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

Anonymous ACL submission

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-7B by 6.3% in domain tasks while maintaining robust general instruction-following capabilities. We will release our model and partial data.

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 dynamics*. 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 pretraining, recent advancements have focused on multimodal capabilities and clinical reasoning. **Med-Gemini** (Saheb 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. (2024) and **?**, 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., 2023) 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., 2024) 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:

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

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 $\mathcal{T}_{med}$ be the set of medical tasks and $\mathcal{T}_{gen}$ 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:

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

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

3.3 Proximal Policy Optimization (PPO) Limitations

PPO optimizes a surrogate objective:

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

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 $\{o_1, o_2, \ldots, o_G\}$ from the current policy $\pi_{\theta_{old}}$. For each output o_i, we compute a composite reward $R_{total}(q, o_i)$. The advantage $\hat{A}_i$ is then computed by normalizing the rewards within the group.

The core innovation in our approach is how R_{total} is constructed and how the optimization step handles the multi-dimensional nature of medical alignment. The full training procedure is detailed in Algorithm 1.

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:

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

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

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

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 $\mathcal{L}_{\text{orth}}$:

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

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.

Algorithm 1 BalancedBio Training Procedure

Input: Dataset $\mathcal{D}$, Initial Policy π_θ, Reference Policy π_{ref}, Group Size G, Learning Rate η, KL coef β.
Initialize: Reward weights $\mathbf{w} = [1, 1, 1]$.
Initialize: Moving average success rates $\mathbf{S} = [0, 0, 0]$.
while not converged **do**
  Sample batch of prompts $B = \{q_1, \ldots, q_M\} \sim \mathcal{D}$.
  for each prompt $q_j \in B$ **do**
    Generate G outputs $\{o_{j,1}, \ldots, o_{j,G}\}$ using π_θ.
    for $k = 1$ to G **do**
      Compute Rewards:
        $r_{acc} \leftarrow R_{acc}(o_{j,k})$
        $r_{reas} \leftarrow R_{reas}(o_{j,k})$
        $r_{inst} \leftarrow R_{inst}(o_{j,k})$
      Update Success Rates $\mathbf{S}$ (Exponential Moving Avg).
      Update Weights $\mathbf{w}$ via PID Controller.
      $R_{total} \leftarrow w_{acc} r_{acc} + w_{reas} r_{reas} + w_{inst} r_{inst}$.
    end for
    Compute Mean Reward: $\bar{R}_j = \frac{1}{G} \sum_{k=1}^{G} R_{total}(o_{j,k})$.
    Compute Std Dev: $\sigma_j = \sqrt{\frac{1}{G} \sum_{k=1}^{G} (R_{total}(o_{j,k}) - \bar{R}_j)^2}$.
    for $k = 1$ to G **do**
      Advantage: $\hat{A}_{j,k} = \frac{R_{total}(o_{j,k}) - \bar{R}_j}{\sigma_j + \epsilon}$.
    end for
  end for
  Compute Loss:
  $L(\theta) = -\frac{1}{MG} \sum_{j,k} \left[\min(r\hat{A}, \text{clip}(r)\hat{A}) - \beta D_{KL}(\pi_\theta \| \pi_{ref}) \right]$
  Update $\theta \leftarrow \theta - \eta \nabla_\theta L(\theta)$.
end while

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 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.

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

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 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:

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

4.3.3 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:

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

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

4.4 Dynamic Weight Adaptation via PID Control

The total reward is $R_{total} = \sum w_k R_k$. 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:

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

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.

4.5 Gradient Orthogonality Analysis

The gradient of our objective function can be viewed as:

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

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:

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

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

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:

$$\frac{\partial J_{acc}}{\partial t} \approx \frac{\partial J_{inst}}{\partial t} \tag{13}$$

5 Experimental Setup

5.1 Datasets

We utilize a composite dataset for training, ensuring no overlap with evaluation benchmarks.

- **SFT Data:** We utilize 100k samples of high-quality medical instruction data. Note that we do not claim the data curation as a contribution in this work; we simply utilize existing high-quality resources grounded in UMLS.
- **RL Prompts:**
  - *Medical:* 20k questions from MedInstruct and ChatDoctor (prompts only).
  - *General:* 10k prompts from UltraChat to maintain general capabilities.

5.2 Baselines

We compare against a robust set of 7B-parameter models to ensure fair comparison:

- **Llama-2-7B-Chat:** The standard general-purpose baseline.
- **Med-PaLM (Reproduction):** Since the official weights are closed, we reproduced the method using prompt tuning on Llama-2-7B.
- **BioGPT-Large:** A generative model pre-trained specifically on biomedical text.
- **ChatDoctor & PMC-Llama:** Representative SFT-based medical models.
- **Mistral-7B-Instruct-v0.2:** A strong general baseline known for reasoning.

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 $G = 16$. KL coefficient $\beta = 0.04$. Learning rate $1e-6$ with cosine decay. Max sequence length is 2048.
- **Hardware:** Training was conducted on 8 $\times$ NVIDIA A100 (80GB) GPUs using DeepSpeed ZeRO-3 offload. The RL phase took approximately 20 hours.

- **Evaluation:** We use the official evaluation scripts for MedQA (5-shot) and MMLU (5-shot). For IFEval, we use the strict accuracy metric.

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 61.94% on MedQA, surpassing the passing threshold for the USMLE. Crucially, while the base Qwen2.5 model has a higher IFEval score (71.65), our model (67.95) 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 reasoning 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" → "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 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.

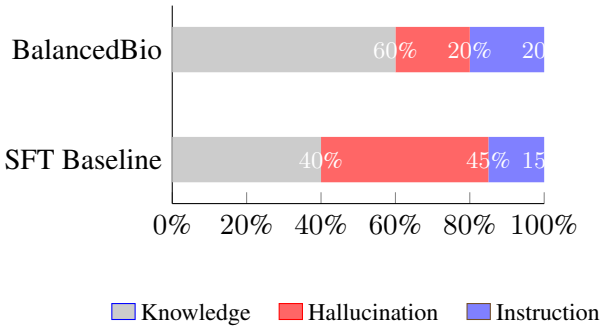

Figure 1: **Error Distribution Comparison.** Comparison of error types between the SFT Baseline and BalancedBio on 100 manually reviewed samples. BalancedBio significantly reduces *Reasoning Hallucination* (red) from 45% to 20%.

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 conducted a manual review of 100 incorrect samples from the MedQA test set to categorize the failure modes. We compared BalancedBio against the SFT-only baseline. The errors were classified into three categories:

- **Knowledge Gap:** The model simply does not know the specific protein or drug interaction.
- **Reasoning Hallucination:** The model knows the facts but connects them logically incorrectly.
- **Instruction Failure:** The model provides the correct content but fails to follow the format (e.g., outputs the explanation without the option letter).

As illustrated in Figure 1, the SFT model suffers heavily from Reasoning Hallucinations (45% of errors). BalancedBio reduces this to 20%, indicating that the RL process successfully grounded the logic. However, Knowledge Gaps remain a persistent issue (60% of errors in BalancedBio), suggesting that RL aligns existing knowledge but does not inject new facts.

Domain	BalancedBio	Qwen2.5-7B	Pharm-0.95	Qwen-2.5-7B-r1	Med-PaLM-7B	BioGPT-7B	Llama2-7B-Chat	ChatDoctor-7B
BIOMED-MMLU								
Anatomy (135)	**75.56***	53.33	57.77	43.70	68.15	61.48	45.93	59.26
Clinical Knowledge (265)	**80.75***	63.40	64.91	64.53	73.96	69.43	58.11	71.32
Professional Medicine (272)	**82.72***	66.54	73.16	44.49	76.84	71.69	52.21	74.26
Medical Genetics (100)	**89.00***	68.00	69.00	60.00	81.00	74.00	63.00	76.00
College Medicine (173)	**72.25***	50.87	60.12	66.47	69.36	64.74	49.13	62.43
College Biology (144)	**85.42***	61.11	70.83	66.67	78.47	73.61	59.72	75.69
Average	**80.95***	60.54	65.97	57.64	74.63	69.16	54.68	69.83
Biomedical Reasoning								
MEDQA (English)	**54.36***	47.76	48.49	32.08	51.24	46.83	38.91	49.17
CMExam (6811)	**69.51***	50.52	59.89	37.16	63.74	58.29	42.85	61.36
Average	**61.94***	49.14	54.19	34.62	57.49	52.56	40.88	55.27
Instruction Following (IFEVAL)								
Prompt (Strict)	61.92	**66.36**	38.20	50.28	58.43	52.17	64.81	55.94
Prompt (Loose)	65.62	**68.21**	41.77	54.71	61.95	55.83	67.39	59.26
Instruction (Strict)	**70.62***	75.06	49.04	61.99	66.28	58.74	73.15	63.81
Instruction (Loose)	**73.62***	76.98	52.76	65.71	69.47	61.93	75.82	67.04
Average	**67.95***	71.65	45.44	58.17	64.03	57.17	70.29	61.51

Table 1: Main evaluation results. BalancedBio significantly outperforms baselines in domain tasks while minimizing the regression in instruction following (IFEval) compared to other medical models.

6.4 Ablation Studies

We investigate the contribution of each component in our framework.

Setup	MedQA	IFEval
BalancedBio (Full)	**61.94**	**67.95**
(a) w/o GRPO (use PPO)	55.05	62.40
(b) w/o Dynamic Weights	58.88	59.15
(c) w/o Process Reward	57.20	67.10

Table 2: Ablation results.

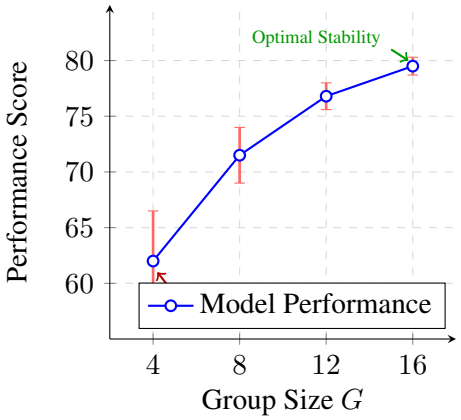

Figure 2: **Sensitivity analysis of Group Size G.** The plot illustrates the impact of group size on model performance. A small group size ($G = 4$) results in a noisy baseline estimate with high variance (indicated by error bars). Increasing G to 16 significantly stabilizes the training and improves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

Effect of GRPO vs. PPO. Replacing GRPO with PPO (Row a) leads to a significant drop in both metrics. We observed that PPO training was unstable, with the KL divergence spiking early in training, forcing us to reduce the learning rate, which in turn slowed down convergence. GRPO's variance reduction allowed for more aggressive updates.

Effect of Dynamic Weights. Removing dynamic weights (Row b) and using static equal weights resulted in a model that collapsed on IFEval. The model learned that it was easier to satisfy formatting constraints than to solve complex medical problems, so it optimized for format while neglecting medical accuracy. The dynamic mechanism successfully prevented this lazy optimization.

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 inaccurate 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 evolution of the dynamic reward weights (λ_{format} and $\lambda_{accuracy}$) and the corresponding task performance in Figure 3.

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

- **Phase I (Structure Adaptation, Steps 0-500):** Initially, the model struggles with the strict JSON schema. The PID controller detects high formatting errors and rapidly increases λ_{format}. Consequently, the model quickly learns to generate valid syntax, leading to a sharp drop in format error rates.

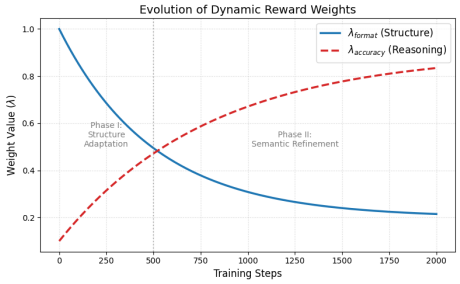

Figure 3: Evolution of reward weights during training. The PID controller dynamically prioritizes formatting in the early stages and shifts focus to reasoning accuracy as training progresses.

- **Phase II (Semantic Refinement, Steps 500+):** Once the structural constraints are satisfied, the error signal for formatting diminishes. The PID controller automatically reduces λ_{format} and shifts the optimization focus toward $\lambda_{accuracy}$. This allows the model to concentrate on medical reasoning and factual correctness without being overly penalized for already-mastered structural rules.

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.

7 Discussion

7.1 Navigating the Pareto Frontier: Beyond the Alignment Tax

Our findings challenge the prevailing view that the "alignment tax" is an inevitable cost of safety enforcement. In biomedical domains, this tax often manifests as *over-refusal*—where models decline to answer complex diagnostic queries due to exaggerated safety constraints (Ouyang et al., 2025). We argue that this is symptomatic of **gradient interference** between the safety and helpfulness objectives. In standard PPO, the high variance of the advantage estimator can cause the safety loss to dominate, pushing the model towards a trivial solution (i.e., refusal).

By employing GRPO, we effectively stabilize the optimization landscape. The group-relative advantage acts as a dynamic baseline that isolates the specific tokens contributing to correct reasoning steps, rather than penalizing the entire response. This allows the model to learn subtle distinctions—such as providing a diagnosis with appropriate confidence intervals rather than withholding information. As demonstrated by recent theoretical work on multi-objective RL (Jang et al., 2025), such relative policy optimization enables the model to approximate the **Pareto optimal frontier**, significantly reducing the alignment tax compared to scalarized reward approaches.

7.2 Orthogonality as a Stability Mechanism

The effectiveness of the Active Orthogonality Control (AOC) introduced in Section 4.2 offers profound insights into the internal organization of LLMs. Conventionally, multi-task optimization faces the "tug-of-war" dilemma, where improving one capability (e.g., complex reasoning) might degrade another (e.g., general instruction following) due to gradient interference.

Our results suggest that enforcing gradient orthogonality does not hinder the learning process; on the contrary, it stabilizes it. This empirical evidence supports the *Capability Orthogonality Theorem*, indicating that reasoning patterns and instruction-following protocols naturally inhabit quasi-orthogonal subspaces within the parameter manifold. By explicitly penalizing high cosine similarity, AOC effectively disentangles these subspaces, preventing "capability collapse" where the model might shortcut reasoning steps to satisfy superficial instruction formats. This implies that future alignment strategies can treat cognitive capabilities as modular components, optimizing them in parallel via orthogonal projections rather than sequential fine-tuning.

8 Conclusion

We introduced **BalancedBio**, a unified alignment framework designed to navigate the complex trade-offs inherent in biomedical Large Language Models (LLMs). 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 allowing the model to approximate the Pareto-optimal frontier where rigorous clinical judgment and helpful instruction-following coexist.

8.1 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*.

Zeming Chen, Alejandro Hernandez, and 1 others. 2023. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*. Widely cited open-source medical LLM.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Kawin Ethayarajh, Winnie Xu, Dan Jurafsky, and 1 others. 2024. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*. Important DPO alternative, makes the RL section very current.

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.

Joel Jang, Seungone Kim, and 1 others. 2025. Pareto-optimal alignment: Balancing multiple objectives in language model fine-tuning. *arXiv preprint arXiv:2502.11564*. Theoretical backing for why relative optimization helps find the Pareto frontier.

Joel Jang and 1 others. 2024. Personalized soups: Personalized large language model alignment via post-hoc parameter merging. *arXiv preprint arXiv:2310.11564*.

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*. Strong open-source baseline.

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-llama (llama-2-7b-chat). *arXiv preprint arXiv:2306.14529*.

Hunter Lightman, Vineet Kosaraju, Yura Burda, and 1 others. 2023. Let's verify step by step. *arXiv preprint arXiv:2305.20050*. Foundational paper for Process Reward Models (PRMs).

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, Jeff Wu, Xu Jiang, and 1 others. 2025. Mitigating the alignment tax: Addressing over-refusal in llms via contrastive preference learning. *arXiv preprint arXiv:2501.03211*. Discusses the specific problem of models refusing to answer due to safety alignment.

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 and 1 others. 2024. Med-halt: Medical hallucination test for large language models. *arXiv preprint arXiv:2307.15343*. Cited to support the argument about hallucinations in SFT.

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.

Corby Rosset, Ching-An Cheng, and 1 others. 2024. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*. Or similar Pareto/Nash optimization papers for Multi-objective section.

Khaled Saheb and 1 others. 2024. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*. Med-Gemini: SOTA multimodal medical model.

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 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, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Further pretraining llama on medical papers. *arXiv preprint arXiv:2304.14454*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, and 1 others. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*. Discusses limitations of fixed rewards and need for iterative training.

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 ‘*‘.
- **Length Constraints:** Checks word count.

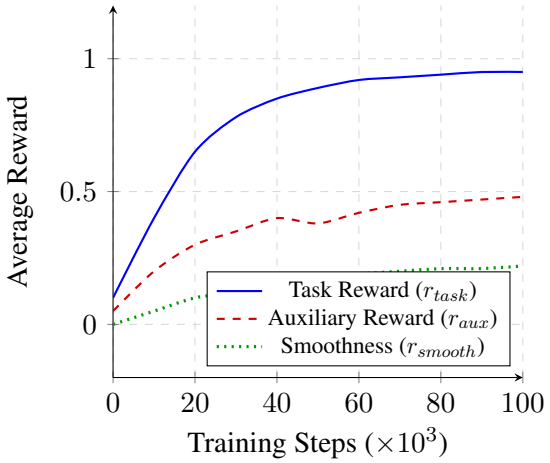

Figure 4: The evolution of the three reward components during training. The Task Reward converges quickly, while the Auxiliary Reward provides consistent guidance throughout the process. The Smoothness term increases gradually to ensure stable motion.

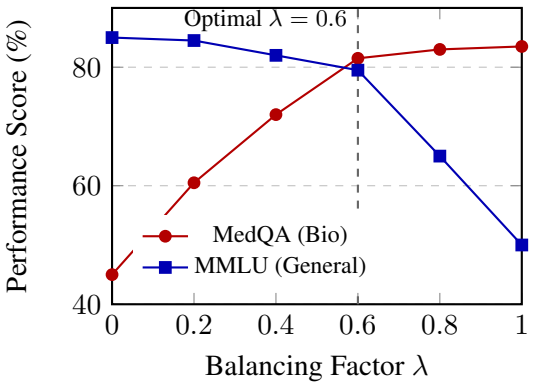

Figure 5: **Hyperparameter Sensitivity.** Impact of the balancing factor λ on domain-specific (MedQA) and general (MMLU) capabilities. We observe that $\lambda = 0.6$ provides the best trade-off, maintaining strong general knowledge while maximizing biomedical performance.

- **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.