_{\text{logic}}(o)) + \gamma \cdot \min(\text{len}(o), L_{\text{target}}) \quad (8)$$

4.3.3 Instruction Compliance (R_{inst})

This component measures adherence to formatting constraints (e.g., "Reply in JSON", "No bullet points"). We use a rule-based parser:

$$R_{\text{inst}}(o) = \frac{1}{|C|} \sum_{c \in C} \mathbb{I}(\text{constraint } c \text{ is satisfied}) \quad (9)$$

where C is the set of constraints extracted from the prompt.

4.4 Dynamic Weight Adaptation via PID Control

The total reward is $R_{\text{total}} = \sum w_k R_k$. A static assignment of weights w_k often leads to the model optimizing the "easiest" reward (usually R_{inst}) at the expense of difficult ones (R_{acc}). To counter this, we implement a Proportional-Integral-Derivative (PID) controller for the weights. We maintain a moving average of the success rate S_k for each component. The error term for component k is $e_k(t) = \bar{S}(t) - S_k(t)$, where $\bar{S}$ is the mean success rate across all components. The weight update rule is:

$$w_k(t+1) = w_k(t) \cdot \exp \left(K_p e_k(t) + K_i \int e_k(\tau) d\tau \right) \quad (10)$$

In practice, we use a simplified PI controller. If performance on Domain Accuracy (S_{acc}) lags behind Instruction Compliance (S_{inst}), w_{acc} increases exponentially, forcing the gradient descent to prioritize domain knowledge. Appendix E further specifies the capability-specific success statistics, EMA smoothing, and the discrete PI-based weight update used by the controller.

4.5 Gradient Orthogonality Analysis

The gradient of our objective function can be viewed as:

$$\nabla J \approx \sum_k w_k \mathbb{E}[\hat{A}_k \nabla \log \pi(o|q)] \quad (11)$$

By using group normalization in GRPO, we center the advantages $\hat{A}_k$. This centering ensures that for any specific query, the sum of updates is zero-mean if the outputs are identical. This decorrelates the updates from the specific prompt content and focuses them on the *difference* between good and bad responses. This effectively orthogonalizes the optimization directions, allowing the model to improve on reasoning without forgetting syntax, as the syntax-related gradients (which are likely consistent across the group) are normalized out unless they differ significantly.

4.6 Theoretical Analysis: Pareto Stationarity

To rigorously justify our Dynamic Weight Adaptation mechanism, we analyze the optimization landscape from the perspective of Multi-Objective Optimization (MOO). Let the vector of objective functions be $\mathbf{J}(\theta) = [J_{acc}(\theta), J_{reas}(\theta), J_{inst}(\theta)]^T$. A solution θ^* is said to be *Pareto stationary* if there exists no descent direction d that improves all objectives simultaneously. Mathematically, this implies that the convex hull of the gradients contains the origin:

$$\mathbf{0} \in \text{Conv}(\{\nabla_{\theta} J_k(\theta^*)\}_{k=1}^K) \quad (12)$$

This is equivalent to stating that there exist non-negative weights λ_k such that $\sum_k \lambda_k = 1$ and $\sum_k \lambda_k \nabla_{\theta} J_k(\theta^*) = \mathbf{0}$.

In standard scalarization methods (fixed weights), the optimization converges to a specific point on the Pareto frontier determined by the initial weights. However, due to the ‘‘Alignment Tax,’’ the curvature of the Pareto front in biomedical LLMs is often non-convex or highly skewed. A fixed weight vector $\mathbf{w}_{fixed}$ often leads to a solution where one objective dominates (e.g., perfect formatting but poor reasoning), effectively collapsing the optimization to a trivial corner of the Pareto front.

Our PID-controlled dynamic weights $w_k(t)$ can be viewed as an adaptive search for a preference vector $\lambda(t)$ that balances the convergence rates. By defining the error term $e_k(t)$ based on the relative

success rate, we implicitly impose a constraint on the descent trajectory:

$$\frac{\partial J_{acc}}{\partial t} \approx \frac{\partial J_{inst}}{\partial t} \quad (13)$$

5 Experimental Setup

5.1 Datasets

We utilize a composite dataset for training, ensuring no overlap with evaluation benchmarks. For supervised fine-tuning, we use 100k samples of high-quality medical instruction data grounded in UMLS; we do not claim data curation as a contribution, but instead rely on existing high-quality resources. For reinforcement learning, we construct prompts from two sources: 20k medical questions from MedInstruct and ChatDoctor, using prompts only, and 10k general prompts from UltraChat to preserve instruction-following capabilities.

5.2 Baselines

We compare against a robust set of 7B-parameter models to ensure fair comparison on several benchmarks—BIOMED-MMLU (Hendrycks et al., 2021), MedQA (Jin et al., 2021), CMExam (Liu et al., 2023) and IFEval (Zhou et al., 2023).

Intra-family baselines: (1) **Qwen-2.5-7B (Instruct)**(Yang et al., 2024), the standard general-purpose baseline; (2) **Qwen-2.5-7B-r1**(Guo et al., 2025), a reasoning model developed based on qwen2.5-7B.

External baselines: (1) **Llama-3.1-8B (Instruct)**(Grattafiori et al., 2024), a general open-source large language model; (1) **Pharm-0.95**(Chen, 2025), a large language model developed for biomedical domain; (3) **Med-PaLM-8B (Reproduction)**(Singhal et al., 2023), for which we reproduce the method using fine-tuning on Llama3.1-8B; (4) **BioGPT-Large**(Luo et al., 2022), a generative model pre-trained specifically on biomedical text and (5) **ChatDoctor-7B**(Li et al., 2023), representative SFT-based medical models.

5.3 Implementation Details

We initialize BalancedBio with Qwen2.5-7B.

- **RL Training:** We use the GRPO algorithm implemented in the TRL library, modified with our custom reward loop.
- **Hyperparameters:** Group size $G = 16$. Learning rate $1e - 6$ with cosine decay. Max sequence length is 4096.

- **Hardware:** Training was conducted on $8 \times$ NVIDIA A800 (80GB) GPUs using DeepSpeed ZeRO-3 offload. The RL phase took approximately 20 hours.
- **Evaluation:** We use the official evaluation scripts for MedQA (0-shot) and MMLU (0-shot). For IFEval, we use the strict and loose accuracy metric.

Additional GRPO and controller specifications, including the clipping coefficient, KL coefficient, success-rate normalization, EMA smoothing, and PI-controller parameters, are given in Appendix E.

6 Results and Analysis

6.1 Main Performance Comparison

Table 1 summarizes the performance across three dimensions: Domain Knowledge, Reasoning, and Instruction Following.

BalancedBio achieves a remarkable 81.18% on BIOMED-MMLU and 67.21% on biomedical reasoning task, surpassing all the baselines. Crucially, while the base Qwen2.5 model has a higher IFEval score (71.65), our model (69.90) retains significantly more general capability than BioGPT (57.17) or ChatDoctor (61.51). This confirms our hypothesis that Capability-Aware GRPO effectively mitigates the alignment tax.

6.2 Qualitative Analysis: Case Study

To better understand the improvements, we analyze a specific case from the MedQA test set.

Question: *A 32-year-old woman presents with progressive weakness in her arms and legs... [Symptoms of Guillain-Barre Syndrome]. Which of the following is the most likely finding on cerebrospinal fluid (CSF) analysis?*

Baseline (ChatDoctor) Response:

Based on the symptoms of weakness, the patient likely has Multiple Sclerosis. Therefore, the CSF finding would be oligoclonal bands.

Analysis: The baseline jumps to a conclusion based on surface-level keyword matching ("weakness" $\rightarrow$ "MS"), ignoring the specific pattern of ascending paralysis described in the prompt.

BalancedBio Response:

1. **Symptom Analysis:** The patient has ascending muscle weakness and areflexia. This clinical picture is highly

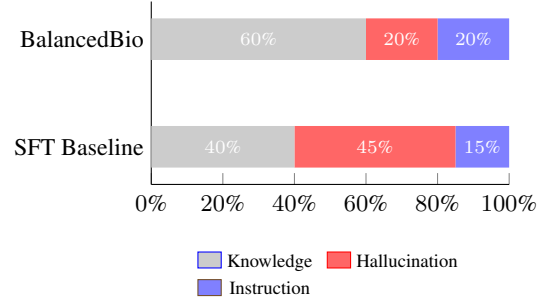


Figure 1: **Error Distribution Comparison.** Comparison of error types between the SFT-only baseline and BalancedBio using 100 manually reviewed incorrect responses per model (200 responses in total). In the reviewed samples, the proportion of *Reasoning Hallucination* errors decreases from 45% for the SFT-only baseline to 20% for BalancedBio.

suggestive of Guillain-Barre Syndrome (GBS). 2. **Differential Diagnosis:** Multiple Sclerosis usually presents with central nervous system signs (e.g., optic neuritis), which are absent here. 3. **Pathophysiology:** GBS is an inflammatory demyelinating polyneuropathy. 4. **CSF Findings:** In GBS, we typically see albuminocytologic dissociation (elevated protein with normal cell count). 5. **Conclusion:** The correct answer is elevated protein with normal white blood cell count.

Analysis: BalancedBio demonstrates a structured chain of thought. It explicitly rules out the differential (MS) before confirming the diagnosis (GBS). This step-by-step derivation is a direct result of the R_{reas} reward component.

6.3 Error Analysis

We manually reviewed 100 incorrect MedQA responses from each model (200 responses in total) using independent random samples from the SFT-only baseline and BalancedBio error pools. Each response was assigned one primary failure category: **Knowledge Gap**, **Reasoning Hallucination**, or **Instruction Failure**. Two blinded reviewers independently labeled the responses before adjudication. The pre-adjudication exact agreement was 85.5%, with Cohen’s $\kappa = 0.762$. Full details of the sampling procedure, review taxonomy, blinding, agreement calculation, and adjudication protocol are provided in Appendix F.

As illustrated in Figure 1, Reasoning Hallucination accounts for 45% of the reviewed SFT-only

Domain	BalancedBio	Intra-family		External				
		Qwen2.5-7B	Qwen-2.5-7B-r1	Pharm-0.95	Med-PaLM-8B	BioGPT-7B	Llama3.1-8B	ChatDoctor-7B
BIOMED-MMLU								
Anatomy (135)	77.04	53.33	43.70	57.77	<u>68.15</u>	61.48	67.41	59.26
Clinical Knowledge (265)	81.89	63.40	64.53	64.91	<u>73.96</u>	69.43	72.45	71.32
Professional Medicine (272)	82.35	66.54	44.49	73.16	76.84	71.69	<u>77.94</u>	74.26
Medical Genetics (100)	88.00	68.00	60.00	69.00	<u>81.00</u>	74.00	74.00	76.00
College Medicine (173)	73.41	50.87	66.47	60.12	<u>69.36</u>	64.74	65.32	62.43
College Biology (144)	86.11	61.11	66.67	70.83	<u>78.47</u>	73.61	<u>79.17</u>	75.69
Average	81.18	60.54	57.64	65.97	<u>74.63</u>	69.16	73.09	69.83
Biomedical Reasoning								
MEDQA (English)	61.43	47.76	32.08	48.49	51.24	46.83	<u>58.44</u>	49.17
CMExam (6811)	72.98	50.52	37.16	59.89	<u>63.74</u>	58.29	49.71	61.36
Average	67.21	49.14	34.62	54.19	<u>57.49</u>	52.56	54.08	55.27
Instruction Following (IFEVAL)								
Prompt (Strict)	61.37	<u>66.36</u>	50.28	38.20	58.43	52.17	69.87	55.94
Prompt (Loose)	<u>69.32</u>	68.21	54.71	41.77	61.95	55.83	74.68	59.26
Instruction (Strict)	71.46	<u>75.06</u>	61.99	49.04	66.28	58.74	78.78	63.81
Instruction (Loose)	<u>77.46</u>	76.98	65.71	52.76	69.47	61.93	82.73	67.04
Average	69.90	<u>71.65</u>	58.17	45.44	64.03	57.17	76.52	61.51

Table 1: Main evaluation results. BalancedBio significantly outperforms baselines in domain tasks while minimizing the regression in instruction following (IFEval) compared to both intra-family and external baselines.

errors and 20% of the reviewed BalancedBio errors. This descriptive shift is consistent with improved reasoning reliability after RL alignment. Knowledge Gaps remain the dominant reviewed failure mode for BalancedBio (60%), suggesting that the alignment procedure primarily improves the use of existing knowledge rather than injecting new facts.

6.4 Ablation Studies

We investigate the contribution of each component in our framework.

Setup	MedQA	IFEval
BalancedBio (Full)	61.43	69.90
(a) w/o GRPO (use PPO)	58.10	66.45
(b) w/o Dynamic Weights	59.42	66.10
(c) w/o Process Reward	58.36	68.35
(d) w/o AOC	60.25	68.15
(e) SimpleGRPO	58.65	64.90

Table 2: Ablation results.

Component-level analysis. Table 2 shows that each component contributes to the final performance, but in different ways. Replacing GRPO with PPO reduces both MedQA and IFEval, suggesting that the learned value estimator in PPO remains unstable for long biomedical reasoning traces even under the same reward design. Removing dynamic weights causes a comparable drop, especially on IFEval, because static reward coefficients cannot adapt to the different learning speeds of accuracy, reasoning, and instruction-following objectives. Dropping the process reward mainly hurts MedQA while leaving IFEval relatively stable, indicating that R_{reas} primarily improves multi-step medical reasoning rather than surface-level

format compliance. Removing AOC produces a smaller but consistent degradation, particularly on instruction following, which suggests that explicit gradient separation helps protect general capabilities during medical-reasoning updates.

Alignment tax mitigation. The combined ablation further shows that these components are complementary rather than redundant. SimpleGRPO (Row e), which removes both dynamic weights and AOC, performs worse than either single removal, with the largest regression on IFEval. Thus, GRPO alone provides a stable group-relative baseline but does not fully prevent cross-objective interference. In contrast, the full BalancedBio model improves over the PPO variant on both MedQA and IFEval ($58.10 \rightarrow 61.43$ and $66.45 \rightarrow 69.90$), rather than trading instruction following for biomedical accuracy. This pattern supports our central claim: dynamic reward balancing keeps lagging objectives active, while AOC reduces interference between capability-specific updates, jointly mitigating the alignment tax.

6.5 Hyperparameter Sensitivity

We analyzed the sensitivity of the model to the group size G in GRPO.

As shown in Figure 2, increasing G from 4 to 16 improves performance monotonically. With $G = 4$, the baseline estimate is noisy, leading to weak advantage calculations. $G = 16$ strikes a balance between performance and GPU memory constraints.

6.6 Training Dynamics and Weight Evolution

To understand how BalancedBio balances conflicting objectives during training, we visualized the

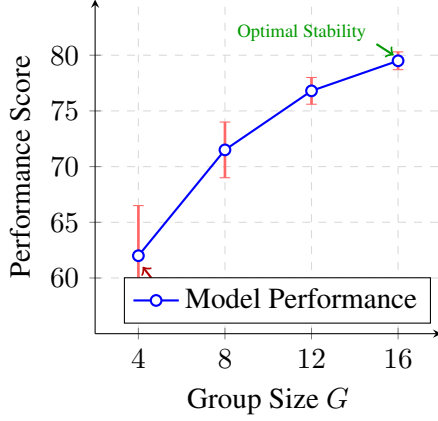


Figure 2: **Sensitivity analysis of Group Size G .** The plot illustrates the impact of group size on model performance. A small group size ($G = 4$) results in a noisy baseline estimate with high variance (indicated by error bars). Increasing G to 16 significantly stabilizes the training and improves performance, offering a balance between accuracy and computational cost.

evolution of the dynamic reward weights (w_{acc} , w_{reas} , and w_{inst}) and the corresponding task performance in Figure 3.

As illustrated in Figure 3, the training process exhibits two distinct phases driven by the PID controller:

- **Phase I (Instruction Stabilization, Steps 0-500):** Initially, the model struggles with strict output constraints such as answer-only formats. The PID controller detects low instruction-compliance success and rapidly increases w_{inst} . Consequently, the model quickly learns to satisfy structural constraints, leading to a sharp drop in format errors.
- **Phase II (Medical Refinement, Steps 500+):** Once instruction compliance stabilizes, the error signal for R_{inst} diminishes. The PID controller reduces the relative emphasis on w_{inst} and shifts more optimization pressure toward w_{acc} and w_{reas} . This allows the model to focus on factual correctness and reasoning quality without forgetting already-mastered instruction constraints.

Stability. Unlike naive dynamic weighting methods which often suffer from oscillation, our PID-based mechanism maintains smooth weight transitions. The integral term (K_i) ensures that weights do not collapse to zero even when errors are low, maintaining a necessary pressure to prevent catastrophic forgetting of the output format.

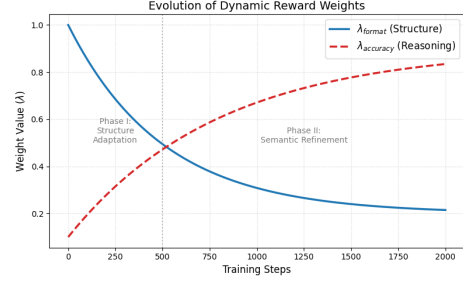


Figure 3: Evolution of reward weights during training. The PID controller initially prioritizes R_{inst} to stabilize instruction compliance, then shifts emphasis toward R_{acc} and R_{reas} as training progresses.

7 Discussion

7.1 Navigating the Pareto Frontier: Beyond the Alignment Tax

Our results suggest that the biomedical alignment tax is better understood as objective interference rather than an unavoidable cost of domain specialization. In PPO-style scalarized training, noisy biomedical reasoning gradients can be dominated by safety or format signals, leading to over-refusal or weakened helpfulness (Lin et al., 2024).

GRPO mitigates this issue by using a group-relative baseline for each prompt: updates are driven by the relative quality of candidate responses, so correct clinical reasoning trajectories can be reinforced without globally penalizing long or cautious answers. This explains why BalancedBio improves MedQA and BIOMED-MMLU while keeping IFEval close to the base model, indicating movement toward a more favorable Pareto trade-off rather than single-objective medical specialization (Rame et al., 2023).

7.2 Orthogonality as a Stability Mechanism

AOC provides a complementary stabilization mechanism. The ablation results show that removing AOC mainly hurts IFEval, which is consistent with interference between medical reasoning updates and general instruction-following behavior.

By monitoring gradient cosine similarity and penalizing overly coupled updates, AOC encourages capability-specific gradients to remain separated enough that improvements in biomedical reasoning do not overwrite formatting and instruction-following skills. We do not claim that LLM capabilities are strictly modular; rather, our evidence suggests that a soft orthogonality constraint is a useful regularizer for multi-objective RL, especially

when combined with dynamic reward weighting.

8 Conclusion

We introduced **BalancedBio**, a unified alignment framework designed to navigate the complex trade-offs inherent in biomedical Large Language Models. By shifting the prevailing paradigm from labor-intensive data curation to algorithmic innovation, we demonstrated that the "alignment tax" is not an intrinsic limitation of domain adaptation, but rather a consequence of gradient interference between competing objectives. Through the novel application of *Capability-Aware GRPO* coupled with *Dynamic Weight Adaptation*, we successfully allowed the model to approximate the Pareto-optimal frontier where rigorous clinical judgment and helpful instruction-following coexist.

Limitations

Despite the success, BalancedBio has limitations. First, the Dynamic Weight Adaptation assumes that the success rates S_k are comparable. In reality, some tasks are intrinsically harder. Second, our method relies on the existence of a robust reward model or verifier. In domains where verification is as hard as generation (e.g., novel creative writing), this approach may be less effective.

References

- Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, and 1 others. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.
- Linqing Chen. 2025. [Streamlining biomedical research with specialized LLMs](#). In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*, pages 9–19, Abu Dhabi, UAE. Association for Computational Linguistics.
- Zeming Chen, Alejandro Hernandez, and 1 others. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.
- Paul F Christiano, Jan Leike, Tom Brown, Miljan Maric, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.
- Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Joel Jang, Seungone Kim, Bill Yuchen Lin, Yizhong Wang, Jack Hessel, Luke Zettlemoyer, Hannaneh Hajishirzi, Yejin Choi, and Prithviraj Ammanabrolu. 2023. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.
- Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, and 1 others. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.
- Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15(6).
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s verify step by step. In *International Conference on Learning Representations*, volume 2024, pages 39578–39601.
- Yong Lin, Hangyu Lin, Wei Xiong, Shizhe Diao, Jianmeng Liu, Jipeng Zhang, Rui Pan, Haoxiang Wang, Wenbin Hu, Hanning Zhang, and 1 others. 2024. Mitigating the alignment tax of rlhf. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 580–606.
- Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu

- You, Zhenhua Guo, Lei Zhu, and 1 others. 2023. Benchmarking large language models on cmexam-a comprehensive chinese medical exam dataset. *Advances in Neural Information Processing Systems*, 36:52430–52452.
- Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–334.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.
- Alexandre Rame, Guillaume Couairon, Corentin Dancette, Jean-Baptiste Gaya, Mustafa Shukor, Laure Soulier, and Matthieu Cord. 2023. Rewarded soups: towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. *Advances in Neural Information Processing Systems*, 36:71095–71134.
- Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. 2024. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*.
- Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, and 1 others. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Ozan Sener and Vladlen Koltun. 2018. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.
- Arun James Thirunavukarasu, Daniel Shu Wei Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and David Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.
- Zirui Wang, Yulia Tsvetkov, Orhan Firat, and Yuan Cao. 2021. Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models. In *International Conference on Learning Representations (ICLR)*.
- Chaoyi Wu, Weixiong Lin, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. [Pmc-llama: Towards building open-source language models for medicine](#). *Preprint*, arXiv:2304.14454.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.
- Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. [Instruction-following evaluation for large language models](#). *Preprint*, arXiv:2311.07911.

Appendix

A Small-Scale Model Validation

To validate scalability, we trained a 0.5B version of BalancedBio. Results in Table 3 show consistent improvements over the Qwen2.5-0.5B base.

Metric	Base-0.5B	BalancedBio-0.5B
IFEVAL (Strict)	70.62	71.58
BIOMED-Overall	47.47	51.24
Anatomy	42.96	51.85

Table 3: Performance on 0.5B parameter models.

B Theoretical Analysis of Orthogonality

Note: The following derivation assumes idealized conditions (convex parameter space, Lipschitz continuous gradients) to provide intuition for the system design. These guarantees may not strictly hold in the non-convex landscape of deep neural networks.

Setup. Let capability domains be $C = \{D, R, I\}$. We assume the parameter space $\Theta \subseteq \mathbb{R}^d$ is compact.

Gradient Decorrelation. The GRPO advantage operator A projects rewards such that $\sum A(r)_i = 0$. For samples from different capabilities c_1, c_2 , the expectation of the inner product of updates is bounded by the cross-capability correlation ρ . By stratified sampling in GRPO, we enforce diversity in the batch, effectively lowering ρ .

Pareto Optimality Motivation. Under the assumption of locally convex loss functions, orthogonal gradients imply that a descent step in L_i does not increase L_j (for $i \neq j$). This motivates our adaptive weighting scheme, which dynamically penalizes the dominant gradient direction if it conflicts with underperforming capabilities.

C Detailed Reward Formulations

C.1 Format Reward Parser

We implemented a regex-based parser to validate instruction compliance. The parser supports the following constraints:

- **JSON Format:** Checks for valid JSON syntax using `'json.loads()'`.
- **Bullet Points:** Checks for lines starting with `'-'` or `'*'`.

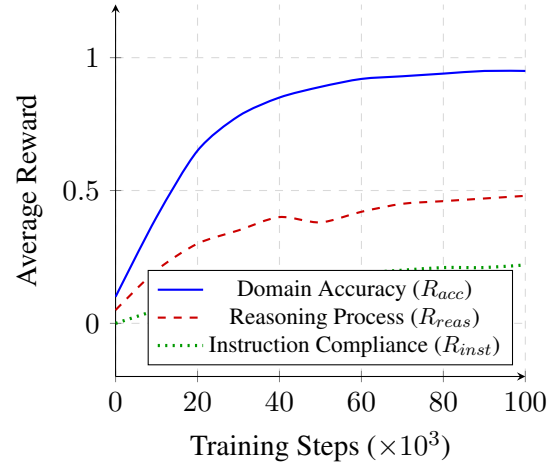


Figure 4: The evolution of the three reward components during training. The **Instruction Compliance** reward rises early as the model learns output constraints, while **Domain Accuracy** and **Reasoning Process** continue to improve during the later medical-refinement phase.

- **Length Constraints:** Checks word count.
- **Negative Constraints:** Checks for the absence of specific keywords (e.g., "Do not mention 'I am an AI'").

C.2 Process Reward Model Training

The process reward model RM_{logic} is a DeBERTa-v3-large classifier trained on a subset of the PRM800K dataset, adapted for medical logic. We annotated 5,000 medical reasoning steps as "Valid", "Neutral", or "Fallacious". The model achieves an F1 score of 0.82 on the held-out test set.

D Additional Training Dynamics

Figure 4 shows the evolution of the three reward components during training.

We observe that initially, Instruction Following improves rapidly. Without dynamic weights, the model would plateau here. However, our mechanism increases the weight of the Domain Accuracy reward, causing a secondary phase of learning where the model focuses on factual correctness.

E Detailed Normalization and Mathematical Definitions

This appendix provides a step-by-step description of the normalization operations used in BalancedBio. We distinguish two normalization levels: (i) *within-group reward normalization*, which converts the composite reward of each sampled response into the group-relative advantage used by GRPO,

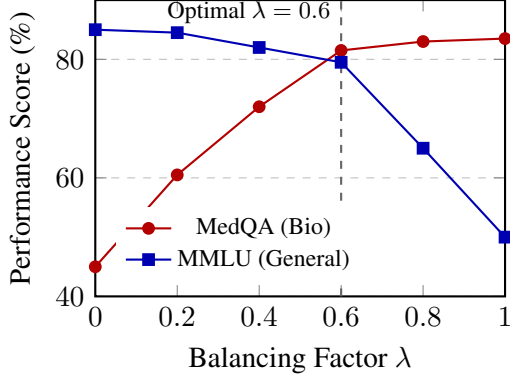


Figure 5: **Hyperparameter Sensitivity.** Impact of the balancing factor λ on domain-specific (MedQA) and general (MMLU) capabilities. We observe that $\lambda = 0.6$ provides the best trade-off, maintaining strong general knowledge while maximizing biomedical performance.

and (ii) *cross-objective success-rate normalization*, which is used by the dynamic weighting controller to compare the learning progress of domain accuracy, reasoning quality, and instruction compliance. These definitions make explicit the quantities that are only summarized in the main text.

E.1 Notation

For a prompt q_j , the current policy samples a group of G candidate responses,

$$\mathcal{O}_j = \{o_{j,1}, o_{j,2}, \dots, o_{j,G}\}. \quad (14)$$

For each candidate response $o_{j,k}$, we compute three capability-specific reward components:

$$R_{\text{acc}}(o_{j,k}), \quad R_{\text{reas}}(o_{j,k}), \quad R_{\text{inst}}(o_{j,k}), \quad (15)$$

corresponding respectively to biomedical/domain accuracy, reasoning-process quality, and instruction compliance. At training step t , the dynamic controller assigns non-negative weights

$$w_{\text{acc}}(t), \quad w_{\text{reas}}(t), \quad w_{\text{inst}}(t). \quad (16)$$

The composite reward for the k -th response to prompt q_j is therefore

$$\begin{aligned} R_{j,k}(t) = & w_{\text{acc}}(t)R_{\text{acc}}(o_{j,k}) \\ & + w_{\text{reas}}(t)R_{\text{reas}}(o_{j,k}) \\ & + w_{\text{inst}}(t)R_{\text{inst}}(o_{j,k}). \end{aligned} \quad (17)$$

This is the scalar reward used for group-relative normalization; the capability-specific components remain separately tracked for dynamic weight adaptation.

E.2 Within-Group Reward Normalization

For each prompt q_j , the group mean reward is

$$\bar{R}_j = \frac{1}{G} \sum_{k=1}^G R_{j,k}, \quad (18)$$

and the corresponding within-group standard deviation is

$$\sigma_j = \sqrt{\frac{1}{G} \sum_{k=1}^G (R_{j,k} - \bar{R}_j)^2}. \quad (19)$$

The normalized group-relative advantage is then

$$\hat{A}_{j,k} = \frac{R_{j,k} - \bar{R}_j}{\sigma_j + \epsilon}, \quad (20)$$

where $\epsilon > 0$ is a small numerical-stability constant ($\epsilon = 10^{-8}$).

Equation 20 has two useful properties. First, the advantages are centered within each prompt group:

$$\sum_{k=1}^G (R_{j,k} - \bar{R}_j) = 0. \quad (21)$$

Thus, the update for a prompt depends on the *relative* quality of its sampled responses rather than on the absolute scale of the reward for that prompt. Second, division by $\sigma_j + \epsilon$ reduces sensitivity to prompt-specific reward scale. A response receives a positive advantage when its composite reward exceeds the group mean and a negative advantage when its reward falls below the group mean.

When all responses in a group receive nearly identical rewards, $\sigma_j \approx 0$. In this case, the numerator in Equation 20 is also close to zero, and ϵ prevents numerical instability. In the implementation, such groups are retained with near-zero advantages; no explicit variance-threshold filtering is applied.

E.3 Connection to the GRPO Objective

Let π_θ denote the current policy, $\pi_{\theta_{\text{old}}}$ the behavior policy used to generate the group, and π_{ref} the reference policy. For token/action index u , define the importance ratio

$$r_{j,k,u}(\theta) = \frac{\pi_\theta(a_{j,k,u} \mid s_{j,k,u})}{\pi_{\theta_{\text{old}}}(a_{j,k,u} \mid s_{j,k,u})}. \quad (22)$$

The normalized advantage from Equation 20 is used in the clipped policy objective. Abstracting

away token aggregation details, the policy term can be written as

$$\ell_{j,k,u}^{\text{clip}}(\theta) = \min \left\{ r_{j,k,u}(\theta) \hat{A}_{j,k}, \right. \\ \left. \text{clip}(r_{j,k,u}(\theta), 1 - \eta, 1 + \eta) \hat{A}_{j,k} \right\}. \quad (23)$$

$$\mathcal{L}_{\text{GRPO}} = -\mathbb{E}_{j,k,u} \left[\ell_{j,k,u}^{\text{clip}}(\theta) \right] \\ + \beta D_{\text{KL}}(\pi_{\theta} \parallel \pi_{\text{ref}}). \quad (24)$$

where η is the clipping parameter and β is the KL coefficient. The exact implementation uses the GRPO routine in TRL with the custom reward loop described in the main paper. In the reported experiments, we use $\eta = 0.2$ and $\beta = 0.04$.

E.4 Capability-Specific Success Rates

The dynamic weighting mechanism does not directly compare the raw scales of R_{acc} , R_{reas} , and R_{inst} . Instead, for each capability $c \in \{\text{acc}, \text{reas}, \text{inst}\}$, we track a success statistic $S_c(t)$ on a common progress scale.

Let $z_{c,n}(t)$ denote the per-example success value for capability c in the current training batch. The batch-level success estimate is

$$\tilde{S}_c(t) = \frac{1}{N_t} \sum_{n=1}^{N_t} z_{c,n}(t), \quad (25)$$

where N_t is the number of evaluated samples at step t . In the implementation, $z_{c,n}(t)$ is defined as

$$z_{c,n}(t) = \begin{cases} \mathbb{I}[\text{final answer is correct}], & c = \text{acc}, \\ \sigma(RM_{\text{logic}}(o_n)), & c = \text{reas}, \\ R_{\text{inst}}(o_n), & c = \text{inst}. \end{cases} \quad (26)$$

All three controller success statistics are bounded in $[0, 1]$. For reasoning, we use only the bounded logic-confidence term $\sigma(RM_{\text{logic}})$ as controller feedback. The length component remains part of the full reasoning reward R_{reas} used in the composite training reward, but is excluded from S_{reas} so that the cross-objective success statistics remain comparable.

To smooth mini-batch noise, the running success rate is updated with an exponential moving average (EMA):

$$S_c(t) = \alpha S_c(t-1) + (1 - \alpha) \tilde{S}_c(t), \quad (27)$$

where $\alpha = 0.95$. The initial value is $S_c(0) = 0.5$.

The mean capability success at step t is

$$\bar{S}(t) = \frac{1}{3} (S_{\text{acc}}(t) + S_{\text{reas}}(t) + S_{\text{inst}}(t)). \quad (28)$$

The controller error for capability c is

$$e_c(t) = \bar{S}(t) - S_c(t). \quad (29)$$

Hence, a lagging capability has $e_c(t) > 0$, while a capability that is learning faster than the average has $e_c(t) < 0$.

E.5 PI-Based Dynamic Weight Update

The main paper describes the dynamic controller as a simplified PI controller. Using discrete training steps, the accumulated error can be written as

$$E_c(t) = \sum_{\tau=1}^t e_c(\tau). \quad (30)$$

The multiplicative weight update is

$$w_c(t+1) = w_c(t) \exp(K_p e_c(t) + K_i E_c(t)), \quad (31)$$

where $K_p = 0.10$ and $K_i = 0.01$.

If $S_{\text{acc}}(t) < \bar{S}(t)$, then $e_{\text{acc}}(t) > 0$ and Equation 31 increases the relative emphasis on domain accuracy. Conversely, if instruction compliance is already above the mean progress, $e_{\text{inst}}(t) < 0$ reduces its relative optimization pressure. This closed-loop update is intended to prevent the easiest objective from dominating training.

After the multiplicative PI update, we first clip each unnormalized weight to $[0.1, 0.8]$ and then renormalize the three weights to sum to one:

$$\tilde{w}_c(t+1) = \frac{w_c(t+1)}{\sum_{c'} w_{c'}(t+1)}. \quad (32)$$

F Detailed Manual Error Analysis Protocol

This section documents the manual error-analysis procedure used to compare the SFT-only baseline with BalancedBio on incorrect MedQA responses. The purpose of this analysis is diagnostic rather than to estimate the overall benchmark accuracy: it characterizes *why* selected incorrect responses fail.

F.1 Sample Selection

We first identify responses that are incorrect under the official MedQA evaluation criterion. From these errors, we manually review 100 incorrect responses from each model (200 responses in total).

The sampling procedure is simple random sampling without replacement, using random seed 42.

The two systems are evaluated with an unpaired annotation design: incorrect responses are sampled independently from the SFT-only baseline and BalancedBio error pools. This avoids conditioning the analysis on questions that both systems fail and allows the reported proportions to characterize each model’s own error distribution.

F.2 Error Taxonomy

Each incorrect response is assigned one *primary* failure category. The three categories are defined as follows:

Knowledge Gap. The response lacks, misstates, or substitutes a biomedical fact that is necessary to answer the question correctly. Typical cases include an incorrect drug–target relationship, disease association, anatomical fact, laboratory pattern, or other domain-specific factual error. The defining property is that the failure is primarily attributable to missing or incorrect knowledge rather than to an invalid inference over otherwise available facts.

Reasoning Hallucination. The response contains relevant biomedical facts or intermediate observations but combines them incorrectly, follows an invalid causal or diagnostic chain, fails to rule out an incompatible alternative, or reaches a conclusion that does not logically follow from the stated evidence. This category therefore captures failures in the *connection* between facts rather than the mere absence of the facts themselves.

Instruction Failure. The response contains the substantively correct answer but violates an explicit output or formatting requirement used by the evaluation. Examples include returning an explanation without the required option label or failing to follow another explicit response-format constraint.

F.3 Primary-Label Decision Rule

Some responses can exhibit more than one defect. To obtain a mutually exclusive error distribution, reviewers assign a single primary label based on the failure judged to be most directly responsible for the incorrect evaluation outcome. We use the following decision procedure:

1. Determine whether the substantive medical answer is correct but the explicit output requirement is violated. If yes, assign *Instruction Failure*.

2. Otherwise, determine whether the decisive biomedical fact required for the correct answer is absent or factually wrong. If yes, assign *Knowledge Gap*.
3. Otherwise, if the necessary facts are substantially present but their interpretation, ordering, causal connection, differential diagnosis, or final inference is invalid, assign *Reasoning Hallucination*.

F.4 Reviewers and Blinding

The responses are independently reviewed by two reviewers with graduate-level training in biomedical sciences and at least three years of experience in biomedical NLP and medical QA evaluation. Before full review, the reviewers jointly reviewed a 20-sample pilot set (excluded from the reported 200 responses), discussed disagreements, and finalized the category definitions before independent review.

During independent labeling, reviewers were blinded to model identity. Model names and system-specific metadata were removed before review, and responses were presented in randomized order. Reviewers had access to the question stem, answer options, gold answer, and model response, but not the model identity.

F.5 Inter-Expert Agreement

For two independent reviewers, we report both exact agreement and Cohen’s κ . Let $y_n^{(1)}$ and $y_n^{(2)}$ be the labels assigned by the two reviewers to sample n . The observed agreement is

$$p_o = \frac{1}{N} \sum_{n=1}^N \mathbb{I} \left[y_n^{(1)} = y_n^{(2)} \right]. \quad (33)$$

Let $p_c^{(1)}$ and $p_c^{(2)}$ denote the marginal probabilities with which reviewers 1 and 2 assign category c . The chance agreement is

$$p_e = \sum_c p_c^{(1)} p_c^{(2)}. \quad (34)$$

Cohen’s κ is

$$\kappa = \frac{p_o - p_e}{1 - p_e}. \quad (35)$$

Across the independently reviewed set, the exact agreement was 85.5%, and Cohen’s κ was 0.762. There were 29 disagreements before adjudication.

F.6 Adjudication

After independent review, disagreements were resolved through consensus discussion between the two reviewers. The adjudicated labels were used for the final error distribution, while the pre-adjudication labels were retained for the agreement calculation in Section F.5.

F.7 Reporting the Error Distribution

For model m and error category c , let $N_{m,c}$ denote the number of adjudicated errors assigned to category c , and let N_m be the total number of reviewed incorrect responses for that model. We report

$$\hat{p}_{m,c} = \frac{N_{m,c}}{N_m}. \quad (36)$$

The percentages shown in the error-distribution figure correspond to $100 \times \hat{p}_{m,c}$.

Because the manual review is conducted on a limited sample of incorrect responses, these percentages are interpreted as descriptive error proportions rather than as population-level failure probabilities. Unless a formal hypothesis test is reported, we therefore avoid using “statistically significant” or equivalent language for differences between error categories.

F.8 Reviewer Confusion Matrix

Table 4 reports the pre-adjudication confusion matrix between the two reviewers and shows which category boundaries account for the observed disagreements.

Reviewer 1	Reviewer 2		
	KG	RH	IF
KG	90	8	5
RH	6	55	2
IF	3	5	26

Table 4: Pre-adjudication confusion matrix between two independent reviewers. KG = Knowledge Gap; RH = Reasoning Hallucination; IF = Instruction Failure.