LEADERBOARD INCENTIVES: MODEL RANKINGS UNDER STRATEGIC POST-TRAINING

Yatong Chen*, **Guanhua Zhang**, **Moritz Hardt**
Max Planck Institute for Intelligent Systems, Tübingen AI Center

ABSTRACT

Influential benchmarks incentivize competing model developers to strategically allocate post-training resources towards improvements on the leaderboard, a phenomenon dubbed *benchmaxxing* or *training on the test task*. In this work, we initiate a principled study of the incentive structure that benchmarks induce. We model benchmarking as a Stackelberg game between a benchmark designer who chooses an evaluation protocol and multiple model developers who compete simultaneously in a subgame given by the designer's choice. Each competitor has a model of unknown latent quality and can inflate its observed score by allocating resources to benchmark-specific improvements. First, we prove that current benchmarks induce games for which no Nash equilibrium between model developers exists. This result suggests one explanation for why current practice leads to misaligned incentives, prompting model developers to strategize in opaque ways. However, we prove that under mild conditions, a recently proposed evaluation protocol, called tune-before-test, induces a benchmark with a unique Nash equilibrium that ranks models by latent quality. This positive result demonstrates that benchmarks need not set bad incentives, even if current evaluations do.

1 INTRODUCTION

Traditionally, machine learning benchmarks came with a fixed training set, requiring that all models under comparison train on the same data. The situation has changed with large language model benchmarks that typically only provide test data, leaving the choice of training data to the model developer. This has raised the concern that model developers can inflate benchmark performance with benchmark-specific tweaks that don't broadly improve model capabilities. The resulting problem, called benchmaxxing or training on the test task, confounds model comparisons and may cause misleading leaderboards, as prior work shows (Dominguez-Olmedo et al., 2024; Singh et al., 2025). But the situation need not be a sign of cheating or wrongdoing. Rather, influential benchmarks incentivize competing model developers to strategically allocate post-training resources towards improvements on the leaderboard. Although widely recognized, there is currently no formal understanding of the incentives that benchmarks set.

In this work, we initiate a principled study of the incentive structure that benchmarks induce. We model benchmarking as a Stackelberg game between a benchmark designer and multiple competing model developers. The designer chooses an evaluation protocol, and the model developers compete in a simultaneous-move subgame given by the designer's choice. Each competitor has a model of a latent quality — unknown to the benchmark designer — and can inflate the observed model score by allocating additional resources to benchmark-specific improvements at a cost. The designer aims to choose an evaluation protocol so that the resulting benchmark yields a ranking by latent quality at equilibrium when competitors best respond to each other.

1.1 OUR RESULTS

Our first result is descriptive and negative: Current benchmarks induce games in which generally no Nash equilibrium between model developers exists. Model developers are always incentivized to strategize in opaque ways, leading to uninterpretable leaderboards that may not reflect a ranking

*Corresponding author: yatong.chen@tuebingen.mpg.de

by latent quality. In contrast, our second result is prescriptive and positive. Under mild conditions, the designer has a cost-effective evaluation strategy that induces a subgame with a unique Nash equilibrium that ranks models by latent quality. In addition, at this equilibrium solution, model developers refrain from benchmaxxing altogether, investing no additional effort in benchmark-specific improvements.

What makes the incentive design problem challenging is that the benchmark creator has no control over the utility each model developer has for gains on the leaderboard, does not know latent capabilities, and cannot limit how much effort model developers invest. The evaluation protocol we study tunes each model on the same small amount of task-specific data before evaluation. The intuition is that a small amount of task-specific preparation levels the playing field by washing out minor benchmark-specific tweaks. Previous work showed empirically that this intervention, called tune-before-test (TbT), leads to consistent model rankings across a wide range of benchmarks (Zhang et al., 2025). Our main result adds a game-theoretic justification: Tune-before-test creates incentives that lead to rankings by latent quality at equilibrium. Surprisingly, a small amount of data suffices to realize the same effect as orders of magnitude more training, as we show.

Complementing our theoretical results, we demonstrate empirically that the assumptions of our theorem hold in a case study of a representative benchmark. After applying tune-before-test with only 3,000 steps, a model developer would have to invest at least 384,668 additional training steps to change model rankings. This captures why TbT can be such an effective intervention: by pushing all models into a diminishing-returns regime, it greatly increases the marginal cost of further score improvements. As a result, overtaking nearby competitors requires substantially more additional effort, amplifying the asymmetry in local overtaking incentives predicted by our theory (see Figure 1 right).

1.2 Related Work

Benchmarks have been the key driver of machine learning progress by enabling frictionless comparison and competition (Donoho, 2024) between models. In the traditional supervised learning paradigm, benchmarks typically come with a fixed training set and a held-out test set (Lyons, 1993; LeCun & Cortes, 2010; Sang & Meulder, 2003; Liberman, 2015; Hardt & Recht, 2022), enabling relatively controlled comparisons across methods. Although test-set reuse (Duda & Hart, 1974) can erode classical statistical guarantees (Dwork et al., 2015a;b; Blum & Hardt, 2015; Mania et al., 2019), fixed train/test splits have historically supported relatively robust model comparison and have helped establish widely adopted baselines, exemplified by the ImageNet challenge (Deng et al., 2009; Russakovsky et al., 2014) and its role in accelerating the adoption of deep learning (Krizhevsky et al., 2012; Goodfellow et al., 2016; He et al., 2015).

A key reason benchmarks are useful in practice is *external validity* (Liao, 2021; Salaudeen et al., 2025): performance and rankings on one benchmark often correlate with performance on related datasets and tasks (Yadav & Bottou, 2019; Recht et al., 2019; Miller et al., 2020a), allowing practitioners to select strong models with some confidence. For instance, studies have shown that ImageNet rankings transfer well to other image datasets (Kornblith et al., 2018; Salaudeen & Hardt, 2024). In the large language model (LLM) era, however, many influential benchmarks provide primarily test instances and a scoring protocol (Patwardhan et al., 2025; Glazer et al., 2024; Zhou et al., 2023a; Jain et al., 2024), leaving training data and post-training choices largely unconstrained (Raffel et al., 2019; Albalak et al., 2024; Guha et al., 2025; Li et al., 2024). This change expands the space of strategic choices available to model providers and can lead to substantial ranking variation across benchmarks (Huan et al., 2025; Zhang & Hardt, 2024; Liang et al., 2023; Fourrier et al., 2024; Hardt, 2025), even among benchmarks that aim to measure similar capabilities.

One mechanism behind the ranking variation is what Dominguez-Olmedo et al. (2024) call *training on the test task*. Model providers conduct benchmark-aware post-training (Touvron et al., 2023) and data curation (Guha et al., 2025) to achieve large gains on measured tasks without comparable improvements in general capability (Zhou et al., 2023b; Singh et al., 2025). Training on the test task differs from training on the test set (Yang et al., 2023), or data contamination (Jiang et al., 2024; Yang et al., 2023; Bordt et al., 2024), where models train directly on the test data. Instead, it captures optimization to the evaluation's task distribution or protocol, such as curating instruction data that matches a benchmark's format and rubrics or tuning with feedback aligned to its scoring

procedure (Dominguez-Olmedo et al., 2024). Because such task-level alignment is difficult to detect and is often not explicitly ruled out by benchmark rules, it can confound leaderboard interpretation: scores conflate latent capability (Ruan et al., 2024) with benchmark-specific effort (Schaeffer et al., 2023). Model providers can strategically choose how much benchmark-specific effort to invest per benchmark, which yield inconsistent rankings.

A natural way to mitigate benchmark-specific advantages is to reduce heterogeneity in how models are prepared for the evaluation. Zhang et al. (2025) propose tune-before-test (TbT), in which all models are fine-tuned on benchmark-specific data before evaluation to equalize preparation. Their empirical results show that TbT can restore ranking consistency across tasks from different domains and reveal that post-TbT scores are dominated by a low-dimensional latent capability factor. While Zhang et al. motivate TbT as a post-hoc correction for observed rankings, we study it as an ex ante mechanism design choice (Manheim & Garrabrant, 2018). In our setting, a benchmark designer commits to an evaluation protocol (including a TbT baseline), anticipating that competing model providers will respond strategically by allocating additional post-training effort. We ask when an equilibrium exists in this competition and whether equilibrium rankings recover a ranking by latent capability.

Relevant to our work is the literature on strategic classification Brückner & Scheffer (2011); Hardt et al. (2016), which studies decision-making when individuals may adapt their features in response to a deployed model. See Rosenfeld (2024) for a survey. These interactions are typically modeled as a Stackelberg game, where a decision-maker commits to a classifier and strategic individuals best respond myopically to the classifier. An important question in this literature is how to incentivize genuine improvements rather than superficial *gaming* Kleinberg & Raghavan (2020); Miller et al. (2020b); Alon et al. (2020); Chen et al. (2023). Our work shares this focus on incentives. In a departure from the classification setting, however, ranking inherently creates competition between participants: when one model moves up in rank, another must move down. The leaderboard therefore induces a game between competitors. This aspect of our work also connects to recent work on markets induced by predictive systems Einav & Rosenfeld (2025); Sommer et al. (2025). Closely related is the work on *strategic ranking* Liu et al. (2022), for which we provide a detailed discussion on the differences between our paper and theirs in the appendix.

2 PRELIMINARIES

Capability, effort, and post-effort score. Consider a set of LLM model developers $N = \{1, \dots, n\}$ who submit a model to a benchmark leaderboard curated by a leaderboard designer. Each model has a latent variable $\theta_i \in \mathbb{R}_{\geq 0}$ capturing general model capabilities. This abstraction is empirically motivated by evidence that performance across diverse language-model benchmarks admits a low-dimensional structure and is often dominated by a general capability factor (Ruan et al., 2024; Zhang et al., 2025). The setting extends straightforwardly to a multi-dimensional latent capabilities vector $\vec{\theta}_i \in \mathbb{R}^d_{\geq 0}$ with a benchmark-specific coefficient vector $\mathbf{w}$ so that $\theta_i = \langle \mathbf{w}, \vec{\theta}_i \rangle$. Without loss of generality, we index models in decreasing order of latent capability, i.e., $\theta_1 > \theta_2 > \cdots > \theta_n$. Capabilities are known to model developers but not to leaderboard designer.

Before evaluation, a model may undergo additional *benchmark-specific adaptation*. We represent the *total* amount of benchmark-specific preparation by a scalar $e \in \mathbb{R}_{\geq 0}$, which may be performed by the model developer, and/or applied uniformly by the leaderboard designer (as in tune-before-test). Following the empirical tune-before-test approach of Zhang et al. (2025), we use the amount of benchmark-specific fine-tuning data (e.g., number of training examples) as a concrete and measurable proxy for this adaptation. We model benchmark performance through a *post-effort score* mapping $v = v(\theta, e)$:

Definition 2.1 (Post-Effort Score)**.** The *post-effort score* of a model with capability θ and total benchmark-specific training effort e is $v(\theta, e) : \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \to [0, 1]$.

The function $v(\theta, e)$ captures how capability and benchmark-specific effort translate into its actual benchmark performance. For simplicity, we model v deterministically; equivalently, our results can be interpreted in terms of expected scores under evaluation noise.

Leaderboard rewards. Let $\{R_j\}_{j=1}^n$ denote the reward assigned to the model ranked j, where rewards are non-increasing in rank, i.e., $R_1 \geq R_2 \geq \cdots \geq R_n \geq 0$. This means higher-ranked models receive greater downstream benefits (e.g., more downstream users, higher visibility) from their leaderboard positions.

Example 2.2. Two canonical leaderboard reward settings:

1. *Winner-take-all*: only the top-ranked model receives a reward, i.e., $R_1 > 0$ and $R_j = 0$ for all $j \neq 1$.
2. *Top-k rewards*: the top k models receive the same reward, i.e., $R_j = R > 0$ for $j \in [k]$ and $R_j = 0$ for $j > k$.

Tune-before-test methodology. We model tune-before-test (TbT) as a designer-chosen baseline effort $\Delta^{tbt} \geq 0$ that applies the same amount of benchmark-specific fine-tuning to every submitted model prior to evaluation. We treat TbT as an explicit component of the evaluation protocol: the leaderboard designer commits to Δ^{tbt}, and model developers choose any additional benchmark-specific effort in response. Setting $\Delta^{tbt} = 0$ recovers standard evaluation without TbT, so the conventional leaderboard protocol is a special case of our framework.

3 Stackelberg Ranking Game

We model leaderboard evaluation as a Stackelberg game with a single leader and multiple followers (Von Stengel & Zamir, 2010). The leader (the leaderboard designer) first commits to an evaluation protocol, after which the followers (model developers) simultaneously choose benchmark-specific effort. We study outcomes in which the developers play a Nash equilibrium of the induced follower game, and the leader chooses its action, anticipating this equilibrium response. Such an outcome is often referred to as a *Stackelberg–Nash equilibrium* (Marchesi, 2021).

Definition 3.1 (Stackelberg Ranking Game)**.** The players are a leaderboard designer (the leader) and a set of model developers $N = \{1, \ldots, n\}$ (the followers). Model developers have latent capabilities $(\theta_i)_{i=1}^n$, which are common knowledge among the model developers but unknown to the designer. A public rank-based reward $\{R_j\}_{j=1}^n$ assigns reward R_j to the model ranked j. The sequence of actions is as follows:

- The designer commits to a benchmark and a tune-before-test baseline $\Delta^{tbt} \geq 0$.
- After observing Δ^{tbt}, developers simultaneously choose additional effort levels $e_i \geq 0$.

Scores are realized as $v_i = v(\theta_i, e_i + \Delta^{tbt})$, and models are ranked by $\mathbf{v} = (v_1, \ldots, v_n)$ (ties broken deterministically). Both parties' utilities are specified in Theorem 3.2 and Theorem 3.4.

3.1 Model Developer's Utility and Externalities

Model developers are modeled as rational agents whose utility equals the rank-based reward minus the cost of additional effort. We use a cost function $c : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ to capture the effort costs [1]. Because rewards depend on rank, incentives are inherently interdependent: improving one model's rank necessarily lowers another's. As a result, a developer's optimal effort depends not only on its own capability and cost, but also on competitors' efforts. Let $\mathbf{e}_{-i}$ denote the vector of the efforts of all models except i. We define the utility of developer i as follows:

Definition 3.2 (Model Developer's Utility)**.** Given a rank-based reward $(R_j)_{j=1}^n$ and a fixed tune-before-test adjustment level $\Delta^{tbt} \geq 0$, the utility of model i with capability θ_i and effort e_i is

$$U_i(e_i; \mathbf{e}_{-i}, \Delta^{tbt}) = R_{\mathrm{rank}(v_i)} - c(e_i), \tag{1}$$

$$v_i = v(\theta_i, \Delta^{tbt} + e_i),\ \mathrm{rank}(v_i) := 1 + \sum_{j \in N \setminus \{i\}} \mathbb{I}\{v_j > v_i\}\,.$$

[1] We assume a common cost function for all model developers for simplicity. In Theorem C.1, we show that *multiplicatively separable* costs ($C_i(e_i; \theta_i) = \gamma_i\, c(e_i)$) yield an equivalent follower game after rescaling rewards.

Nash equilibrium of the induced follower game. For a fixed TbT level Δ^{tbt}, the developers' interaction forms the induced follower game, for which we use pure-strategy Nash equilibrium (Nash, 1950) as the solution concept. Intuitively, an effort profile is at equilibrium if no developer can improve its utility by unilaterally changing the amount of benchmark-specific effort it invests:

Definition 3.3 (Follower Game's Nash Equilibrium)**.** Fix a TbT level Δ^{tbt}. An effort profile $\mathbf{e}^* = (e_1^*, \cdots, e_n^*)$ is a *pure-strategy Nash equilibrium* (PNE) of the induced follower game if, for every model developer $i \in N$,

$$e_i^* \in \arg\max_{e \geq 0} U_i(e; \mathbf{e}_{-i}^*, \Delta^{tbt}),$$

where $\mathbf{e}_{-i}^*$ denotes the equilibrium efforts of other models.

Although real-world leaderboard competition involves repeated submissions over time, equilibrium analysis still provides a useful baseline: it clarifies whether incentives can ever settle, or whether the leaderboard instead induces persistent "arms-race" behavior.

3.2 Leaderboard Designer's Utility

While model developers seek to maximize rank-based rewards, the leaderboard designer evaluates a benchmark by how well its induced ranking reflects the models' latent capabilities. Because the designer only observes benchmark scores, it aims to choose an evaluation protocol that yields a ranking aligned with the capability ordering, while incurring a cost for interventions such as tune-before-test. A minimal way to capture this objective is through a *ranking correctness* criterion:

Definition 3.4. (Leaderboard Designer's Utility) Given a tune-before-test level Δ^{tbt} and a follower's effort profile $\mathbf{e} = (e_1, \cdots, e_n)$ inducing post-effort scores $v_i = v(\theta_i, e_i + \Delta^{tbt})$, the leaderboard designer's utility is:

$$U^L(\Delta^{tbt}; \mathbf{e}) = R^L \cdot \mathbb{I}[\text{Rank}(v_i) = \text{Rank}(\theta_i), \forall i] - n \cdot c^L(\Delta^{tbt}), \tag{2}$$

where $R^L \gg 0$ is the reward from achieving a capability-consistent ranking, and $c^L(\Delta^{tbt})$ is the per-model cost of applying tune-before-test level Δ^{tbt}.

This binary formulation captures the designer's core concern: whether the leaderboard correctly orders models by latent capability. One could alternatively model the designer's objective more smoothly using a rank-correlation metric such as Kendall's τ (Kendall, 1938), which rewards partial agreement between score-based and capability rankings, minus the cost of tune-before-test adjustment.

We assume that achieving a capability-consistent ranking is the designer's primary objective, in the sense that R^L is large enough that any capability-consistent ranking is preferred to any inconsistent one, even after accounting for tune-before-test costs. We also assume that $c^L(\Delta^{tbt})$ is increasing in Δ^{tbt}. Under these preferences, among all tune-before-test levels that induce a follower equilibrium preserving the capability ordering, the designer prefers the cheapest one. The leader's optimal choice then corresponds to the *Stackelberg–Nash equilibrium* defined below.

3.3 Stackelberg-Nash Equilibrium

Formally, a pair $(\Delta^{tbt\star}, \mathbf{e}^\star)$ is a *Stackelberg–Nash equilibrium* if: (i) $\mathbf{e}^\star$ is a pure-strategy Nash equilibrium of the induced follower game under $\Delta^{tbt\star}$, and (ii) $\Delta^{tbt\star}$ maximizes the designer's utility anticipating this response.

We begin by fixing Δ^{tbt} and analyzing the induced follower game. Understanding the existence of follower equilibria and the rankings they generate is a prerequisite for determining the designer's optimal policy. In Section 5, we then study how to choose the optimal TbT level $\Delta^{tbt\star}$ to maximize the leaderboard designer's utility.

4 Equilibrium Analysis of the Follower Game

What incentives does a rank-based leaderboard create for the strategic model developers? When model developers can invest in additional benchmark-specific post-training to improve performance,

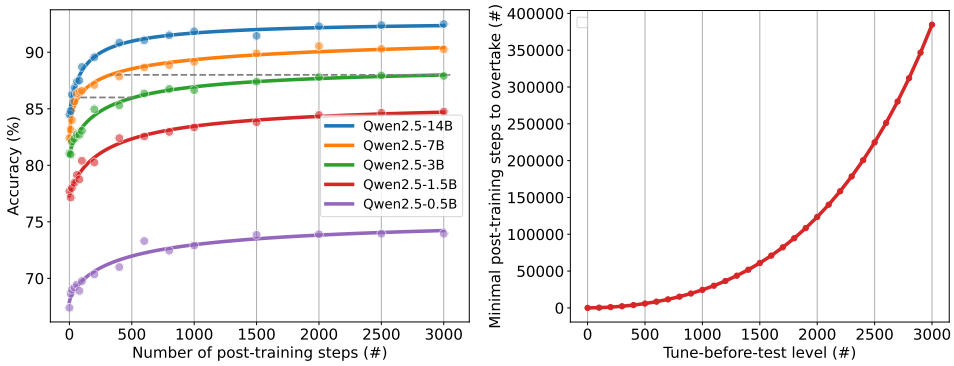

Figure 1: **Left:** Continued post-training trajectories of *Qwen2.5* models of different sizes on `Winogrande`. Here, we use model size as a proxy for the model's latent capability θ. The x-axis denotes the amount of post-training steps (each step corresponds to 8 data points), reflecting post-training effort e. The y-axis denotes accuracy on the validation set, i.e., $v(\theta, e)$. For each model, we fit a curve following Equation (3). The empirical results align with the assumptions of *monotonicity in capability*, *diminishing returns and saturation in effort*, and *non-decreasing effort gaps* in Assumption 4.2. See Appendix F for additional details and results for the other eight benchmarks. **Right:** For each tune-before-test level Δ^{tbt} (the amount of benchmark-specific finetuning steps, x-axis), we calculate the minimal additional steps required (y-axis) to change the ranking for at least one model, i.e., $\min_{r\in\{2,\dots,n\}} e_r^{\text{req}}(\Delta^{tbt})$, based on the fitted curves on the left. With $\Delta^{tbt} = 3,000$, at least 384,668 training steps are needed to change the ranking of one model.

a central question is whether the induced competition admits a pure-strategy Nash equilibrium, that is, whether incentives can settle at a stable effort profile.

Perhaps surprisingly, our first result is partially optimistic: whenever a pure-strategy equilibrium of the induced follower subgame exists, the resulting leaderboard ranking must preserve the latent capability ordering (Theorem 4.3). In this case, any strategic post-training should still lead to a stable and correct ranking. However, as we point out later, the more subtle issue is that such equilibria need not exist. We show that equilibrium existence depends critically on the *reward gaps* between adjacent ranks (Theorem 4.6). When rewards are sufficiently flat, developers face persistent incentives to "just overtake" nearby competitors, leading to arms-race dynamics.

4.1 Assumptions and Empirical Verifications

To analyze equilibrium behavior in the induced follower game, we need to make some structural assumptions on (i) the cost of benchmark-specific post-training, and (ii) how benchmark performance depends on a model's capability and post-training effort.

Initial benchmark-specific data curation may be inexpensive, but getting more data typically becomes progressively more costly (DatologyAI, 2024). To capture this, we assume that the cost of effort is non-decreasing and convex:

Assumption 4.1 (Cost Function c)**.** The cost of benchmark-specific post-training $c : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is non-decreasing and convex, with $c(0) = 0$ and $\lim_{e\to\infty} c(e) = \infty$.

Next, we impose structure on the post-effort score function:

Assumption 4.2 (Post-effort score function v)**.** Let $v : \Theta \times \mathbb{R}_{\geq 0} \to [0, 1]$ denote the post-effort score function, where $\theta \in \Theta \subseteq \mathbb{R}_{\geq 0}$ is the capability and $e \in \mathbb{R}_{\geq 0}$ is the total effort. Assume v is continuous and that for all (θ, e):

C1. (Monotonicity in capability) Holding effort fixed, higher capability yields a higher score: $\partial_\theta v(\theta, e) > 0$.

C2. (Diminishing returns and saturation in effort) Effort weakly improves performance but with diminishing marginal returns, and scores converge to a finite limit: $\partial_e v(\theta, e) \geq 0$, $\partial_{ee} v(\theta, e) \leq 0$, and $v^\infty(\theta) := \lim_{e\to\infty} v(\theta, e)$ exists and is finite.

C3. (Non-decreasing effort gaps) Let $e^{\text{req}}(s;\theta) := \inf\{e \geq 0 : \ v(\theta, e) \geq s\}$ denote the minimal effort required for capability θ to reach target score $s \in [0, 1]$. For any $\theta' > \theta$, the effort advantage of higher capability, $e^{\text{req}}(s;\theta) - e^{\text{req}}(s;\theta')$, is (weakly) nondecreasing in s.

These conditions capture three widely observed regularities in post-training scaling behavior: higher-capability models perform better at any fixed effort (C1), benchmark-specific training exhibits diminishing returns and saturation as models approach the benchmark's performance ceiling (C2). C3 further requires that as the target score increases, it does not become easier for a lower-capability model to close the gap through post-training alone. This is a standard single-crossing–type regularity condition in economic theory (Topkis, 1998), and is consistent with empirical observations that stronger foundation models tend to make more effective use of additional training, especially at a higher target score (Wei et al., 2021).

Example: generalized power-law scaling. As an illustrative functional form consistent with Assumption 4.2, consider a generalized scaling law motivated by empirical studies of post-training behavior (Ruan et al., 2024; Finnveden, 2020; Owen, 2024). Let

$$\tilde{v}(\theta, e) := \frac{v(\theta, e) - L(\theta)}{U(\theta) - L(\theta)} \in [0, 1]$$

denote a normalized score, where $L(\theta)$ and $U(\theta)$ represent model-specific lower and upper performance levels. Suppose

$$\sigma^{-1}\big(\tilde{v}(\theta, e)\big) = \alpha(\theta) + \beta(\theta)\log(1 + e), \tag{3}$$

where σ^{-1} is the logit link function, $\alpha(\theta)$ is the baseline performance on the logit scale, and $\beta(\theta) > 0$ is the scaling coefficient governing how efficiently extra compute improves performance. If $\alpha(\theta)$ and $\beta(\theta)$ are both weakly increasing in θ, this specification satisfies all conditions in Theorem 4.2 (C1–C3).

We also provide empirical evidence for Assumption 4.2 using controlled post-training experiments within a single model family. Because latent capability θ and prior benchmark-specific post-training effort are unobserved, we restrict attention to *Qwen2.5* models (Yang et al., 2024). Within this family, we use model size as a proxy for capability θ, and we treat the models' pre-existing post-training as approximately the same across sizes. We then apply additional benchmark-specific post-training to each model on the benchmark's training set, plotting performance on the validation set as a function of added effort e (measured by the incremental number of training steps).

Figure 1 (left) shows the results on `Winogrande`, a large-scale commonsense pronoun resolution benchmark introduced by Sakaguchi et al., while results for the other eight benchmarks are in Section F. For each model size, we fit the generalized power-law specification in Equation (3). The fitted curves track the observed points closely and are consistent with Theorem 4.2. First, at any fixed e, larger models achieve better performance (monotonicity in θ, C1). Second, gains from additional effort diminish and scores approach a plateau (concavity and saturation in e, C2). Third, the horizontal distance between curves—interpretable as the extra effort required for a smaller model to match a larger model's target score—does not shrink at higher target accuracies (increasing effort gaps, C3). For instance, the implied effort gap between Qwen2.5-3B and Qwen2.5-7B is larger at 88% accuracy than at 86%, as illustrated by the dashed guides in Figure 1 (left).

4.2 Order Properties of Equilibrium Profiles

Our first result shows that if the induced follower game admits a pure-strategy equilibrium $\mathbf{e}^*$, the resulting leaderboard ordering must respect the latent capability ordering (Theorem 4.3). Intuitively, this means that if all developers choose benchmark-specific post-training optimally, strategic fine-tuning alone cannot cause a lower-capability model to strictly outrank a higher-capability one:

Proposition 4.3. *Under Theorem 4.1 and Theorem 4.2, fix any tune-before-test adjustment level $\Delta^{tbt} \geq 0$. If the follower game admits a pure-Nash equilibrium $\boldsymbol{e}^*$, then for any i, j,*

$$\theta_i > \theta_j \ \Rightarrow \ v(\theta_i, \Delta^{tbt} + e_i^*) \geq v(\theta_j, \Delta^{tbt} + e_j^*).$$

In particular, post-effort scores at equilibrium preserve the latent capability ordering up to ties.

Theorem 4.3 offers reassuring news: if the leaderboard competition ever settles, it will settle in the *right* order. However, this positive result comes with an important caveat. As we show next, such equilibria need not exist at all. In particular, if the reward gap between some adjacent ranks is large relative to the cost of overtaking, namely, if there exists $r \in \{2, \cdots, n\}$ with $R_{r-1} - R_r$ sufficiently large, the induced follower game fails to admit any pure-strategy Nash equilibrium. In this regime, developers face persistent incentives to "just overtake" nearby competitors, leading to arms-race dynamics rather than convergence.

4.3 Existence of Pure-Nash Equilibrium

Fixing a TbT Δ^{tbt}, we analyze the induced follower game among developers and characterize when it admits a pure-strategy Nash equilibrium, and when it does not.

All-Zero additional effort profile. A natural reference point is the all-zero additional-effort profile $\mathbf{e} = (0, \cdots, 0)$, where model i attains the baseline post-TbT score

$$s_i(\Delta^{tbt}) = v(\theta_i, \Delta^{tbt}).$$

By monotonicity in capability (C1), we have $s_1(\Delta^{tbt}) > \cdots > s_n(\Delta^{tbt})$. Because rewards depend only on rank and effort costs are nondecreasing, any profitable deviation must be an *overtaking move*, and the cheapest such deviation is to overtake the adjacent competitor directly above:

Definition 4.4 (Just-Overtake Effort at TbT Level Δ^{tbt})**.** Fix a TbT level $\Delta^{tbt} \geq 0$. For any rank $r \in \{2, \cdots, n\}$, define the *"just-overtake "effort* as

$$e_r^{\text{req}}(\Delta^{tbt}) := \inf\Big\{e \geq 0 :\ v(\theta_r, \Delta^{tbt} + e) > s_{r-1}(\Delta^{tbt})\Big\}.$$

Equivalently, $e_r^{\text{req}}(\Delta^{tbt})$ measures how difficult it is for the model designer r to "climb" from rank r to rank $r-1$ starting from any common TbT baseline. When $e_r^{\text{req}}(\Delta^{tbt})$ is small, even minor fine-tuning can change rank, creating strong overtaking incentives. When it is large, small investments are unlikely to affect the leaderboard ordering. The following proposition precisely states when the all-zero effort profile is a pure Nash equilibrium strategy of the induced follower game:

Proposition 4.5 (Zero-effort equilibrium condition)**.** *Fix any $\Delta^{tbt} \geq 0$. The all-zero profile $\mathbf{e} = (0, \cdots, 0)$ is a PNE if and only if, for every $r \in \{2, \ldots, n\}$,*

$$c\big(e_r^{req}(\Delta^{tbt})\big) \ \geq\ R_{r-1} - R_r. \tag{4}$$

In words, with TbT level Δ^{tbt}, no agent can profitably "just overtake" the model immediately above it: the minimal cost required to surpass the adjacent baseline score exceeds the corresponding reward gap.

Theorem 4.5 further implies that if the adjacent "just-overtake" condition fails for some model designer, then no pure-strategy equilibrium exists:

Theorem 4.6 (Equilibrium structure and nonexistence)**.** *Fix any $\Delta^{tbt} \geq 0$, and suppose ties are broken deterministically in a way that does not favor the lower-capability model* [2]*. Then any PNE, if it exists, must be the all-zero profile $\mathbf{e} = (0, \cdots, 0)$. Moreover, if there exists $r \in \{2, \ldots, n\}$ such that*

$$c\big(e_r^{req}(\Delta^{tbt})\big) < R_{r-1} - R_r, \tag{5}$$

then the induced follower game admits no PNE.

The theorem shows that incentives are governed locally by the adjacent reward gap $R_{r-1} - R_r$: moving from rank r to rank $r-1$ is profitable only if this gain is large enough relative to the corresponding overtaking cost. This gives a simple interpretation of common reward schemes. In winner-take-all schemes, incentives are concentrated at the top rank; in top-k schemes, they are concentrated near the cutoff into the rewarded set. By contrast, under a smoothly decaying reward

[2] Under random tie-breaking, a lower-capability model may benefit from exerting positive effort to enter a tie lottery, so the conclusion that every PNE is all-zero need not hold without further qualification. This is mainly a technical issue: exact ties are rare in practice because leaderboard scores are usually reported at high precision or resolved by fixed secondary rules.

scheme, adjacent reward gaps are spread more evenly across ranks and may all be small. In that case, no single rank improvement creates a large reward jump, so overtaking incentives are correspondingly weaker throughout the leaderboard.

Our analysis also suggests that the main failure mode of leaderboards is not that they converge to a *stable but incorrect* ranking. Instead, leaderboards can fail because they may incentivize *continuous competitive fine-tuning*: developers are motivated to repeatedly invest in benchmark-specific optimization simply to gain (or defend) a small rank advantage (Equation (5)).

This regime is particularly relevant when benchmarks provide limited separation between models, so that even small performance gains translate into meaningful rank changes. Such situations can arise on *saturated* benchmarks where frontier models perform similarly (e.g., MMLU (Hendrycks et al., 2020) or HellaSwag (Zellers et al., 2019b)), as well as on *very difficult or early-stage* benchmarks where all models perform poorly and remain tightly clustered (e.g., Humanity's Last Exam (Phan et al., 2025)).

5 TUNE-BEFORE-TEST ALIGNS BENCHMARK INCENTIVES

We now turn to the leaderboard designer's perspective and ask: if small reward gaps can preclude equilibrium in the induced follower game, how should the designer choose Δ^{tbt} to stabilize incentives? At a high level, tune-before-test shifts the leaderboard's "operating point" by applying the same amount of benchmark-specific training to all models, so the baseline scores become $s_i(\Delta^{tbt}) = v(\theta_i, \Delta^{tbt})$. This has two conceptually distinct effects.

By moving models closer to their benchmark-specific performance limits, TbT *directly* reduces the leaderboard's sensitivity to who prepared more aggressively from scratch. In the limit as models approach their saturation levels $v^\infty(\theta) = \lim_{e\to\infty} v(\theta, e)$, rankings depend only on capability. Under monotonicity in capability (C1), saturation yields a capability-consistent ordering:

Proposition 5.1 (Rank preservation at saturation)**.** *Suppose the post-effort score $v(\theta, e)$ satisfies $\partial_\theta v(\theta, e) > 0$ (C1). Then the saturated score $v^\infty(\theta) := \lim_{e\to\infty} v(\theta, e)$ is increasing in θ.*

More importantly, TbT also weakens the incentives for strategic post-training. Since $v(\theta, e)$ exhibits diminishing returns in effort (C2), increasing Δ^{tbt} pushes all models into a regime where further improvements are harder.

5.1 TBT INCREASES LEADERBOARD CLIMBING COST

To formalize this incentive effect, recall $e_r^{\text{req}}(\Delta^{tbt})$, the minimal additional effort required for model r to overtake the adjacent competitor at baseline Δ^{tbt} (Theorem 4.4). This precisely captures the marginal difficulty of climbing the leaderboard. The next lemma shows that increasing the TbT baseline monotonically raises this overtaking cost:

Lemma 5.2 (TbT Increases the Leaderboard Climbing Cost)**.** *For any $\Delta^{tbt} \geq 0$ and $r \in \{2, \ldots, n\}$, the minimal just-overtake effort $e_r^{req}(\Delta^{tbt})$, and thus also the corresponding cost $c(e_r^{req}(\Delta^{tbt}))$, is non-decreasing in Δ^{tbt}.*

Figure 1 (right) plots the minimal just-overtake effort across all models $\min_{r\in\{2,\ldots,n\}} e_r^{\text{req}}(\Delta^{tbt})$ (y-axis), as a function of Δ^{tbt} (x-axis). We estimate $e_r^{\text{req}}(\Delta^{tbt})$ using the fitted post-training trajectories shown in the left panel. For example, for model Qwen2.5-7B at $\Delta^{tbt} = 500$, we first compute the accuracy of Qwen2.5-14B at $\Delta^{tbt} = 500$, 91.0%. We then calculate how much additional training data Qwen2.5-7B requires to reach 91.0% based on the fitted curve, which is 5,890. We repeat this for all models (except for the best model Qwen2.5-14B) and plot the minimum across all rank $r \geq 2$. The resulting curve shows that leaderboard climbing becomes rapidly more difficult as Δ^{tbt} grows. At $\Delta^{tbt} = 0$, only 18 additional steps are needed to change the ranking, whereas at $\Delta^{tbt} = 3,000$, the required effort rises to $384,668$ steps.

5.2 TBT RESTORES EQUILIBRIUM EXISTENCE

Since increasing TbT raises the effort required to overtake higher-ranked models, the all-zero equilibrium condition in Theorem 4.5 becomes easier to satisfy at higher TbT levels:

Proposition 5.3. *Given two tune-before-test adjustment levels* $0 \le \Delta_1^{tbt} \le \Delta_2^{tbt}$*, if the all-zero effort profile* $\boldsymbol{e} = (0, \dots, 0)$ *is a PNE under* Δ_1^{tbt}*, then it is also a PNE under* Δ_2^{tbt}*.*

Thus, TbT has a monotone stabilizing effect on post-training incentives: once no model developer finds it profitable to exert additional effort, this remains true under any larger TbT intervention.

5.3 How Much TbT is Enough?

In practice, a benchmark evaluator may not have enough compute to apply a large TbT adjustment to every submitted model, due to resource, latency, or cost constraints. This raises a natural question: what is the smallest TbT level that eliminates incentives for further strategic fine-tuning? This motivates us to view TbT as an incentive-control parameter with a corresponding threshold.

Definition 5.4 (Stabilizing TbT Threshold)**.** Define the *stabilizing TbT threshold* as

$$\Delta^{tbt\star} := \inf\Big\{\Delta^{tbt} \ge 0 : c\big(e_r^{\text{req}}(\Delta^{tbt})\big) \;\ge\; R_{r-1} - R_r, \forall r \in \{2, \cdots, n\}\Big\},$$

that is, the smallest TbT adjustment for which no model can profitably overtake the model directly above it at the baseline level.

At $\Delta^{tbt\star}$, the cost of overtaking the immediate neighbor weakly exceeds the corresponding reward gain for every adjacent pair. By Theorem 5.3, once this condition holds, it continues to hold for all larger TbT levels. Hence, any $\Delta^{tbt} \ge \Delta^{tbt\star}$ eliminates profitable overtaking incentives:

Corollary 5.5. *For any* $\Delta^{tbt} \ge \Delta^{tbt\star}$*, the induced follower game admits a PNE in which all model developers choose zero additional effort, i.e.,* $\boldsymbol{e}^* = (0, \dots, 0)$*. In particular, no model developer engages in strategic post-training.*

Under the leaderboard designer objective in Theorem 3.4, $\Delta^{tbt\star}$ is also the optimal TbT choice: since it is the smallest intervention that induces a follower equilibrium preserving the true capability ordering, it therefore achieves the designer's ranking objective at minimum cost.

In Section E.1, we illustrate the effectiveness of the TbT intervention under generalized power-law scaling, showing that the stabilizing threshold is controlled by the hardest adjacent pair and increases polynomially with the effective overtaking incentive.

6 Conclusion, Limitations and Future Work

We study benchmarking as a mechanism design problem and show that the resulting competition among model developers can induce persistent "just-overtake" incentives and may prevent any Nash equilibrium from existing. We demonstrate that tune-before-test acts as an effective incentive-control lever: by pushing models into a diminishing-returns regime, TbT raises the marginal cost of further benchmark-specific post-training and amplifies the effort required to overtake nearby competitors. Our results show that even a small amount of TbT can have a disproportionately large stabilizing effect, which we demonstrate both empirically and theoretically. In practice, leaderboard designers could estimate the minimal stabilizing TbT baseline $\Delta^{tbt\star}$ using historical score gaps and the observed fine-tuning effort required to change rankings.

These insights, however, rest on several simplifying assumptions. In particular, our model abstracts from many features of real benchmarking environments in order to isolate the core incentive structure induced by rank-based evaluation. A natural next step is to understand how these incentives interact with features that are common in practice, such as noisy evaluations, uncertainty about competitors, and more flexible budget or cost structures. Extending the framework in these directions would help clarify the robustness of our insights in more realistic benchmarking environments.

At the same time, TbT is not costless: it requires additional evaluation resources and may blur the distinction between a model's underlying generalization ability and its capacity to adapt during evaluation. More broadly, these trade-offs highlight that benchmark design does not merely determine how models are tested, but also shapes the incentives for how they are developed. TbT should therefore be understood as a useful but imperfect design lever, rather than a complete solution to strategic behavior in evaluation.

REFERENCES

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, Colin Raffel, Shiyu Chang, Tatsunori Hashimoto, and William Yang Wang. A survey on data selection for language models. *ArXiv*, abs/2402.16827, 2024. URL https://api.semanticscholar.org/CorpusID:268032975.

Tal Alon, Magdalen Dobson, Ariel Procaccia, Inbal Talgam-Cohen, and Jamie Tucker-Foltz. Multiagent evaluation mechanisms. *Proc. AAAI Conference on Artificial Intelligence*, 34(02):1774–1781, Apr. 2020. URL https://ojs.aaai.org/index.php/AAAI/article/view/5543.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language. In *AAAI Conference on Artificial Intelligence*, 2019. URL https://api.semanticscholar.org/CorpusID:208290939.

Avrim Blum and Moritz Hardt. The ladder: A reliable leaderboard for machine learning competitions. In *International Conference on Machine Learning*, 2015. URL https://api.semanticscholar.org/CorpusID:1493191.

Sebastian Bordt, Suraj Srinivas, Valentyn Boreiko, and Ulrike von Luxburg. How much can we forget about data contamination? *ArXiv*, abs/2410.03249, 2024. URL https://api.semanticscholar.org/CorpusID:273163321.

Michael Brückner and Tobias Scheffer. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 547–555, 2011.

Yatong Chen, Jialu Wang, and Yang Liu. Learning to incentivize improvements from strategic agents. *Transactions on Machine Learning Research*, 2023.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457, 2018. URL https://api.semanticscholar.org/CorpusID:3922816.

Karl Cobbe, Vineet Kosaraju, Mo Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168, 2021. URL https://api.semanticscholar.org/CorpusID:239998651.

DatologyAI. Technical deep-dive: Curating our way to a state-of-the-art text dataset. https://www.datologyai.com/blog/technical-deep-dive-curating-our-way-to-a-state-of-the-art-text-dataset, November 2024.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. URL https://api.semanticscholar.org/CorpusID:57246310.

Ricardo Dominguez-Olmedo, Florian E Dorner, and Moritz Hardt. Training on the test task confounds evaluation and emergence. *arXiv preprint arXiv:2407.07890*, 2024.

David Donoho. Data science at the singularity. *Harvard Data Science Review*, 6(1), 2024.

Richard O. Duda and Peter E. Hart. Pattern classification and scene analysis. In *A Wiley-Interscience publication*, 1974. URL https://api.semanticscholar.org/CorpusID:12946615.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Generalization in adaptive data analysis and holdout reuse. *ArXiv*, abs/1506.02629, 2015a. URL https://api.semanticscholar.org/CorpusID:14762349.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349:636 – 638, 2015b. URL https://api.semanticscholar.org/CorpusID:15569600.

Ohad Einav and Nir Rosenfeld. A market for accuracy: Classification under competition. *arXiv preprint arXiv:2502.18052*, 2025.

Lukas Finnveden. Extrapolating gpt-n performance, 2020. URL https://www.lesswrong.com/posts/k2SNji3jXaLGhBeYP/extrapolating-gpt-n-performance.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2024.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai. *ArXiv*, abs/2411.04872, 2024. URL https://api.semanticscholar.org/CorpusID:273877467.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Etash Kumar Guha, Ryan Marten, Sedrick Scott Keh, Negin Raoof, Georgios Smyrnis, Hritik Bansal, Marianna Nezhurina, Jean-Pierre Mercat, Trung Vu, Zayne Sprague, Ashima Suvarna, Ben Feuer, Liangyu Chen, Zaid Khan, Eric Frankel, Sachin Grover, Caroline Choi, Niklas Muennighoff, Shiye Su, Wanjia Zhao, John Yang, Shreyas Pimpalgaonkar, Kartik Sharma, Charlie Cheng-Jie Ji, Yichuan Deng, Sarah Pratt, Vivek Ramanujan, Jon Saad-Falcon, Jeffrey Li, Achal Dave, Alon Albalak, Kushal Arora, Blake Wulfe, Chinmay Hegde, Greg Durrett, Sewoong Oh, Mohit Bansal, Saadia Gabriel, Aditya Grover, Kai-Wei Chang, Vaishaal Shankar, Aaron Gokaslan, Mike A. Merrill, Tatsunori Hashimoto, Yejin Choi, Jenia Jitsev, Reinhard Heckel, Maheswaran Sathiamoorthy, Alexandros G. Dimakis, and Ludwig Schmidt. Openthoughts: Data recipes for reasoning models. *ArXiv*, abs/2506.04178, 2025. URL https://api.semanticscholar.org/CorpusID:279154475.

Moritz Hardt. The emerging science of machine learning benchmarks. Online at https://mlbenchmarks.org, 2025. Manuscript.

Moritz Hardt and Benjamin Recht. *Patterns, predictions, and actions: Foundations of machine learning*. Princeton University Press, 2022.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pp. 111–122, 2016.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015. URL https://api.semanticscholar.org/CorpusID:206594692.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021. URL https://api.semanticscholar.org/CorpusID:235458009.

Maggie Huan, Yuetai Li, Tuney Zheng, Xiaoyu Xu, Seungone Kim, Minxin Du, Radha Poovendran, Graham Neubig, and Xiang Yue. Does math reasoning improve general llm capabilities? understanding transferability of llm reasoning. *ArXiv*, abs/2507.00432, 2025. URL https://api.semanticscholar.org/CorpusID:280146966.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *ArXiv*, abs/2403.07974, 2024. URL https://api.semanticscholar.org/CorpusID:268379413.

Minhao Jiang, Ken Ziyu Liu, Ming Zhong, Rylan Schaeffer, Siru Ouyang, Jiawei Han, and Sanmi Koyejo. Investigating data contamination for pre-training language models. *ArXiv*, abs/2401.06059, 2024. URL https://api.semanticscholar.org/CorpusID:266933004.

M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, 1938. ISSN 00063444. URL http://www.jstor.org/stable/2332226.

Jon Kleinberg and Manish Raghavan. How do classifiers induce agents to invest effort strategically? *ACM Transactions on Economics and Computation (TEAC)*, 8(4):1–23, 2020.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better? *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2656–2666, 2018. URL https://api.semanticscholar.org/CorpusID:43928547.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84 – 90, 2012. URL https://api.semanticscholar.org/CorpusID:195908774.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc V. Le, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019. URL https://api.semanticscholar.org/CorpusID:86611921.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. http://yann.lecun.com/exdb/mnist/, 2010. URL http://yann.lecun.com/exdb/mnist/.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Kumar Guha, Sedrick Scott Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean-Pierre Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic, Sham M. Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke S. Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldani, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: In search of the next generation of training sets for language models. *ArXiv*, abs/2406.11794, 2024. URL https://api.semanticscholar.org/CorpusID:270560330.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher R'e, Diana Acosta-Navas, Drew A. Hudson, E. Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan S. Kim, Neel Guha, Niladri S. Chatterji, O. Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas F. Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525:140 – 146, 2023. URL https://api.semanticscholar.org/CorpusID:253553585.

Thomas Liao. Are we learning yet? a meta review of evaluation failures across machine learning. In *NeurIPS Datasets and Benchmarks*, 2021. URL https://api.semanticscholar.org/CorpusID:244907059.

Marc Liberman. Reproducible research and the common task method, 2015. URL https://www.simonsfoundation.org/lecture/reproducible-research-and-thecommon-task-method.

Lydia T Liu, Nikhil Garg, and Christian Borgs. Strategic ranking. In *International Conference on Artificial Intelligence and Statistics*, pp. 2489–2518. PMLR, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. URL https://api.semanticscholar.org/CorpusID:53592270.

John W Lyons. Darpa timit acoustic-phonetic continuous speech corpus. *National Institute of Standards and Technology*, 1993.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, Benjamin Bossan, and Marian Tietz. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

David Manheim and Scott Garrabrant. Categorizing variants of goodhart's law. *ArXiv*, abs/1803.04585, 2018. URL https://api.semanticscholar.org/CorpusID:4715794.

Horia Mania, John Miller, Ludwig Schmidt, Moritz Hardt, and Benjamin Recht. Model similarity mitigates test set overuse. *ArXiv*, abs/1905.12580, 2019. URL https://api.semanticscholar.org/CorpusID:168169971.

Alberto Marchesi. *Leadership Games: Multiple Followers, Multiple Leaders, and Perfection*, pp. 107–118. Springer International Publishing, Cham, 2021. ISBN 978-3-030-62476-7. doi: 10.1007/978-3-030-62476-7_10. URL https://doi.org/10.1007/978-3-030-62476-7_10.

John Miller, Karl Krauth, Benjamin Recht, and Ludwig Schmidt. The effect of natural distribution shift on question answering models. *ArXiv*, abs/2004.14444, 2020a. URL https://api.semanticscholar.org/CorpusID:216867120.

John Miller, Smitha Milli, and Moritz Hardt. Strategic classification is causal modeling in disguise. In *International Conference on Machine Learning*, pp. 6917–6926. PMLR, 2020b.

John F Nash. Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences*, 36(1):48–49, 1950.

David Owen. How predictable is language model benchmark performance? *arXiv preprint*, arXiv:2401.04757, 2024. URL https://arxiv.org/abs/2401.04757.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. In *ACM Conference on Health, Inference, and Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:247763070.

Tejal Patwardhan, Rachel Dias, Elizabeth Proehl, Grace Kim, Michele Wang, Olivia Watkins, Sim'on Posada Fishman, Marwan Aljubeh, Phoebe Thacker, Laurance Fauconnet, Natalie S. Kim, Patrick Chao, Samuel Miserendino, Gildas Chabot, David Li, Michael Sharman, Alexandra Barr, Amelia Glaese, and Jerry Tworek. Gdpval: Evaluating ai model performance on real-world economically valuable tasks. *ArXiv*, abs/2510.04374, 2025. URL https://api.semanticscholar.org/CorpusID:281843768.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2019. URL https://api.semanticscholar.org/CorpusID:204838007.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 2019. URL https://api.semanticscholar.org/CorpusID:67855879.

Nir Rosenfeld. Strategic ml: How to learn with data that "behaves". In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 1128–1131. ACM, 2024.

Yangjun Ruan, Chris J Maddison, and Tatsunori B Hashimoto. Observational scaling laws and the predictability of langauge model performance. *Advances in Neural Information Processing Systems*, 37:15841–15892, 2024.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115:211 – 252, 2014. URL https://api.semanticscholar.org/CorpusID:2930547.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. Winogrande: An adversarial winograd schema challenge at scale, 2019. URL https://arxiv.org/abs/1907.10641.

Olawale Salaudeen and Moritz Hardt. Imagenot: A contrast with imagenet preserves model rankings. *ArXiv*, abs/2404.02112, 2024. URL https://api.semanticscholar.org/CorpusID:268857319.

Olawale Salaudeen, Anka Reuel, Ahmed M. Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Ben Domingue, Angelina Wang, and Oluwasanmi Koyejo. Measurement to meaning: A validity-centered framework for ai evaluation. *ArXiv*, abs/2505.10573, 2025. URL https://api.semanticscholar.org/CorpusID:278715024.

E. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Conference on Computational Natural Language Learning*, 2003. URL https://api.semanticscholar.org/CorpusID:2470716.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. Socialiqa: Commonsense reasoning about social interactions, 2019. URL https://arxiv.org/abs/1904.09728.

Rylan Schaeffer, Brando Miranda, and Oluwasanmi Koyejo. Are emergent abilities of large language models a mirage? *ArXiv*, abs/2304.15004, 2023. URL https://api.semanticscholar.org/CorpusID:258418299.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'Souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A Smith, et al. The leaderboard illusion. *arXiv preprint arXiv:2504.20879*, 2025.

Yonatan Sommer, Ivri Hikri, Lotan Amit, and Nir Rosenfeld. Learning classifiers that induce markets. In *Proceedings of the 42nd International Conference on Machine Learning (ICML 2025)*, 2025.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *ArXiv*, abs/1811.00937, 2019. URL https://api.semanticscholar.org/CorpusID:53296520.

Donald M. Topkis. *Supermodularity and Complementarity*. Princeton University Press, 1998. ISBN 9780691032443. URL http://www.jstor.org/stable/j.ctt7s83q.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Niko lay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melissa Hall Melanie Kambadur, Sharan Narang, Aur'elien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288, 2023. URL https://api.semanticscholar.org/CorpusID:259950998.

Bernhard Von Stengel and Shmuel Zamir. Leadership games with convex strategy sets. *Games and Economic Behavior*, 69(2):446–457, 2010.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

Chhavi Yadav and Léon Bottou. Cold case: The lost mnist digits. In *Neural Information Processing Systems*, 2019. URL https://api.semanticscholar.org/CorpusID:166227957.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhenru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024. URL https://api.semanticscholar.org/CorpusID:274859421.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph Gonzalez, and Ion Stoica. Rethinking benchmark and contamination for language models with rephrased samples. *ArXiv*, abs/2311.04850, 2023. URL https://api.semanticscholar.org/CorpusID:265050721.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Annual Meeting of the Association for Computational Linguistics*, 2019a. URL https://api.semanticscholar.org/CorpusID:159041722.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019b.

Guanhua Zhang and Moritz Hardt. Inherent trade-offs between diversity and stability in multi-task benchmarks. *ArXiv*, abs/2405.01719, 2024. URL https://api.semanticscholar.org/CorpusID:269587753.

Guanhua Zhang, Ricardo Dominguez-Olmedo, and Moritz Hardt. Train-before-test harmonizes language model rankings. *arXiv preprint arXiv:2507.05195*, 2025.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models. *ArXiv*, abs/2311.07911, 2023a. URL https://api.semanticscholar.org/CorpusID:265157752.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. Don't make your llm an evaluation benchmark cheater. *ArXiv*, abs/2311.01964, 2023b. URL https://api.semanticscholar.org/CorpusID:265019021.

A APPENDIX

B PRIMARY NOTATION TABLE

Symbol	Description
$\theta \in \mathbb{R}_{\geq 0}$	one-dimensional latent model capability
$e \in \mathbb{R}_{\geq 0}$	benchmark-specific training effort
$v(\theta, e) \in [0, 1]$	post-effort score (benchmark performance)
$c(e)$	cost of exerting effort e
$(R_j)_{j=1}^n$	rank-based reward scheme: the model ranked j receives reward R_j
$\Delta^{tbt} \in \mathbb{R}_{\geq 0}$	tune-before-test (TbT) baseline chosen by the leaderboard designer
$s_r(\Delta^{tbt}) := v(\theta_r, \Delta^{tbt})$	baseline score of model r under TbT level Δ^{tbt}
$e^{\text{req}}(s; \theta)$	minimal effort required for a model with capability θ to reach score s, i.e., $e^{\text{req}}(s;\theta) := \inf\{e \geq 0 :\ v(\theta, e) \geq s\}$
$e_r^{\text{req}}(\Delta^{tbt})$	minimal additional effort required for model r to overtake model $r-1$ at TbT level Δ^{tbt}, i.e., $e_r^{\text{req}}(\Delta^{tbt}) := \inf\{e \geq 0 :\ v(\theta_r, \Delta^{tbt} + e) > s_{r-1}(\Delta^{tbt})\}$
$\sigma : \mathbb{R} \to [0, 1]$	logit link function
$L(\theta)$	lower attainable performance level for a model with capability θ
$U(\theta)$	upper attainable performance level for a model with capability θ
$\tilde{v}(\theta, e) := \frac{v(\theta,e)-L(\theta)}{U(\theta)-L(\theta)} \in [0, 1]$	normalized post-effort score
$\alpha(\theta)$	logit of the baseline performance when $e = 0$
$\beta(\theta) > 0$	coefficient governing how effectively extra effort translates into performance
$\kappa > 0$	constant lower bound on the marginal cost of effort, i.e., $c(e) \geq \kappa e$
$\rho_r := (R_{r-1} - R_r)/\kappa$	effective reward gap (in effort units) for model r
$\lambda_r := \rho_r / e_r^{\text{req}}(0)$	effective incentive parameter
$\gamma_r := \beta(\theta_r)/\beta(\theta_{r-1}) \leq 1$	learning-rate ratio for the pair $(r-1, r)$
$\Delta_r^{tbt^\star}$	stabilizing TbT threshold for the adjacent pair $(r-1, r)$
$\Delta^{tbt^\star} := \max_{r \in \{2,\ldots,n\}} \Delta_r^{tbt^\star}$	global stabilizing TbT threshold

Table 1: Primary notation.

C MISSING PROOF IN SECTION 3

Equivalence of heterogeneous separable costs to a homogeneous same-family cost

Proposition C.1 (Equivalence of heterogeneous separable costs to a homogeneous same-family cost)**.** *Fix a common cost shape $c : \mathbb{R}_+ \to \mathbb{R}_+$ that is continuous, strictly increasing, and satisfies $c(0) = 0$. Consider each model i chooses effort $e_i \in \mathbb{R}_+$, is ranked by a score $v(\theta_i, e_i)$, and has a heterogeneous but multiplicatively separable cost*

$$C_i(e_i; \theta_i) = \gamma_i \, c(e_i) \qquad \textit{with } \gamma_i > 0.$$

Then there exists an equivalent *game in which every model developer chooses $z_i \in \mathbb{R}_+$, all model developers share the* same *cost function of the* same form *$c(\cdot)$ (i.e., homogeneous cost $c(z_i)$), and the only change is a type-dependent relabeling of the score:*

$$\hat{v}(\theta_i, z_i) \;:=\; v\Big(\theta_i,\; c^{-1}\Big(\frac{c(z_i)}{\gamma_i}\Big)\Big).$$

Specifically, the map

$$\Phi_i : \; e_i \;\Leftrightarrow\; z_i := c^{-1}\big(\gamma_i \, c(e_i)\big)$$

is a bijection for each i, and for every profile $(e_j)_j$ with image $(z_j)_j$ we have

$$R_{\mathrm{rank}(v(\theta_i, e_i))} \;-\; \gamma_i \, c(e_i) \;=\; R_{\mathrm{rank}(\hat{v}(\theta_i, z_i))} \;-\; c(z_i) \quad \textit{for all } i.$$

Consequently, the two games induce identical rankings, payoffs, best responses, and Nash equilibria up to the one-to-one reparametrization.

Proof. Because c is strictly increasing, c^{-1} exists and is strictly increasing. For each i, define the type-wise bijection $\Phi_i(e) := c^{-1}(\gamma_i \, c(e))$, with inverse

$$\Phi_i^{-1}(z) \;=\; c^{-1}\Big(\frac{c(z)}{\gamma_i}\Big).$$

Under Φ_i, the heterogeneous cost transforms as

$$\gamma_i \, c(e_i) \;=\; c\big(\Phi_i(e_i)\big) \;=\; c(z_i),$$

so all model developers share the same cost function $c(\cdot)$ in the z-parametrization. Define $\hat{v}(\theta_i, z_i) \;:=\; v\big(\theta_i, \Phi_i^{-1}(z_i)\big) \;=\; v\big(\theta_i, c^{-1}(c(z_i)/\gamma_i)\big)$. Then for any profile $(e_j)_j$ with the reparametrization variable $(z_j)_j$, we have:

$$R_{\mathrm{Rank}(v(\theta_i, e_i))} - \gamma_i c(e_i) \;=\; R_{\mathrm{Rank}\left(v(\theta_i, \Phi_i^{-1}(z_i))\right)} - c(z_i) \;=\; R_{\mathrm{Rank}\left(\hat{v}(\theta_i, z_i)\right)} - c(z_i),$$

so the utilities are preserved. Hence, the best responses and equilibria correspond one-to-one via Φ, establishing the equivalence between the two settings. □

D PROOFS FOR SECTION 4

Proof for Theorem 4.3

Proof. Fix any $\Delta^{tbt} \geq 0$ and let $\mathbf{e}^*$ be a PNE of the induced follower game. Suppose for contradiction that there exist models i, j such that $\theta_i > \theta_j$ but

$$v_i^* \;:=\; v(\theta_i, \Delta^{tbt} + e_i^*) \;<\; v(\theta_j, \Delta^{tbt} + e_j^*) \;=:\; v_j^*. \tag{6}$$

Counterfactual efforts. For any target score $s \in [0, 1]$, recall the minimal additional effort (beyond Δ^{tbt}) needed for capability θ to reach s is:

$$e^{\mathrm{req}}(s; \theta, \Delta^{tbt}) := \inf\{e \geq 0 : \; v(\theta, \Delta^{tbt} + e) \geq s\}.$$

Define the counterfactual efforts

$$\tilde{e}_i := e^{\mathrm{req}}(v_j^*; \theta_i, \Delta^{tbt}), \qquad \tilde{e}_j := e^{\mathrm{req}}(v_i^*; \theta_j, \Delta^{tbt}).$$

Then by definition,

$$v(\theta_i, \Delta^{tbt} + \tilde{e}_i) \geq v_j^*, \qquad v(\theta_j, \Delta^{tbt} + \tilde{e}_j) \geq v_i^*.$$

Moreover, since $v_j^* > v_i^*$ and $v(\theta, \cdot)$ is nondecreasing in effort (C2), we have $\tilde{e}_i > e_i^*$ and $\tilde{e}_j \leq e_j^*$.

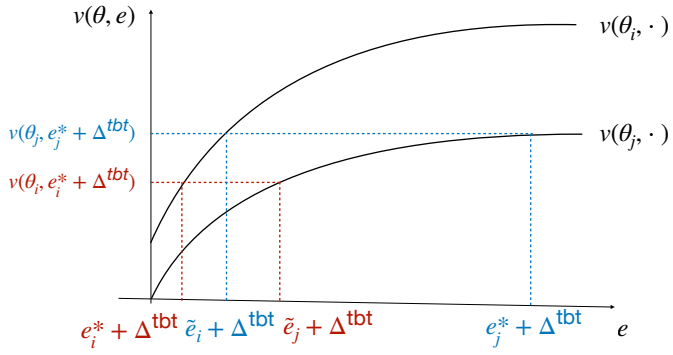

Figure 2: Illustration of the notations and functions used in the proof of Theorem 4.3. The post-effort score function $v(\theta, e)$ satisfies the conditions in Theorem 4.2. Here, e_i^* and e_j^* denote the equilibrium efforts of model developers i and j, respectively. The counterfactual efforts $\tilde{e}_i$ and $\tilde{e}_j$ are defined as the efforts each model developer would need to match the other model developer's post-effort score given a TbT effort level Δ^{tbt}, i.e., $v(\theta_i, \tilde{e}_i + \Delta^{tbt}) = v(\theta_j, e_j^* + \Delta^{tbt}), v(\theta_j, \tilde{e}_j + \Delta^{tbt}) = v(\theta_i, e_i^* + \Delta^{tbt})$.

Effort-gap comparison. Let

$$\Delta_{\theta_i,\theta_j}(s) := e^{\text{req}}(s; \theta_j, \Delta^{tbt}) - e^{\text{req}}(s; \theta_i, \Delta^{tbt})$$

denote the effort gap required for the lower-capability model j to achieve score s relative to the higher-capability model i (at TbT level Δ^{tbt}). By (C3), $\Delta_{\theta_i,\theta_j}(s)$ is (weakly) nondecreasing in s. Applying this with $s = v_j^* > v_i^*$ gives

$$\Delta_{\theta_i,\theta_j}(v_j^*) \;\geq\; \Delta_{\theta_i,\theta_j}(v_i^*),$$

i.e.,

$$e^{\text{req}}(v_j^*; \theta_j, \Delta^{tbt}) - e^{\text{req}}(v_j^*; \theta_i, \Delta^{tbt}) \;\geq\; e^{\text{req}}(v_i^*; \theta_j, \Delta^{tbt}) - e^{\text{req}}(v_i^*; \theta_i, \Delta^{tbt}).$$

Using $e^{\text{req}}(v_i^*; \theta_i, \Delta^{tbt}) = e_i^*$ and $e^{\text{req}}(v_j^*; \theta_j, \Delta^{tbt}) = e_j^*$, and the definitions $\tilde{e}_i := e^{\text{req}}(v_j^*; \theta_i, \Delta^{tbt})$ and $\tilde{e}_j := e^{\text{req}}(v_i^*; \theta_j, \Delta^{tbt})$, this becomes

$$\tilde{e}_i - e_i^* \;\leq\; e_j^* - \tilde{e}_j. \tag{7}$$

By convexity and monotonicity of c (Theorem 4.1), Equation (7) implies

$$c(\tilde{e}_i) - c(e_i^*) \;\leq\; c(e_j^*) - c(\tilde{e}_j). \tag{8}$$

Nash inequalities. Since $\mathbf{e}^*$ is a PNE, neither model can gain by deviating unilaterally. Let $\mathbf{v}^* = (v_1^*, \ldots, v_n^*)$ be the equilibrium score profile and let $r_k(\cdot)$ denote the rank of model k under a given score profile (with an arbitrary but fixed deterministic tie-breaking rule). Then

$$R_{r_i(\mathbf{v}^*)} - c(e_i^*) \geq R_{r_i\left(v(\theta_i, \Delta^{tbt} + \tilde{e}_i), \mathbf{v}_{-i}^*\right)} - c(\tilde{e}_i),$$
$$R_{r_j(\mathbf{v}^*)} - c(e_j^*) \geq R_{r_j\left(v(\theta_j, \Delta^{tbt} + \tilde{e}_j), \mathbf{v}_{-j}^*\right)} - c(\tilde{e}_j).$$

Rearranging terms gives

$$R_{r_i\left(v(\theta_i, \Delta^{tbt} + \tilde{e}_i), \mathbf{v}_{-i}^*\right)} - R_{r_i(\mathbf{v}^*)} \leq c(\tilde{e}_i) - c(e_i^*), \tag{9}$$
$$R_{r_j(\mathbf{v}^*)} - R_{r_j\left(v(\theta_j, \Delta^{tbt} + \tilde{e}_j), \mathbf{v}_{-j}^*\right)} \geq c(e_j^*) - c(\tilde{e}_j). \tag{10}$$

Here Equation (9) is the reward increase when i raises its effort to (at least) match j's score v_j^*, and Equation (10) is the reward decrease when j lowers its effort to (at least) match i's score v_i^*.

Reward comparison via strict overtaking. Fix any $\varepsilon > 0$, and define the ε-overtake effort for model i:

$$\tilde{e}_i^{\varepsilon} := \inf\{e \geq 0 : \ v(\theta_i, \Delta^{tbt} + e) \geq v_j^* + \varepsilon\}.$$

Then $v(\theta_i, \Delta^{tbt} + \tilde{e}_i^{\varepsilon}) > v_j^*$, so model i strictly outranks model j under *any* deterministic tie-breaking rule. Hence the reward increase of model i from deviating to $\tilde{e}_i^{\varepsilon}$ is at least the reward decrease of model j when it deviates to achieve v_i^*:

$$R_{r_i\left(v(\theta_i, \Delta^{tbt} + \tilde{e}_i^{\varepsilon}), \mathbf{v}^*_{-i}\right)} - R_{r_i(\mathbf{v}^*)} \ \geq \ R_{r_j(\mathbf{v}^*)} - R_{r_j\left(v(\theta_j, \Delta^{tbt} + \tilde{e}_j), \mathbf{v}^*_{-j}\right)}. \tag{11}$$

Combining equation 9, equation 10, and equation 11 yields

$$c(\tilde{e}_i^{\varepsilon}) - c(e_i^*) \ \geq \ c(e_j^*) - c(\tilde{e}_j).$$

Letting $\varepsilon \downarrow 0$ and using continuity of $v(\theta, \cdot)$ and $c(\cdot)$ gives

$$c(\tilde{e}_i) - c(e_i^*) \ \geq \ c(e_j^*) - c(\tilde{e}_j).$$

Together with equation 8, we must have

$$c(\tilde{e}_i) - c(e_i^*) = c(e_j^*) - c(\tilde{e}_j).$$

However, since $v_i^* < v_j^*$ by assumption, model i can strictly improve its rank by exerting effort slightly above $\tilde{e}_i$ (i.e., achieving $v_j^* + \varepsilon$), which yields a *strictly* higher reward whenever rewards are strictly decreasing in rank. This would force a strict inequality in the Nash bound for i, contradicting equality.

Therefore Equation (6) is impossible, and for any $\theta_i > \theta_j$ we must have

$$v(\theta_i, \Delta^{tbt} + e_i^*) \ \geq \ v(\theta_j, \Delta^{tbt} + e_j^*).$$

In particular, equilibrium scores are capability-consistent (up to ties).

□

Proof for Theorem 4.5

Proof. Fix any tune-before-test adjustment level $\Delta^{tbt} \geq 0$. At the all-zero *additional-effort* profile $\mathbf{e} = (0, \ldots, 0)$, the realized scores are $s_i(\Delta^{tbt}) = v(\theta_i, \Delta^{tbt})$ for $i = 1, \ldots, n$. Since models are indexed so that $\theta_1 > \theta_2 > \cdots > \theta_n$ and $v(\theta, e)$ is increasing in θ (C1), we have $s_1(\Delta^{tbt}) > s_2(\Delta^{tbt}) > \cdots > s_n(\Delta^{tbt})$. Hence the payoff of model k at this profile is $U_k(0; \mathbf{0}_{-k}, \Delta^{tbt}) = R_k$ for $k = 1, \ldots, n$.

Fix a model $r \in \{2, \ldots, n\}$. Consider any unilateral deviation $e'_r > 0$ by model r. Because rewards depend only on ranks and c is nondecreasing, among all deviations that improve model r's rank, the cheapest such deviation is to *just overtake* some model currently above it, i.e., to move from rank r to some rank $k < r$ by achieving a score strictly greater than $s_k(\Delta^{tbt})$. Define the minimal additional effort needed for model r to strictly beat score $s_k(\Delta^{tbt})$ as

$$e^{\text{req}}\big(s_k(\Delta^{tbt}); \theta_r, \Delta^{tbt}\big) := \inf\{e \geq 0 : \ v(\theta_r, \Delta^{tbt} + e) > s_k(\Delta^{tbt})\}.$$

Such a deviation yields utility

$$R_k - c\big(e^{\text{req}}\big(s_k(\Delta^{tbt}); \theta_r, \Delta^{tbt}\big)\big).$$

Therefore, $e_r = 0$ is a best response for model r at the all-zero profile if and only if

$$R_r \ \geq \ R_k - c\big(e^{\text{req}}\big(s_k(\Delta^{tbt}); \theta_r, \Delta^{tbt}\big)\big) \qquad \text{for all } k < r,$$

equivalently,

$$c\big(e^{\text{req}}\big(s_k(\Delta^{tbt}); \theta_r, \Delta^{tbt}\big)\big) \ \geq \ R_k - R_r \qquad \text{for all } k < r.$$

and we shorthand $e_r^{\text{req}}(\Delta^{tbt}) := e^{\text{req}}\big(s_{r-1}(\Delta^{tbt}); \theta_r, \Delta^{tbt}\big)$.

Finally, since overtaking a higher-ranked model requires weakly more effort than overtaking the adjacent model $r-1$, it suffices to check the adjacent deviation $k = r-1$, yielding Equation (4). □

Proof for Theorem 4.6

Proof. We first show that if a PNE exists, it must be the all-zero additional-effort profile; we then show that under the stated condition, the all-zero profile cannot be a PNE.

Fix $\Delta^{tbt} \geq 0$ and suppose $\mathbf{e}^*$ is a PNE. By Theorem 4.3, post-effort scores are weakly ordered by capability at equilibrium: if $\theta_i > \theta_j$, then $v(\theta_i, \Delta^{tbt} + e_i^*) \geq v(\theta_j, \Delta^{tbt} + e_j^*)$. Under deterministic tie-breaking that favors higher capability, this implies that model n (the lowest capability θ_n) is last-ranked at effort level e^*. Since at any effort profile, the last-ranked model receives reward R_n regardless of its own effort, thus model n can weakly improve its utility by reducing its effort to zero: its reward cannot decrease, while its cost weakly decreases, since c is nondecreasing and $c(0) = 0$. Therefore, in any PNE, we must have $e_n^* = 0$.

Now consider model $n-1$. Given $e_n^* = 0$, if model $n-1$ sets effort to zero as well, its score becomes $v(\theta_{n-1}, \Delta^{tbt})$, which is weakly higher than $v(\theta_n, \Delta^{tbt})$ by the assumption that $v(\theta, e)$ is monotonic in capability (C1). Under the same tie-breaking rule, model $n-1$ cannot fall below model n by choosing zero effort, and its cost decreases. Hence $e_{n-1}^* = 0$ as well. Proceeding inductively, we obtain $e_i^* = 0$ for all i, so any PNE must be $\mathbf{e}^* = (0, \ldots, 0)$.

Next, we show that under the stated condition eq. (5), all-zero is not a PNE. At $\mathbf{e} = (0, \ldots, 0)$, baseline scores are $s_i(\Delta^{tbt}) = v(\theta_i, \Delta^{tbt})$. If for some $r \in \{2, \ldots, n\}$,

$$c\Big(e_r^{\text{req}}\big(\Delta^{tbt}\big)\Big) < R_{r-1} - R_r,$$

then model r has a profitable deviation, namely that it can exert effort $\hat{e} = e_r^{\text{req}}\big(\Delta^{tbt}\big)$, so that $v(\theta_r, \Delta^{tbt} + \hat{e}) > s_{r-1}(\Delta^{tbt})$, thereby overtaking model $r-1$ and improving its reward from R_r to at least R_{r-1}. The resulting utility gain is at least $R_{r-1} - R_r$ while the incurred cost is $c(\hat{e})$, which is strictly smaller by assumption. Therefore, the deviation is profitable, so the all-zero profile is not a PNE. As a result, no PNE exists in this case. □

E PROOF IN SECTION 5

Proof for Theorem 5.1

Proof. Fix any pair (i, j) with $\theta_i > \theta_j$ and define the score gap at effort level $e \geq 0$ as $d_{ij}(e) := v(\theta_i, e) - v(\theta_j, e)$. By C1 ($\partial_\theta v > 0$), we have $d_{ij}(e) > 0$ for all e. Hence,

$$v^\infty(\theta_i) - v^\infty(\theta_j) = \lim_{e \to \infty} \big(v(\theta_i, e) - v(\theta_j, e)\big) = \lim_{e \to \infty} d_{ij}(e) > 0,$$

so $v^\infty(\theta)$ is increasing in θ. □

Proof for Theorem 5.2

Proof. Fix $r \in \{2, \ldots, n\}$ with $\theta_{r-1} > \theta_r$, recall the baseline score for model $r-1$ with TbT adjustment level Δ^{tbt} is

$$s_{r-1}(\Delta^{tbt}) \;:=\; v(\theta_{r-1}, \Delta^{tbt}).$$

By C2, $s_{r-1}(\Delta^{tbt})$ is non-decreasing in Δ^{tbt}.

Let $e^{\text{req}}(s; \theta) := \inf\{e \geq 0 :\; v(\theta, e) \geq s\}$ be the minimal *total* effort to reach score s. Define the (weak) catch-up effort for model r at baseline Δ^{tbt} by

$$\delta_r(\Delta^{tbt}) \;:=\; e^{\text{req}}\big(s_{r-1}(\Delta^{tbt}); \theta_r\big) \;-\; \Delta^{tbt}\,.$$

Since $s_{r-1}(\Delta^{tbt})$ is achieved by capability θ_{r-1} at effort Δ^{tbt}, we also have $e^{\text{req}}(s_{r-1}(\Delta^{tbt}); \theta_{r-1}) = \Delta^{tbt}$, hence

$$\delta_r(\Delta^{tbt}) = e^{\text{req}}\big(s_{r-1}(\Delta^{tbt}); \theta_r\big) - e^{\text{req}}\big(s_{r-1}(\Delta^{tbt}); \theta_{r-1}\big).$$

By (C3), the effort gap $e^{\text{req}}(s; \theta_r) - e^{\text{req}}(s; \theta_{r-1})$ is nondecreasing in S; composing with the nondecreasing map $\Delta^{tbt} \mapsto s_{r-1}(\Delta^{tbt})$ implies $\delta_r(\Delta^{tbt})$ is nondecreasing in Δ^{tbt}.

Finally, the strict "just-overtake" effort $e_r^{\text{req}}(\Delta^{tbt})$ is obtained by requiring $v(\theta_r, \Delta^{tbt}+e) > s_{r-1}(\Delta^{tbt})$ instead of $\geq$; by continuity of $v(\theta_r, \cdot)$ this differs only by an arbitrarily small $\varepsilon > 0$, so $e_r^{\text{req}}(\Delta^{tbt})$ is also nondecreasing in Δ^{tbt}.

Since c is nondecreasing, $c(e_r^{\text{req}}(\Delta^{tbt}))$ is nondecreasing as well. □

Proof for Theorem 5.3

Proof. Given $0 \leq {\Delta^{tbt}}_1 < {\Delta^{tbt}}_2$, and assume the all-zero additional-effort profile $\mathbf{e} = (0, \ldots, 0)$ is a PNE of the induced follower game under baseline ${\Delta^{tbt}}_1$, we know from Theorem 4.5 that $\forall r \in \{2, \cdots n\}$, we have

$$c(e_r^{\text{req}}({\Delta^{tbt}}_1)) \geq R_{r-1} - R_r.$$

Then according to Theorem 5.2, the cost of "just-overtake" effort $c(e_r^{\text{req}}(\Delta^{tbt}))$ is monotonically increasing in Δ^{tbt}, which means:

$$c(e_r^{\text{req}}({\Delta^{tbt}}_2)) \geq c(e_r^{\text{req}}({\Delta^{tbt}}_1)) \geq R_{r-1} - R_r,$$

which implies that $\mathbf{e} = (0, \cdots, 0)$ is also a PNE under ${\Delta^{tbt}}_2$. □

Proof for Theorem 5.5

Proof. By definition of $\Delta^{tbt\star}$, for any $\Delta^{tbt} \geq \Delta^{tbt\star}$ we have $c(e_r^{\text{req}}(\Delta^{tbt})) \geq R_{r-1} - R_r$ for $r = \{2, \ldots, n\}$, so the all-zero profile is a PNE by Theorem 4.5. □

E.1 Case Study Under Generalized Power-Law

To build intuition for the scale of the stabilizing TbT level $\Delta^{tbt\star}$, we consider the generalized scaling law introduced above:

$$\sigma^{-1}\big(\tilde{v}(\theta, e)\big) = \alpha(\theta) + \beta(\theta)\log(1+e),$$

and assume effort costs satisfy $c(e) \geq \kappa e$ for some constant $\kappa > 0$, so that effort is costly at least linearly. We focus on the regime in which overtaking remains feasible. For an adjacent pair $(r-1, r)$, define $\rho_r := \frac{R_{r-1}-R_r}{\kappa}$, which expresses the reward gap in units of effort. Let $e_r^{\text{req}}(0)$ denote the minimal effort required for model r to catch up to model $r-1$ when $\Delta^{tbt} = 0$. We then define

$$\lambda_r := \frac{\rho_r}{e_r^{\text{req}}(0)},$$

which measures the reward gap relative to the baseline catch-up difficulty.

The next proposition shows that, under this scaling law, the stabilizing TbT level for pair $(r-1, r)$ grows polynomially in λ_r, with exponent $\gamma_r := \frac{\beta(\theta_r)}{\beta(\theta_{r-1})} \leq 1$ interpreted as a relative learning-rate ratio between the two models.

Proposition E.1 (Stabilizing TbT under generalized scaling)**.** *Under the generalized scaling law, and the cost satisfies $c(e) \geq \kappa e$, there exists a stabilizing TbT threshold ${\Delta^{tbt\star}}_r$ such that any $\Delta^{tbt} \geq {\Delta^{tbt\star}}_r$ eliminates profitable overtaking deviations from rank r to rank $r-1$. Moreover, in the regime where catch-up remains feasible,*

$$\Delta_r^{tbt^\star} = O\big(\lambda_r^{\gamma_r}\big),$$

up to a constant factor depending only on $e_r^{req}(0)$.

The exponent γ_r admits a natural interpretation as a relative learning-rate ratio. When γ_r is small, the stronger model benefits more from additional effort than the weaker one, so the ranking becomes easier to stabilize. Conversely, when $\gamma_r \approx 1$, the two models benefit similarly from effort, and a larger TbT baseline is needed to deter overtaking. Thus, the required TbT level grows polynomially in the effective incentive λ_r, and the growth is slower when the effort-response gap between the two models is larger.

Finally, to stabilize the full leaderboard, it suffices to choose

$$\Delta^{tbt\star} := \max_{r \in \{2, \cdots, n\}} \Delta_r^{tbt^\star},$$

so that the global threshold is determined by the hardest adjacent pair to stabilize.

A concrete example. We estimate the baseline catch-up difficulty $e_r^{\text{req}}(0)$ and the learning-rate ratio γ_r from the fitted trajectories in Figure 1. Among all adjacent pairs, the hardest-to-stabilize pair maximizes $\Delta^{tbt\star} = \max_r \Delta_r^{tbt\star}$. For this pair, we estimate $e_r^{\text{req}}(0) \approx 13.7$ and $\gamma_r \approx 0.33$. Suppose the corresponding reward gap translates to $\rho_r = 1000$ units of effort, so the effective incentive is $\lambda_r = \frac{\rho_r}{e_r^{\text{req}}(0)} \approx \frac{1000}{13.7} \approx 73$. Using Proposition E.1, a stabilizing TbT level is

$$\Delta_r^{tbt\star} \approx \left(\frac{\rho_r}{e_r^{\text{req}}(0)}\right)^{\gamma_r} = \left(\frac{1000}{13.7}\right)^{0.33} \approx 4,$$

That is, adding roughly 4 units of baseline TbT effort is already enough to deter overtaking for this hardest-to-stabilize adjacent pair.

Proof for Theorem E.1

Proof. Since σ is strictly increasing, the catch-up condition $v(\theta_r, \Delta^{tbt}+\delta) \geq v(\theta_{r-1}, \Delta^{tbt})$ is equivalent to

$$\tilde{v}(\theta_r, \Delta^{tbt}+\delta) \;\geq\; \frac{v(\theta_{r-1}, \Delta^{tbt}) - L(\theta_r)}{U(\theta_r) - L(\theta_r)}.$$

Applying σ^{-1} and the scaling law for $\tilde{v}(\theta_r, \cdot)$, this becomes

$$\alpha(\theta_r) + \beta(\theta_r)\log(1 + \Delta^{tbt}+\delta) \;\geq\; \sigma^{-1}\Big(\frac{v(\theta_{r-1}, \Delta^{tbt}) - L(\theta_r)}{U(\theta_r) - L(\theta_r)}\Big).$$

Solving with equality yields the minimal additional effort:

$$1 + \Delta^{tbt} + \delta = \exp\Big(\tfrac{1}{\beta(\theta_r)}\Big[\sigma^{-1}\big(\tfrac{v(\theta_{r-1}, \Delta^{tbt}) - L(\theta_r)}{U(\theta_r) - L(\theta_r)}\big) - \alpha(\theta_r)\Big]\Big).$$

Define $e_r^{\text{req}}(\Delta^{tbt})$ to be this minimal δ, then subtracting $(1 + \Delta^{tbt})$ gives

$$e_r^{\text{req}}(\Delta^{tbt}) = \exp\Big(\tfrac{1}{\beta(\theta_r)}\Big[\sigma^{-1}\big(\tfrac{v(\theta_{r-1}, \Delta^{tbt}) - L(\theta_r)}{U(\theta_r) - L(\theta_r)}\big) - \alpha(\theta_r)\Big]\Big) - (1 + \Delta^{tbt}).$$

Using the scaling law for θ_{r-1},

$$\sigma^{-1}\big(\tilde{v}(\theta_{r-1}, \Delta^{tbt})\big) = \alpha(\theta_{r-1}) + \beta(\theta_{r-1})\log(1 + \Delta^{tbt}),$$

and the definition of $\gamma_r = \beta(\theta_r)/\beta(\theta_{r-1})$, we can rewrite the leading term as

$$\exp\Big(\frac{\bar{\alpha}_r(\Delta^{tbt})}{\beta(\theta_r)}\Big)\,(1 + \Delta^{tbt})^{1/\gamma_r},$$

where $\bar{\alpha}_r(\Delta^{tbt}) = \sigma^{-1}(\tilde{v}(\theta_{r-1}, \Delta^{tbt})) - \alpha(\theta_r)$, yielding the stated expression.

With costs $c(e) \geq \kappa e$, any deviation that uses additional effort δ incurs cost at least $\kappa\delta$, so an adjacent-overtake deviation is unprofitable whenever $e_r^{\text{req}}(\Delta^{tbt}) \geq \rho_r$.

Finally, in the regime where the leading term in $e_r^{\text{req}}(\Delta^{tbt})$ dominates $(1 + \Delta^{tbt})$, the condition $e_r^{\text{req}}(\Delta^{tbt}) \geq \rho_r$ is well-approximated by

$$\exp\Big(\frac{\bar{\alpha}_r(\Delta^{tbt})}{\beta(\theta_r)}\Big)\,(1 + \Delta^{tbt})^{1/\gamma_r} \;\gtrsim\; \rho_r.$$

Treating $\bar{\alpha}_r(\Delta^{tbt})$ as approximately constant over the relevant range (or evaluating it at $\Delta^{tbt} = 0$ to obtain a conservative rule of thumb) gives

$$1 + \Delta^{tbt} \;\gtrsim\; \left(\frac{\rho_r}{\exp(\bar{\alpha}_r(0)/\beta(\theta_r))}\right)^{\gamma_r} = \left(\frac{\rho_r}{e_r^{\text{req}}(0) + 1}\right)^{\gamma_r},$$

which implies the stated scaling. Since $\lambda_r = \rho_r / e_r^{\text{req}}(0)$, we also have $\Delta^{tbt\star}_r = O(\lambda_r^{\gamma_r})$ up to a constant factor depending only on $e_r^{\text{req}}(0)$. □

Estimating $e_r^{\text{req}}(0)$, γ_r, and λ_r from the post-training trajectories. For each model i, we fit the generalized scaling form $\sigma^{-1}(\tilde{v}_i(e)) = \alpha_i + \beta_i \log(1+e)$ by linear regression of $\sigma^{-1}(\tilde{v}_i)$ on $\log(1+e)$ over the observed training range, yielding $(\hat{\alpha}_i, \hat{\beta}_i)$. For an adjacent pair $(r-1, r)$ we set

$$\hat{\gamma}_r := \frac{\hat{\beta}_r}{\hat{\beta}_{r-1}}.$$

The baseline catch-up difficulty $e_r^{\text{req}}(0)$ is defined as the minimal additional effort at $\Delta^{tbt} = 0$ required for model r to match model $r-1$; under the fitted scaling law,

$$\widehat{e_r^{\text{req}}(0)} = \exp\Big(\tfrac{\hat{\alpha}_{r-1} - \hat{\alpha}_r}{\hat{\beta}_r}\Big) - 1.$$

Given a reward gap (converted to effort units) $\rho_r = (R_{r-1} - R_r)/\kappa$, we form the effective incentive $\hat{\lambda}_r := \rho_r / \widehat{e_r^{\text{req}}(0)}$.

As a nonparametric check, we also estimate a local slope ratio at a common baseline $e = \Delta^{tbt}$ using finite differences on the fitted trajectories:

$$\widehat{s}_i(\Delta^{tbt}) \approx \frac{\hat{v}_i(\Delta^{tbt} + h) - \hat{v}_i(\Delta^{tbt})}{h}, \qquad \widehat{\gamma}_r^{\text{slope}}(\Delta^{tbt}) := \frac{\widehat{s}_r(\Delta^{tbt})}{\widehat{s}_{r-1}(\Delta^{tbt})}.$$

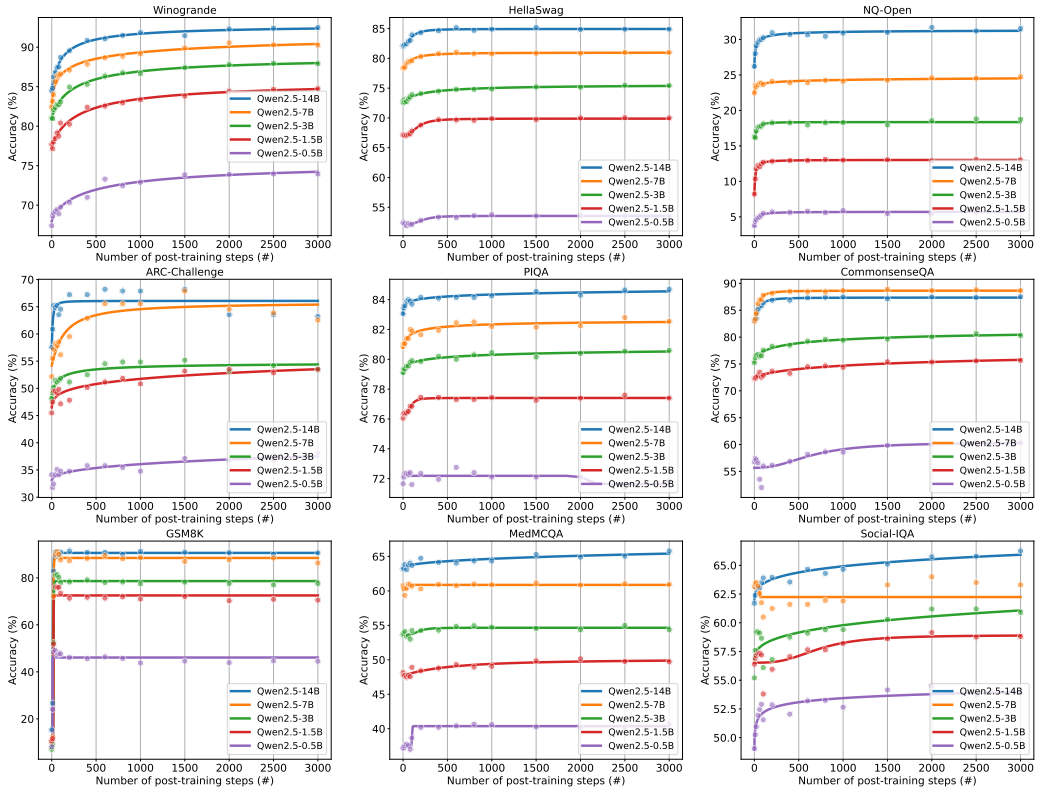

Figure 3: Continued post-training trajectories of *Qwen* models of different sizes on nine benchmarks. Here, we use model size as a proxy for the model's latent capability θ. The x-axis denotes the number of post-training steps, reflecting post-training effort e. The y-axis denotes accuracy on the validation set, i.e., $v(\theta, e)$. For each model, we fit a curve following Equation equation 3.

F Additional Empirical Results

Setting We conduct our emprical study on nine benchmarks, `Winogrande` (Sakaguchi et al., 2019), `HellaSwag` (Zellers et al., 2019a), `NQ-Open` (Kwiatkowski et al., 2019), `ARC-Challenge` (Clark et al., 2018), `Piqa` (Bisk et al., 2019), `CommonsenseQA` (Talmor et al., 2019), `Gsm8k` (Cobbe et al., 2021), `MedMcQA` (Pal et al., 2022), and `Social-IQA` (Sap et al., 2019). Each model is trained with LoRA (Hu et al., 2021; Mangrulkar et al., 2022) (rank 8, α=32) and AdamW optimizer (Loshchilov & Hutter, 2017) (weight decay 0.01 and learning rate 5e-5). We use a batch size of 8, so each training step corresponds to eight training data points. For those benchmarks without a validation split, we randomly allocate 20% of the training data as the validation set.

Results In Figure 3, we present the results of controlled post-training experiments for all nine benchmarks. The assumption 4.2 still largely holds in most benchmarks, despite a few anomalies. In addition, on benchmarks like `HellaSwag`, continued post-training cannot make a smaller model catch up to a larger one. This is not surprising, since model sizes (the proxy for θ) differ significantly in our controlled setting. As a result, there is no incentive to conduct strategic post-training in this scenario. We argue, however, that in real-world settings with models from diverse sources, θ will not differ so dramatically.