

000
001
002
003
004
005
006
007
008
009
010
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053

SCALING LAWS FOR GENERATIVE REWARD MODELS

Anonymous authors

Paper under double-blind review

ABSTRACT

We study the scaling behavior of generative reward models (GenRMs) for reinforcement learning from AI feedback (RLAIF) when used as drop-in replacements for Bradley-Terry models to optimize policies. Building on established scaling laws for reward model overoptimization, we investigate whether GenRMs, particularly those employing chain-of-thought reasoning, exhibit different robustness properties as policies drift from their training distribution during gradient updates. Using the Qwen3 model family (0.6B–14B), our study includes systematic evaluation of thinking GenRMs (trained via GRPO) against answer-only variants (trained via SFT) across policy size, reward model size, reward model type, training budget, and the β parameter in online DPO. Our results reveal a consistent *evaluator-rewarder gap*: thinking GenRMs outperform answer-only variants by 1–2% on validation tasks, yet these gains diminish—and often reverse—during policy optimization, where answer-only GenRMs achieve higher Gold Elo and more stable proxy-Gold alignment. We find that reward model scale is the most decisive factor for policy quality, with gains continuing even when the GenRM far exceeds the policy in parameters. Moreover, intermediate GRPO checkpoints of thinking judges can outperform fully-trained checkpoints as rewarders, despite worse static accuracy. We track these dynamics with Elo arenas under both proxy and Gold evaluation, providing a fine-grained proxy–Gold alignment diagnostic beyond saturated validation metrics.

1 INTRODUCTION

Policies trained via reinforcement learning from human feedback systematically exploit imperfections in their reward models, degrading true quality even as proxy metrics improve (Gao et al., 2022; Rafailov et al., 2024). Generative reward models (GenRMs) have emerged as a promising response: by formulating evaluation as conditional text generation rather than scalar prediction, GenRMs can leverage pre-training priors and chain-of-thought reasoning and achieve strong performance on static preference benchmarks (Mahan et al., 2024; Zhang et al., 2025; Whitehouse et al., 2025). This success has fostered an implicit assumption that more accurate static evaluators should translate to more effective online rewarders.

Yet GenRMs have been studied almost exclusively as judges on fixed distributions, not as in-loop rewarders where the policy distribution shifts during training. In this work, we test that assumption and target the practitioner-facing question it conceals: *if one chooses a GenRM as the reward signal, how should compute and architecture be allocated to maximize downstream policy quality?* We study three design dimensions in the online setting where they matter: (i) thinking versus answer-only judging, (ii) judge scale, and (iii) policy scale to maximize downstream policy quality rather than static accuracy? We instantiate these axes in creative writing using LITBENCH (Fein et al., 2025), a non-verifiable preference domain where correctness cannot be programmatically checked. In such domains, static evaluator accuracy can saturate while reward hacking and proxy–Gold divergence remain salient, making creative writing an ideal testbed for whether stronger static evaluators translate to better in-loop reward signals.

Contributions. (1) We present the first systematic scaling study of GenRMs deployed as online reward signals, sweeping judge format, judge size (0.6B–14B), policy size (0.6B–8B) and training budgets under online DPO. (2) We introduce a proxy–Gold Elo diagnostic that measures alignment under optimization pressure and supports cross-configuration comparison. (3) We provide empirical

guidance for GenRM deployment: judge scale and training duration dominate, answer-only GenRMs yield more stable optimization than thinking variants, and intermediate GenRM checkpoints can outperform fully-trained models as rewarders.

Our findings reveal a consistent *evaluator–rewarder gap*. Thinking GenRMs improve in-distribution accuracy by 1–2% across all scales, yet these gains diminish—and often reverse—when the same judges drive online optimization. Answer-only GenRMs yield higher peak Gold Elo and more stable proxy–Gold alignment. In contrast, scaling judge capacity proves decisive: larger judges continue improving policy quality even when far exceeding the policy in parameters, suggesting that judging is at least as hard as generating in this domain.

2 METHODOLOGY

We begin with a human preference dataset of the form

$$\mathcal{D}_{\text{human}} = \{x^{(i)}, y_A^{(i)}, y_B^{(i)}, I_H^{(i)}\}_{i=1}^N,$$

where $x^{(i)} \in \mathcal{X}$ are prompts, $y_A^{(i)}, y_B^{(i)} \in \mathcal{Y}$ are pairs of responses to the prompts, and $I_H^{(i)} \in \{A, B\}$ denotes the human-preferred response.

Gold evaluator. Since human preferences are not available on demand during policy training, we cannot directly query them to measure policy performance at each checkpoint. To enable continuous evaluation throughout the optimization process, we train a large generative reward model (GenRM), denoted V_{gold} , on $\mathcal{D}_{\text{human}}$. This model serves as a *Gold* evaluator that provides a stable, queryable preference distribution anchored to human judgments. Grounding V_{gold} in human preferences ensures that our experimental setup reflects real-world alignment scenarios where the ultimate objective is to match human judgment. However, our study does not require V_{gold} to perfectly replicate human preferences. Rather, we require a *consistent anchor distribution* against which we can measure proxy GenRM behavior as policies drift during training. The central question we investigate is how well smaller proxy GenRMs generalize when used as reward signals, relative to this anchor—not whether the anchor itself is perfectly human-aligned.

To make evaluation computationally feasible, we choose V_{gold} to be an *Answer-Only* model that outputs a single indicator token. Using V_{gold} , we construct a Gold Preference dataset:

$$\mathcal{D}_{\text{gold}} = \{x^{(i)}, y_A^{(i)}, y_B^{(i)}, I_G^{(i)}\}_{i=1}^M,$$

which is then used to train a variety of smaller GenRMs.

Proxy GenRMs. We consider two types of GenRMs: *Answer-Only* and *Thinking*. The Answer-Only models output a direct judgment token (A or B):

$$I \sim v_{\text{ans}}(\cdot \mid x, y_A, y_B),$$

while the Thinking models first generate a reasoning trace z before producing the final verdict:

$$z \sim v_{\text{think}}(\cdot \mid x, y_A, y_B), \quad I \sim v_{\text{think}}(\cdot \mid x, y_A, y_B, z).$$

Training objectives. We train Answer-Only models using supervised fine-tuning (SFT) and Thinking models using GRPO. This asymmetry reflects the training pipelines used by practitioners in real-world settings. For Answer-Only models, the supervision signal is unambiguous: given a preference label, the model should predict the corresponding token. SFT directly optimizes this objective and is the natural choice. For Thinking models, however, the supervision problem is fundamentally different. A chain-of-thought has no single correct trajectory, and current practice overwhelmingly relies on reinforcement learning objectives like GRPO. Our goal is not to perform an architectural ablation under identical training algorithms, but rather to compare the best-supported pipeline for each model type—training each with the method that empirically performs best for its structure.

For GRPO, we employ two reward signals:

1. **Accuracy reward:** A binary reward (1 if the model reaches the correct verdict, 0 otherwise).

108
 109
 110
 111
 112
 113
 114

2. **Positional consistency reward r_{pos} :** Inspired by prior work, we observe that models often produce contradictory judgments when the order of (y_A, y_B) is swapped in the prompt. To mitigate this, we explicitly place both orderings (A, B) and (B, A) into the *same GRPO group*, and compute a group-level majority vote over the sampled completions. A reward of 1 is assigned only if the majority verdicts under both orderings are consistent and match the correct label. To avoid introducing noise, r_{pos} is only given to completions that end at the correct verdict.

115
 116 We refer to these trained GenRMs as *Proxy models*, and evaluate them with respect to V_{gold} preferences.
 117

118 **Policy optimization.** We train policies π_θ of varying sizes using *Online Direct Preference Optimization* (Online DPO). DPO is a natural choice for preference-based policy optimization because
 119 GenRMs produce pairwise preferences as their native output, which DPO can consume directly
 120 without requiring scalarization or additional transformations. In *offline* DPO, response pairs are
 121 sampled once from a fixed reference policy and reused across many gradient updates, keeping the
 122 supervision distribution static throughout training. In *online* DPO, response pairs are repeatedly
 123 resampled from the current policy after each update, so the supervision distribution evolves as the
 124 policy changes. This online setting is particularly important for our study, as we aim to understand
 125 GenRM robustness under distribution shift induced by policy optimization. Empirically, online pref-
 126 erence optimization has been shown to yield more stable and effective alignment dynamics when
 127 implemented carefully (Guo et al., 2024).

128 The policies are trained on prompts sampled from the same distribution \mathcal{X} used for V_{gold} and the
 129 Proxy models. During training, we periodically save checkpoints and sample responses on a fixed
 130 validation set of prompts. To evaluate these checkpoints, we compute ELO ratings based on pairwise
 131 comparisons of their responses. ELO evaluation provides a more fine-grained measurement than
 132 raw win rates against a fixed reference distribution. We compute ELOs with respect to both Proxy
 133 models and the Gold model, enabling us to analyze the relationship between Proxy-based evaluation
 134 and Gold-standard evaluation.

135
 136

3 EXPERIMENTAL SETUP

 137

138 **Dataset construction.** We construct three disjoint datasets for our experiments:
 139

140
 141
 142
 143

- **Human preference dataset $\mathcal{D}_{\text{human}}$:** We sample 21,000 preference triplets (x, y_A, y_B, I_H)
 144 from LITBENCH Fein et al. (2025), a large-scale preference dataset over human-written
 145 stories from Reddit. We split these into 20,000 for training and 1,000 for validation. This
 146 dataset is used exclusively for training the Gold evaluator.

147
 148

- **Gold preference dataset $\mathcal{D}_{\text{gold}}$:** We sample a separate, non-overlapping set of 21,000
 149 triplets from LITBENCH and re-annotate them using the trained Gold evaluator, yielding
 150 (x, y_A, y_B, I_G) triplets. Critically, while both $\mathcal{D}_{\text{human}}$ and $\mathcal{D}_{\text{gold}}$ are sourced from LIT-
 151 BENCH, they represent different preference distributions: the former reflects human judg-
 152 ments, while the latter reflects the Gold model’s judgments. We split $\mathcal{D}_{\text{gold}}$ into 20,000 for
 153 training and 1,000 for validation. This dataset is used for training and evaluating Proxy
 154 GenRMs.

155
 156

- **Policy training dataset:** We source prompts from the WRITINGPROMPTS dataset¹, the an-
 157 cestor dataset from which LITBENCH was constructed. To avoid data leakage, we exclude
 158 all prompts that appear in LITBENCH, yielding 199,000 candidate prompts. We randomly
 159 sample 91,000 prompts (90,000 for training and 1,000 for validation) unless specified oth-
 160 erwise in specific experiments.

161 Building on prior work that has extensively evaluated various reward models on LITBENCH facil-
 162 itates future research and cross-comparisons between our results and other studies.

163 **Model families.** We use the QWEN3 model family for all GenRMs and policies. For Proxy Gen-
 164 RMs, QWEN3 is the only open-weight model family that supports both Thinking and Answer-Only

¹https://huggingface.co/datasets/euclaise/WritingPrompts_preferences

162 modes across multiple parameter sizes. This enables controlled comparisons where we can vary
 163 GenRM mode and scale while holding the backbone architecture and instruction-tuning distribution
 164 fixed. For policy models, using the same backbone as the GenRMs allows us to cleanly isolate the
 165 interaction between policy size and GenRM size, which is central to our scaling analysis.

166 **Gold GenRM training.** We train the Gold evaluator from QWEN3-32B using supervised fine-
 167 tuning on $\mathcal{D}_{\text{human}}$. To improve positional robustness, we train on both orderings (y_A, y_B) and
 168 (y_B, y_A) , resulting in 40,000 training pairs. We use a batch size of 128 and a learning rate of
 169 1×10^{-5} .

170 **Proxy GenRM training.** We train Proxy GenRMs from the QWEN3 series at sizes $\{0.6\text{B}, 1.7\text{B}, 4\text{B}, 8\text{B}, 14\text{B}\}$ using $\mathcal{D}_{\text{gold}}$. We refer to the off-the-shelf, untrained QWEN3 models as
 171 Baseline- $\{\text{size}\}$ -Ans and Baseline- $\{\text{size}\}$ -Think for Answer-Only and Thinking
 172 modes, respectively. For Answer-Only models, we use supervised fine-tuning with batch size 128
 173 and learning rate 1×10^{-5} . As with the Gold model, we train on both positional orders, yielding
 174 40,000 training pairs. For Thinking models, we use GRPO with the following configuration: 16
 175 prompts per step, group size 8, minibatch size 64, no KL penalty, and learning rate 1×10^{-6} . We
 176 train on both positional orders. We refer to the trained Proxy GenRMs as GenRM- $\{\text{size}\}$ -Ans for
 177 Answer-Only models and GenRM- $\{\text{size}\}$ -Think for Thinking models, where $\{\text{size}\}$ denotes
 178 the parameter scale.

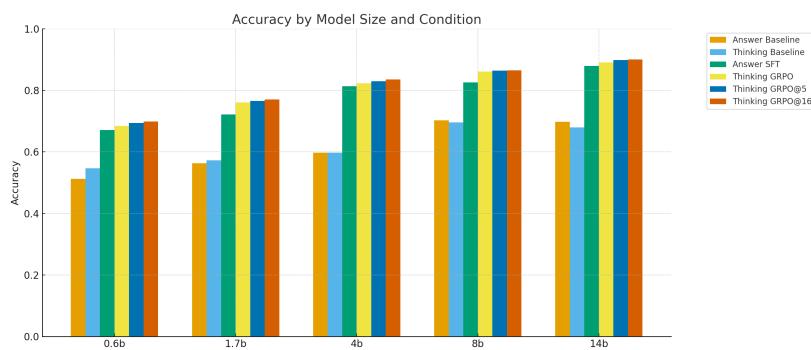
179 **Policy training.** We train policies π_θ from the QWEN3 series at sizes $\{0.6\text{B}, 1.7\text{B}, 4\text{B}, 8\text{B}\}$ using
 180 Online DPO. All policies are trained with thinking disabled. We use a training batch size of 192,
 181 minibatch size of 64, learning rate of 1×10^{-6} , and $\beta = 0.02$ unless specified otherwise in specific
 182 experiments. For each sampled prompt, we generate response pairs from the current policy and
 183 evaluate them using a Proxy GenRM. To ensure positional consistency, we compute preferences
 184 under both orderings (y_A, y_B) and (y_B, y_A) , retaining only pairs where the two orderings agree. We
 185 restrict prompt length to 512 tokens and response length to 2048 tokens.

186 **Evaluation.** During policy training, we save checkpoints every 10 training steps. For each check-
 187 point, we sample responses on the fixed validation set and compute ELO ratings based on pairwise
 188 comparisons judged by both the Proxy GenRM used during training and the Gold evaluator. Pol-
 189 icy training runs for 460 steps in the main experiments (yielding 47 checkpoints), though specific
 190 ablations may use different training durations.

191 **Reproducibility and LLM usage.** Beyond the GenRM and policy models described above, we did
 192 not use LLMs for dataset filtering, rubric creation, label consolidation, or prompt generation. All
 193 prompts used for policy generation and GenRM pairwise evaluation are provided in Appendix B.

194 4 EXPERIMENTS AND RESULTS

195 4.1 GENRM TRAINING



216 Figure 1: Performance of trained GenRMs of different sizes and training methods on the Gold Preference
 217 dataset. Thinking GenRMs consistently outperform Answer-Only models across all scales, while both sub-
 218 stantially exceed baseline performance. **Takeaway:** Training yields $\sim 20\%$ accuracy gains; Thinking models
 219 provide a consistent 1–2% edge as static evaluators.

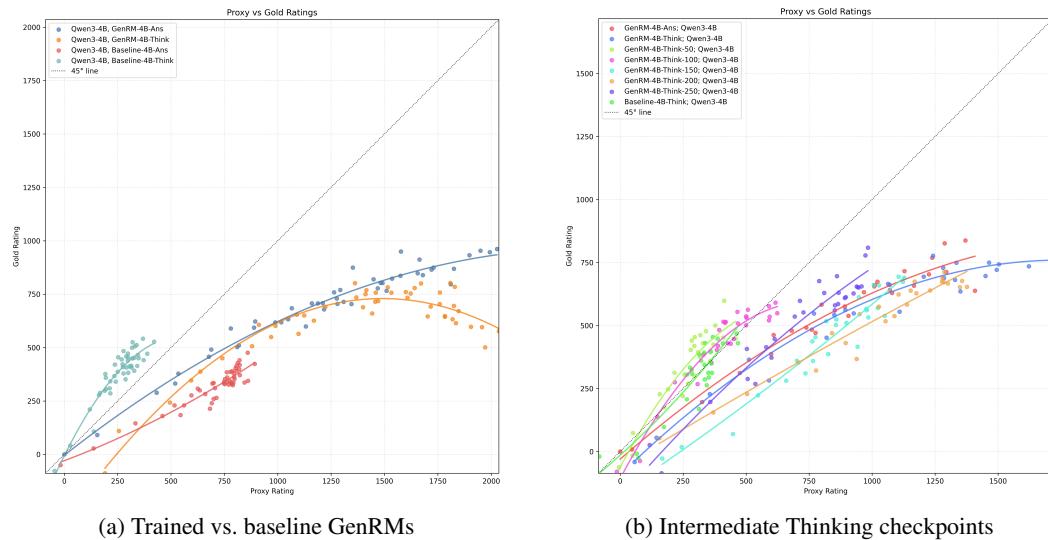
216 **Gold evaluator alignment.** The Gold GenRM achieves 79% agreement with human preferences
 217 on the held-out validation split of $\mathcal{D}_{\text{human}}$. Prior work on LITBENCH reports peak accuracies of
 218 approximately 78% Fein et al. (2025), indicating that our Gold evaluator achieves strong alignment
 219 with the human preference distribution. While our study only requires a stable reference distribution
 220 rather than perfect human alignment, the 79% agreement demonstrates that the Gold evaluator is
 221 well-grounded in human judgments.

222 **Trained GenRMs substantially outperform baselines.** We train GenRMs of sizes 0.6B, 1.7B,
 223 4B, 8B, and 14B, with both Answer-Only (SFT) and Thinking (GRPO) variants (Figure 1). The
 224 untrained baseline models (both Answer-Only and Thinking) achieve similar accuracies across all
 225 sizes, suggesting no apparent bias in the Gold evaluator toward either format. Across all scales,
 226 trained models show an approximately 20% accuracy improvement over their respective baselines,
 227 confirming substantial room for optimization and indicating that the Gold anchor distribution is
 228 sufficiently distinct from the untrained proxy distributions.

229 **GRPO is an effective training method for Thinking GenRMs.** To validate our choice of GRPO
 230 for training Thinking models, we conducted an ablation study on the 4B model. When evaluated
 231 on the Gold Preference dataset, the off-the-shelf model achieves 60.0%, STaR achieves 60.7%,
 232 distillation from QWEN3-32B thinking traces (SFT-style) achieves 64.7%, and GRPO with accuracy
 233 and positional consistency rewards achieves 83.2%. Only GRPO yields substantial improvements,
 234 validating its use as the primary training method for Thinking GenRMs.

235 **Thinking GenRMs outperform Answer-Only models on in-distribution evaluation.** On average
 236 across all scales, trained Thinking GenRMs achieve 1.7% higher accuracy than their Answer-Only
 237 counterparts when evaluated on the Gold Preference dataset. To evaluate whether sampling multiple
 238 completions yields additional gains, we compute majority-vote accuracy over $k = 5$ and $k = 16$
 239 samples from the Thinking models (Figure 1). This yields modest further improvements of 0.5%
 240 and 0.8% respectively (e.g., at 14B, 89.1% \rightarrow 89.6% \rightarrow 90.0%). Given the computational overhead
 241 and minimal gains, we do not employ multi-sampling during policy training.

242 4.2 TRAINED GENRMS VS. BASELINE MODELS IN POLICY TRAINING



263 Figure 2: Proxy vs. Gold ELO ratings for policies trained with different GenRMs. (a) Comparison of trained
 264 vs. baseline GenRMs. (b) Policies trained on 50,000 prompts with Thinking GenRMs, where the GenRM is
 265 checkpointed at intermediate stages of GRPO training (50–250 steps). In both plots, the x-axis shows Proxy
 266 ELO and the y-axis shows Gold ELO, with the dotted line indicating perfect alignment (45°).

267 *Does training the GenRM improve downstream policy quality compared to off-the-shelf models?*

268 We compare policy training using trained versus baseline (off-the-shelf) GenRMs. In this ex-
 269 periment, both the policy and GenRM are 4B models, with Answer-Only and Thinking vari-

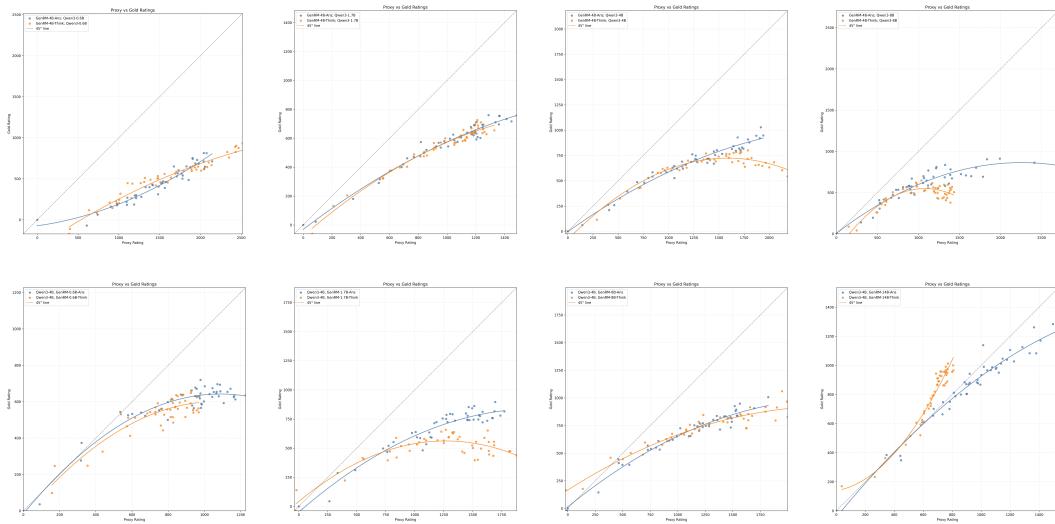
270 ants (Figure 2a). Trained GenRMs dramatically outperform their baseline counterparts. Notably, while Thinking outperforms Answer-Only at baseline, this advantage reverses after training: 271 GenRM-4B-Ans exceeds GenRM-4B-Think by over 200 Elo points. We discuss this reversal 272 further in the next subsection.

274

275

4.3 ANSWER-ONLY VS THINKING

276



294

295 Figure 3: Proxy vs. Gold ELO ratings for policies trained with Answer-Only (blue) vs. Thinking GenRMs 296 (orange) across multiple policy sizes and GenRM sizes. Top row: policy sizes 0.6B, 1.7B, 4B, 8B with fixed 297 4B GenRM. Bottom row: 4B policy with GenRM sizes 0.6B, 1.7B, 8B, 14B. The dotted line indicates perfect 298 alignment. **Takeaway:** Answer-Only GenRMs consistently achieve higher Gold Elo and tighter proxy-Gold 299 alignment than Thinking GenRMs across all configurations.

300

301

Does chain-of-thought reasoning improve GenRM effectiveness as online rewarders?

302

303 Across all combinations of policy sizes (0.6B, 1.7B, 4B, 8B) and GenRM sizes (0.6B–14B), policies 304 trained with Answer-Only GenRMs achieve both higher maximum Gold ELO and smaller discrepancies 305 between Proxy and Gold ratings compared to those trained with Thinking GenRMs (Figure 3).

306

307 This finding is surprising for two reasons. First, in-distribution evaluation (Section 4.1) showed 308 Thinking GenRMs outperforming Answer-Only models by \sim 1–2% accuracy. Second, baseline (un- 309 trained) models displayed the opposite trend, with Thinking variants aligning better with the Gold 310 model than Answer-Only. Despite these initial advantages, Thinking GenRMs prove less reliable 311 when used as reward models for online policy optimization.

312

313 We interpret this as evidence that Thinking GenRMs are more vulnerable to off-distribution shifts 314 introduced by policy training. Two factors may explain this gap.

315

316 First, higher in-distribution GenRM accuracy does not necessarily translate to better policy training. 317 Since Online DPO changes the response distribution during training, GenRM robustness across the 318 full domain matters more than performance on a fixed validation set. Figure 2b provides direct 319 evidence: we trained policies using intermediate checkpoints of a 4B Thinking GenRM during GRPO 320 training (at 0, 50, 100, 150, 200, and 250 steps). While GenRM accuracy steadily improves on the 321 in-distribution set (from 59.7% to 82.3%), policy Gold Elo peaks at an intermediate checkpoint and 322 then degrades. Notably, the 250-step checkpoint (79.2% accuracy) produces policies that perform 323 close to the Answer-Only baseline in Gold Elo, with a more favorable slope between Proxy and Gold 324 ratings. This suggests overfitting to the evaluation distribution hurts generalization during policy 325 optimization, and that intermediate Thinking checkpoints may provide more effective learning 326 signals than fully-trained models.

327

328 Second, GRPO optimizes final preferences rather than reasoning correctness, allowing flawed 329 reasoning paths that still reach correct labels. These weaknesses may be tolerated in-distribution but 330

exploited out-of-distribution. Because Thinking models split computation between latent activations and explicit language traces, this may introduce additional failure modes compared to Answer-Only models, which perform computation entirely in the latent space.

In summary, Answer-Only reward models are more token efficient and demonstrably more robust for preference-based policy optimization in this domain, though intermediate Thinking checkpoints retain potential if their training dynamics are better understood.

4.4 EFFECT OF GENRM SIZE

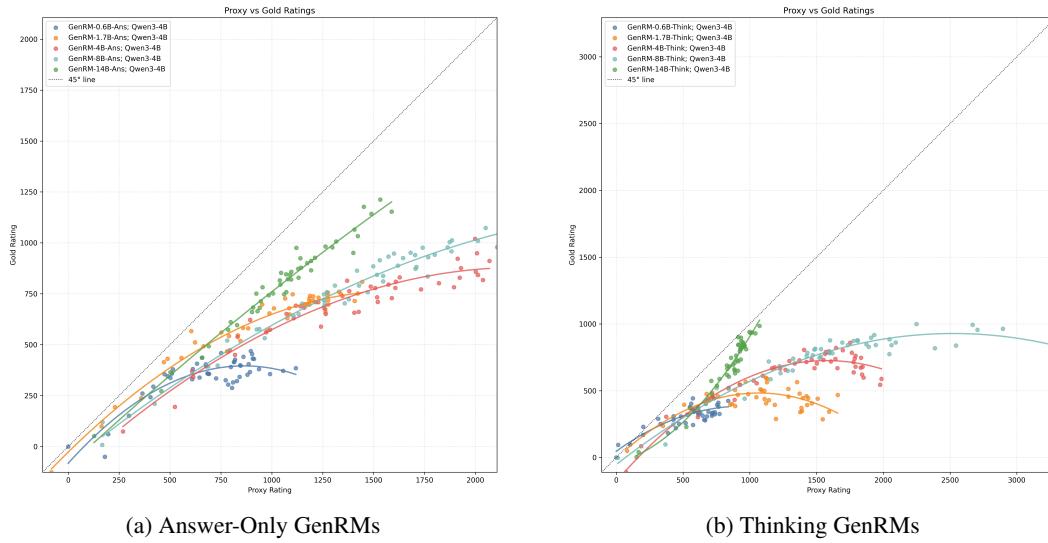


Figure 4: Proxy vs. Gold ELO ratings for policies trained with GenRMs of different sizes. The policy is fixed at 4B. Left: Answer-Only GenRMs (0.6B–14B). Right: Thinking GenRMs (0.6B–14B). The dotted line indicates perfect alignment. **Takeaway:** Increasing GenRM size improves Gold Elo and tightens proxy-Gold alignment, with gains continuing even when the GenRM far exceeds the policy in parameters.

Does scaling the reward model improve policy quality, and do gains continue when the GenRM exceeds the policy in size?

We investigate the effect of scaling GenRM size while fixing the policy size at 4B. Figure 4 shows results for both Answer-Only and Thinking GenRMs across scales from 0.6B to 14B.

We observe clear and consistent gains from increasing GenRM size. Notably, performance continues to improve even when the GenRM is much larger than the policy: at 14B, both Answer-Only and Thinking models yield substantial gains over smaller counterparts. This trend suggests that the capacity of the evaluator plays a decisive role in stabilizing and guiding preference-based training. The result highlights an important asymmetry: within this domain, “judging” appears to be as hard—or harder—than “generating.”

4.5 EFFECT ON POLICY SIZE

Does scaling the policy improve quality, and does the effect differ between Answer-Only and Thinking supervision?

We examine the effect of scaling policy size while fixing the GenRM. Figure 5 shows results for policies ranging from 0.6B to 8B trained with both Answer-Only and Thinking GenRMs.

For Answer-Only supervision, the trend is straightforward: larger policies consistently achieve higher Gold ELO, with steady improvements across scales. This aligns with the expectation that larger policies better exploit the reward signal and generalize more effectively.

For Thinking supervision, however, the picture is less clear. While the largest policy does achieve the highest peak performance, its Gold ELO curve saturates and bends downward earlier than smaller

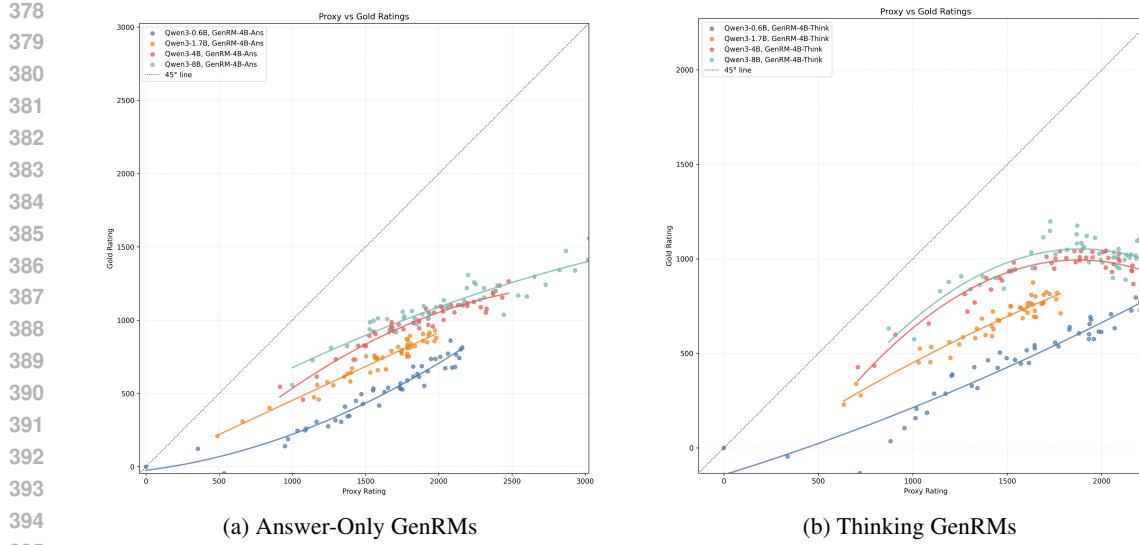


Figure 5: Proxy vs. Gold ELO ratings for policies of different sizes trained with fixed GenRMs. Left: Answer-Only GenRMs. Right: Thinking GenRMs. The dotted line indicates perfect alignment. **Takeaway:** Larger policies consistently improve under Answer-Only supervision, but under Thinking supervision, large policies saturate earlier, suggesting a more fragile interaction between policy capacity and GenRM capacity.

policies, which continue to improve steadily. This suggests two possible explanations: (i) large policies may more quickly exhaust the effective capacity of the GenRM, reaching its “ceiling” earlier, or (ii) beyond a certain scale, further increasing policy size without correspondingly stronger GenRMs may become counterproductive.

Additional training is required to disentangle these explanations, but the evidence points toward an important asymmetry: scaling policy size is reliably beneficial under Answer-Only supervision, but under Thinking supervision, the interaction between policy capacity and GenRM capacity is more fragile.

4.6 EFFECT OF THE β COEFFICIENT IN ONLINE DPO

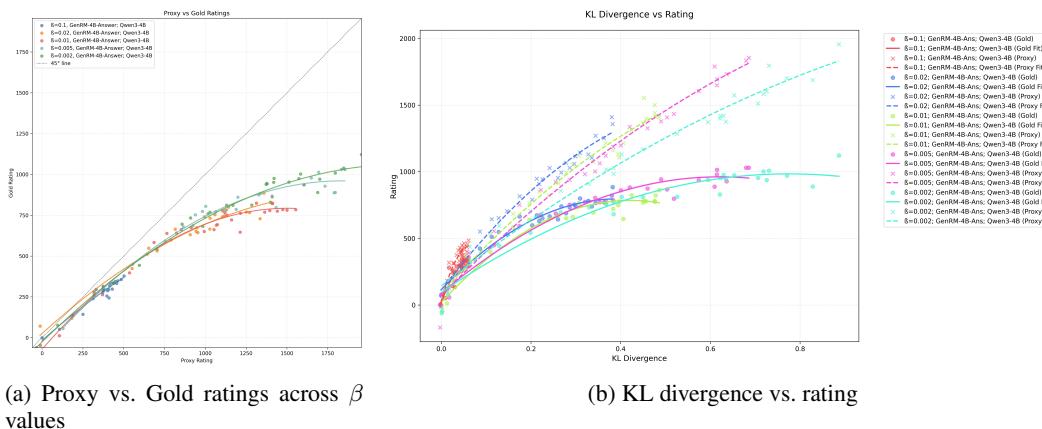


Figure 6: Effect of the β coefficient in Online DPO. Left: Proxy vs. Gold ELO ratings for Answer-Only GenRMs at different β values. Right: KL divergence vs. rating tradeoff, showing how larger β suppresses exploration. The dotted line indicates perfect alignment. **Takeaway:** Smaller β values (0.002–0.01) yield higher Gold Elo by allowing more exploration, but the effect is secondary compared to GenRM or policy scaling. All policies have been trained on 50,000 prompts.

432 *How does the DPO regularization strength affect policy quality and the exploration-exploitation*
 433 *tradeoff?*

434
 435 We analyze the effect of the β coefficient in Online DPO. Figure 6(a) shows Proxy vs. Gold ELO
 436 ratings for policies trained with $\beta \in \{0.1, 0.02, 0.01, 0.005, 0.002\}$ under Answer-Only supervision.
 437 Overall, smaller β values consistently yield better Gold performance. While gains become increas-
 438 ingly incremental as β decreases, the smallest value ($\beta = 0.002$) achieves the highest Gold ELO.

439 Importantly, lower β does not simply increase Proxy ratings at the expense of Gold performance.
 440 Instead, the lowest β exhibits the slope closest to the 45° line, indicating the least degree of over-
 441 optimization among all settings.

442 Figure 6(b) further illustrates the KL-performance tradeoff. Larger β values overly constrain the
 443 policy to the reference distribution, limiting attainable Gold ratings. Reducing β relaxes this con-
 444 straint and improves alignment with Gold outcomes, without inducing severe divergence.

445 Overall, smaller β values are strictly preferable in this regime. Although the marginal gains diminish
 446 at very low β , the smallest values simultaneously achieve the highest Gold ELO and the most faithful
 447 Proxy–Gold correspondence.

448

449

450 5 LIMITATIONS AND DISCUSSION

451

452 **Gold evaluator is a learned proxy, not humans.** We mitigate label noise by fine-tuning a 32B
 453 Gold evaluator on human preferences, re-annotating pairwise data with this model, and discarding
 454 items with inconsistent Gold verdicts (Gao et al., 2022). This improves consistency but couples our
 455 objective to the Gold model’s inductive biases. At 32B parameters, the Gold judge is stronger than
 456 our proxies, leaving headroom for improvement, yet it still reflects preferences from its training set.
 457 *Takeaway.* Interpret results as optimization toward a strong, learned proxy.

458 **Answer-only format for the Gold evaluator.** The Gold judge emits a single verdict token. This
 459 choice improves throughput and simplifies adjudication, but might favor answer-only proxies in
 460 subtle ways. We investigated this concern empirically and found no evidence of systematic bias.
 461 First, off-the-shelf (untrained) answer-only and thinking judges show similar agreement patterns on
 462 the Gold-labeled set, with differences small relative to the trained-vs-baseline gap (Figure 1): at
 463 smaller sizes Thinking models slightly outperform, at mid-size they are comparable, and at larger
 464 sizes Answer-Only holds a small edge. Second, in Figure 2a (baseline proxies), policies trained
 465 with thinking baseline judges trace a slope closer to the diagonal, indicating tighter proxy–Gold
 466 agreement even though Gold is answer-only. These observations suggest that the Answer-Only
 467 advantage we report in trained GenRMs reflects genuine optimization dynamics rather than evaluator
 468 format bias. Nonetheless, we do not evaluate an alternative thinking-style Gold, which would further
 469 strengthen robustness. *Takeaway.* Empirical checks show no format bias at baseline; a rationale-
 470 producing Gold remains valuable future work.

471

472 **Domain scope.** All experiments use creative writing; we do not systematically evaluate other do-
 473 mains such as safety, helpfulness, or verifiable tasks (math, code). Replicating our full experimental
 474 grid in additional domains would scale cost roughly linearly: each (judge, policy) configuration
 475 requires an online DPO run plus Elo evaluation, costing 6–20 hours on 8×H100, not including
 476 judge training. Given our ablation count (policy size \times judge size \times judge mode \times training-
 477 budget/ β /checkpoint sweeps), full multi-domain replication would require thousands of additional
 478 H100-hours. To provide preliminary evidence, we ran cross-domain checks on OpenOrca using
 479 off-the-shelf Qwen3 models (Appendix A.1); these support the same qualitative pre-training trend
 (thinking judges start stronger as static evaluators). *Takeaway.* Our conclusions are scoped to non-
 480 verifiable creative writing; domain transfer is an important open question.

481

482 **Relation to prior creative-writing studies.** Creative-writing preferences are subjective. Optimiz-
 483 ing win rate can reduce stylistic diversity. We observe some style convergence (Chung et al., 2025)
 484 in late-stage policies against both judge modes, while draw rates do not show large collapses. Fein
 485 et al. (2025) report that chain-of-thought can degrade verification accuracy for creative writing, and
 486 that trained BT and generative verifiers outperform zero-shot judges on their benchmark. In our set-
 487 ting, distilling Qwen3-32B traces into Qwen3-4B also yields negligible gains as evaluators, which
 488 aligns with these observations. However, after training, thinking judges significantly improve as

486 static evaluators. The gap between thinking and answer-only during policy optimization therefore
 487 cannot be explained by domain “unsuitability” alone; it indicates different optimization dynamics.
 488 *Takeaway.* Our evaluation–optimization divergence is a property of the training loop in this domain,
 489 not only a property of static judging.

490 **Family and algorithm scope, and behavioral priors.** All models use Qwen3 backbones where
 491 pretraining data are not public. Cross-family studies suggest that behavioral priors, including syn-
 492 synthetic data that instantiate verification or backtracking, can modulate RL improvements and collapse
 493 family gaps (Gandhi et al., 2025). Such priors could shift our coefficients. We do not evaluate
 494 PPO-style RLHF for policies or alternative thinking-judge recipes. Additionally, we do not include
 495 Best-of-N (BoN) inference-time comparisons; our focus is on GenRMs as *training signals* for policy
 496 optimization rather than inference-time scaling, where the primary gains in our setting arise from
 497 gradient-based learning rather than reranking at generation time. BoN comparisons remain valuable
 498 future work for understanding the full compute-performance tradeoff. *Takeaway.* Our coefficients
 499 and inflection points are conditional on online DPO and Qwen3 underlying behavioral priors.

500 **Elo anchoring and schedule.** Elo is anchored to the Gold evaluator and depends on the match
 501 schedule (Chiang et al., 2024). We report both global and size-stratified arenas, but we do not study
 502 alternative anchors or tournament designs. *Takeaway.* Absolute Elo levels can shift with the anchor,
 503 while within-arena orderings are more stable.

504 6 RELATED WORK

505 Alignment from preferences exhibits predictable overoptimization: gold reward degrades as policies
 506 drift from a reference, following smooth scaling laws for both RLHF (Ouyang et al., 2022; Gao et al.,
 507 2022) and direct methods like DPO that remove explicit reward heads (Rafailov et al., 2023; 2024).
 508 This motivated architecturally unified judges. Generative Reward Models (GenRMs) replace scalar
 509 heads with next-token prediction, enabling rationales alongside verdicts (Mahan et al., 2024; Zhang
 510 et al., 2025). Mahan et al. (2024) trained GenRMs via iterative self-taught reasoning with DPO,
 511 achieving strong out-of-distribution generalization. Subsequent work scales these approaches: J1
 512 extends with GRPO and positional consistency rewards (Whitehouse et al., 2025), DeepSeek-GRM
 513 adds Self-Principled Critique Tuning with meta-aggregation for inference-time scaling (Liu et al.,
 514 2025), and Heimdall demonstrates test-time improvements via majority voting in verification tasks
 515 (Wang et al., 2025; Shi & Jin, 2025). Complementary supervision strategies include self-generated
 516 critiques (Yu et al., 2025) and criteria trees (Liang et al., 2025), while EvalPlanner frames evaluation
 517 as plan-and-reason generation (Saha et al., 2025). Despite extensive work on judge reliability (Ye
 518 et al., 2024) and benchmarks (Lambert et al., 2024), the field conflates static evaluation accuracy
 519 with rewarder effectiveness. We disambiguate these roles through controlled scaling experiments
 520 with answer-only (SFT) versus thinking (GRPO) GenRMs as both evaluators and online DPO re-
 521 warders, using Elo arenas (Chiang et al., 2024) for unified comparison. Our results reveal when
 522 inference-time reasoning helps evaluation but hinders policy optimization under matched FLOPs
 523 and KL budgets.

525 7 CONCLUSION

526 We present the first systematic study of generative reward models as online training signals. Our
 527 experiments reveal a consistent *evaluator-rewarder gap*: Thinking GenRMs outperform Answer-
 528 Only variants on static evaluation, yet this advantage reverses during policy optimization, where
 529 Answer-Only models yield higher Gold Elo and more stable alignment.

530 Three findings offer practical guidance: (1) reward model scale dominates other factors, with gains
 531 continuing even when the GenRM far exceeds the policy in size; (2) intermediate GenRM check-
 532 points can outperform fully-trained models as rewarders; and (3) in-distribution accuracy does not
 533 predict training effectiveness. These results suggest that optimizing GenRMs for static benchmarks
 534 may be counterproductive for downstream alignment.

540 REFERENCES

541

542 Wei-Lin Chiang, Lianmin Zheng, et al. Chatbot arena: An open platform for evaluating llms by hu-
543 man preference. *arXiv:2403.04132*, 2024. URL <https://arxiv.org/abs/2403.04132>.

544 John Joon Young Chung, Vishakh Padmakumar, Melissa Roemmele, Yuqian Sun, and Max
545 Kreminski. Modifying large language model post-training for diverse creative writing. *arXiv*
546 *preprint arXiv:2503.17126*, 2025. URL <https://arxiv.org/abs/2503.17126>.

547

548 Daniel Fein, Sebastian Russo, Violet Xiang, Kabir Jolly, Rafael Rafailov, and Nick Haber. Lit-
549 bench: A benchmark and dataset for reliable evaluation of creative writing. *arXiv preprint*
550 *arXiv:2507.00769*, 2025. doi: 10.48550/arXiv.2507.00769. URL <https://arxiv.org/abs/2507.00769>.

551

552 Kanishk Gandhi, Ayush Chakravarthy, Anikait Singh, Nathan Lile, and Noah D. Goodman. Cogni-
553 tive behaviors that enable self-improving reasoners, or, four habits of highly effective STaRs.
554 *arXiv preprint arXiv:2503.01307*, 2025. doi: 10.48550/arXiv.2503.01307. URL <https://arxiv.org/abs/2503.01307>.

555

556 Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. *arXiv*
557 *preprint arXiv:2210.10760*, 2022. doi: 10.48550/arXiv.2210.10760. URL <https://arxiv.org/abs/2210.10760>.

558

559

560 Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre
561 Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, Johan Ferret, and Mathieu Blondel. Direct lan-
562 guage model alignment from online AI feedback. *arXiv preprint arXiv:2402.04792*, 2024. doi:
563 10.48550/arXiv.2402.04792. URL <https://arxiv.org/abs/2402.04792>.

564

565 Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu,
566 Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. Re-
567 wardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*,
568 2024. URL <https://arxiv.org/abs/2403.13787>.

569

570 Xiaobo Liang, Haoke Zhang, Juntao Li, Kehai Chen, Qiaoming Zhu, and Min Zhang. Generative
571 reward modeling via synthetic criteria preference learning. In *Proceedings of the 63rd Annual*
572 *Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26755–
573 26769, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1297. URL <https://aclanthology.org/2025.acl-long.1297>.

574

575 Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu.
576 Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495*, 2025.
577 doi: 10.48550/arXiv.2504.02495. URL <https://arxiv.org/abs/2504.02495>.

578

579 Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castricato,
580 Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. Generative reward models. *arXiv preprint*
581 *arXiv:2410.12832*, 2024. doi: 10.48550/arXiv.2410.12832. URL <https://arxiv.org/abs/2410.12832>.

582

583 Long Ouyang et al. Training language models to follow instructions with human feedback. *arXiv*
584 *preprint arXiv:2203.02155*, 2022. URL <https://arxiv.org/abs/2203.02155>.

585

586 Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and
587 Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model.
588 *arXiv preprint arXiv:2305.18290*, 2023. URL <https://arxiv.org/abs/2305.18290>.

589

590 Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W. Bradley Knox,
591 Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct align-
592 ment algorithms. *arXiv preprint arXiv:2406.02900*, 2024. doi: 10.48550/arXiv.2406.02900. URL
593 <https://arxiv.org/abs/2406.02900>. NeurIPS 2024.

594

595 Swarnadeep Saha, Xian Li, Marjan Ghazvininejad, Jason Weston, and Tianlu Wang. Learning to
596 plan & reason for evaluation with thinking-llm-as-a-judge. *arXiv preprint arXiv:2501.18099*,
597 2025. URL <https://arxiv.org/abs/2501.18099>. Introduces FollowBenchEval.

594 Wenlei Shi and Xing Jin. Heimdall: Test-time scaling on the generative verification. *arXiv preprint*
 595 *arXiv:2504.10337*, 2025. doi: 10.48550/arXiv.2504.10337. URL <https://arxiv.org/abs/2504.10337>.

597

598 Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin,
 599 Ellie Evans, Yi Dong, and Oleksii Kuchaiev. Helpsteer3-preference: Open human-annotated
 600 preference data across diverse tasks and languages. *arXiv preprint arXiv:2505.11475*, 2025. doi:
 601 10.48550/arXiv.2505.11475. URL <https://arxiv.org/abs/2505.11475>.

602

603 Chenxi Whitehouse, Tianlu Wang, Ping Yu, Xian Li, Jason Weston, Ilia Kulikov, and Swarnadeep
 604 Saha. J1: Incentivizing thinking in llm-as-a-judge via reinforcement learning. *arXiv preprint*
 605 *arXiv:2505.10320*, 2025. URL <https://arxiv.org/abs/2505.10320>.

606

607 Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner
 608 Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. Justice or prejudice?
 609 quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*, 2024. URL <https://arxiv.org/abs/2410.02736>.

610

611 Yue Yu, Zhengxing Chen, Aston Zhang, Liang Tan, Chenguang Zhu, Richard Yuanzhe Pang, Yundi
 612 Qian, Xuewei Wang, Suchin Gururangan, Chao Zhang, Melanie Kambadur, Dhruv Mahajan, and
 613 Rui Hou. Self-generated critiques boost reward modeling for language models. *arXiv preprint*
 614 *arXiv:2411.16646*, 2025. doi: 10.48550/arXiv.2411.16646. URL <https://arxiv.org/abs/2411.16646>.

615 Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal.
 616 Generative verifiers: Reward modeling as next-token prediction. In *ICLR*, 2025. URL <https://arxiv.org/abs/2408.15240>. ICLR 2025.

617

618

619

620 **A APPENDIX**

621

622 Additional experimental details and supplementary results.

623

624 **A.1 PRELIMINARY CROSS-DOMAIN EVALUATION: OPENORCA**

625

626 To provide preliminary evidence on domain generalization (see §5), we ran cross-domain checks on
 627 OpenOrca using off-the-shelf Qwen3 policies, judges, and Gold evaluator. No fine-tuning was per-
 628 formed. These results support the same qualitative pre-training trend observed in the main creative-
 629 writing domain: prior to GenRM training, Thinking judges show stronger alignment with the Gold
 630 evaluator than Answer-Only judges.

631

632 **A.2 ELO RATING COMPUTATION**

633

634 Given a set of policy checkpoints \mathcal{S} , we generate pairwise matches $(i, j) \in \mathcal{S} \times \mathcal{S}$ on held-out
 635 prompts and obtain Gold evaluator decisions $w_{ij} \in \{0, 1\}$. Elo ratings $\{R_s\}_{s \in \mathcal{S}}$ are estimated by
 636 maximizing the logistic likelihood:

637

638

$$\max_{\{R_s\}} \sum_{(i,j)} \left[w_{ij} \log \sigma\left(\frac{R_i - R_j}{s}\right) + (1 - w_{ij}) \log \sigma\left(\frac{R_j - R_i}{s}\right) \right],$$

639 with scale s fixed. Proxy Elo ratings are computed analogously using proxy GenRM decisions in
 640 place of Gold decisions.

641

642

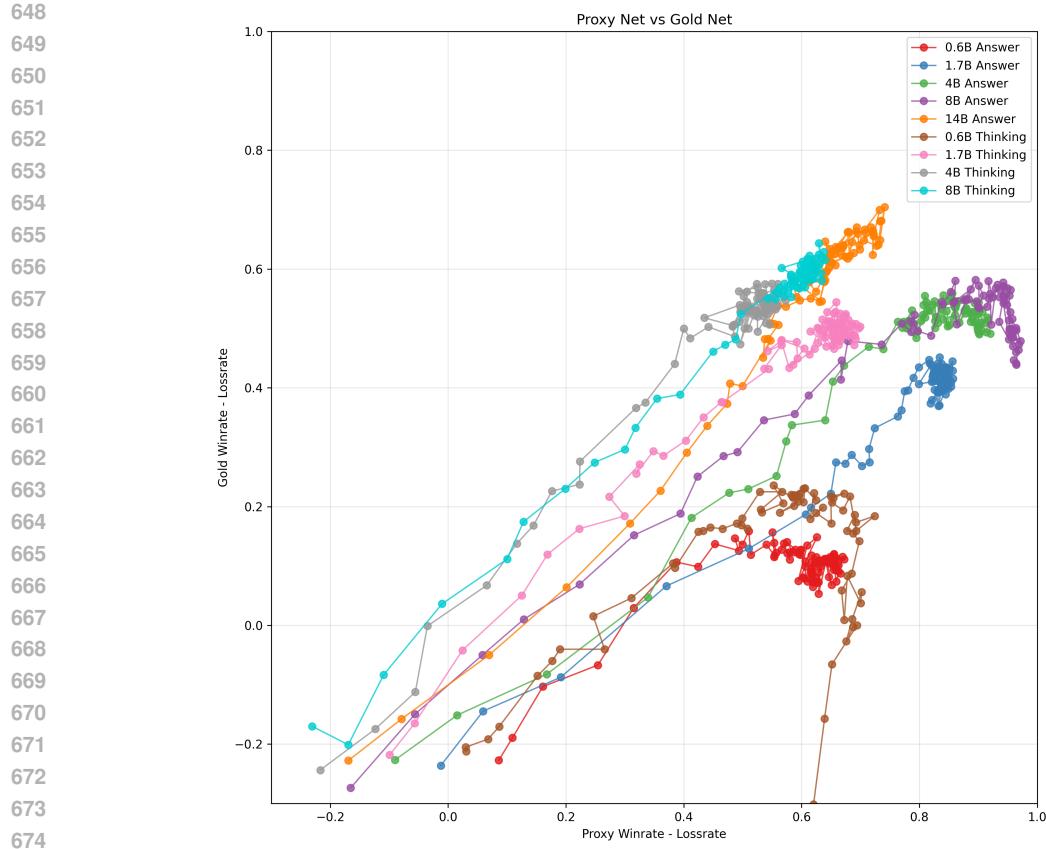
643

644

645

646

647



675 Figure 7: Proxy vs. Gold net score on OpenOrca. All models are off-the-shelf (no GenRM training). A full
 676 replication with trained GenRMs and policy optimization remains future work.
 677

679 B PROMPT TEMPLATES

681 We provide the exact prompt templates used for policy generation and GenRM pairwise evaluation.
 682 All models receive these prompts without modification.
 683

684 B.1 POLICY PROMPT

686 The policy model receives the following prompt template for creative writing generation:

688 Creative Writing Prompt:
 689 {QUESTION}
 690 Goal: write an original short story that directly answers the prompt.
 691 Requirements: Output only the story text. no titles, notes, commentary
 692 or meta-text.

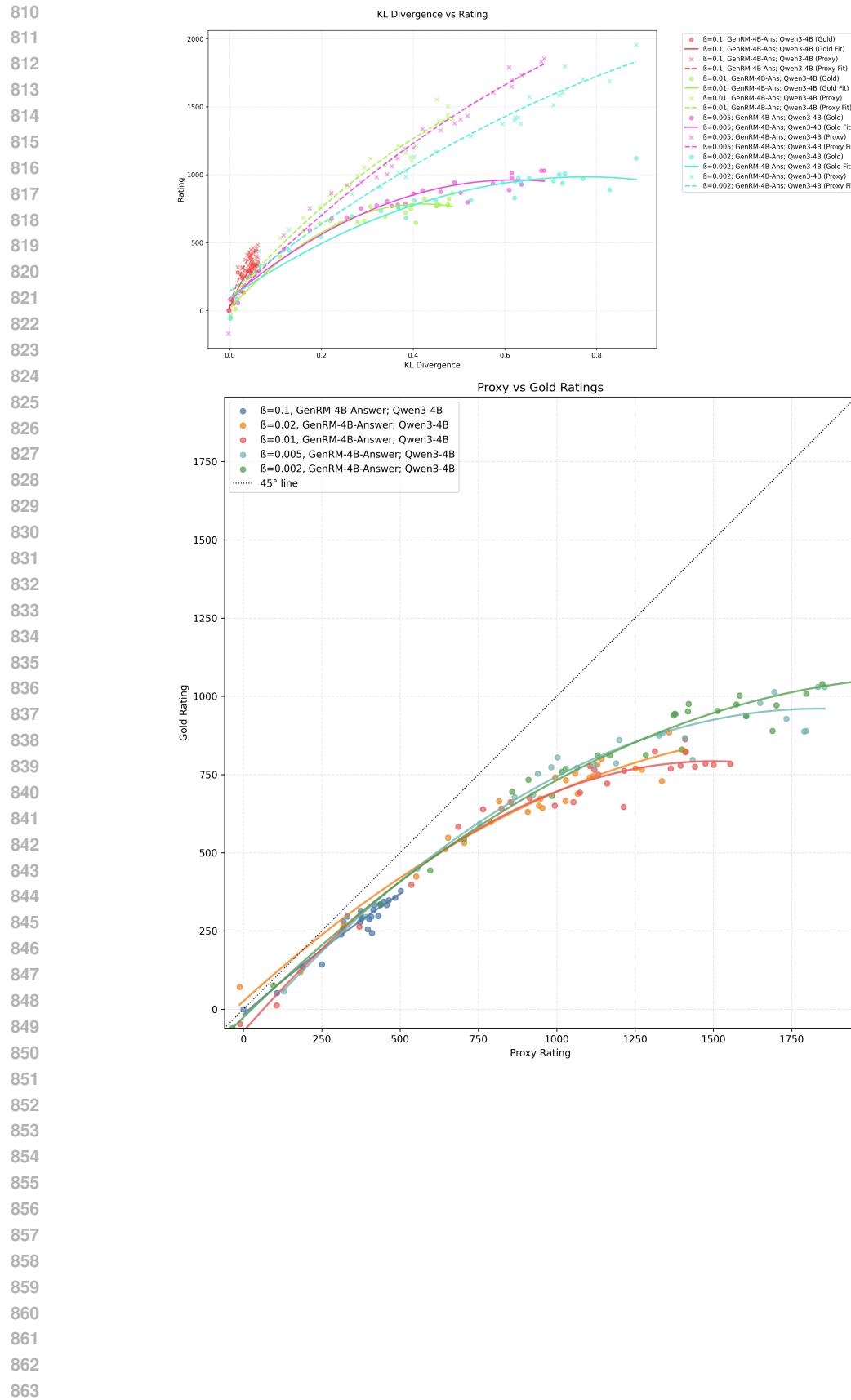
693 where {QUESTION} is replaced with the specific creative writing instruction from the evaluation
 694 set.
 695

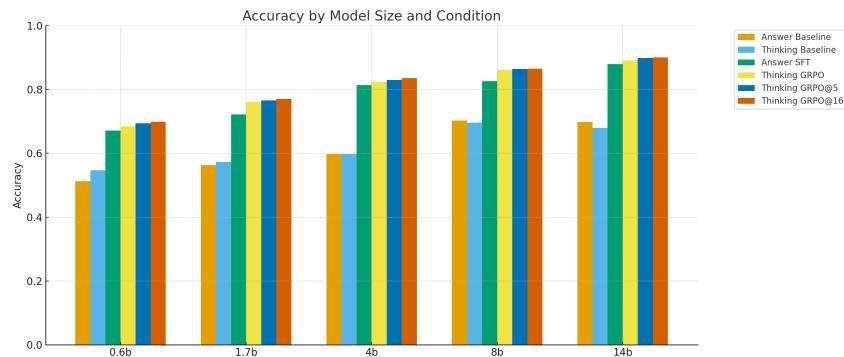
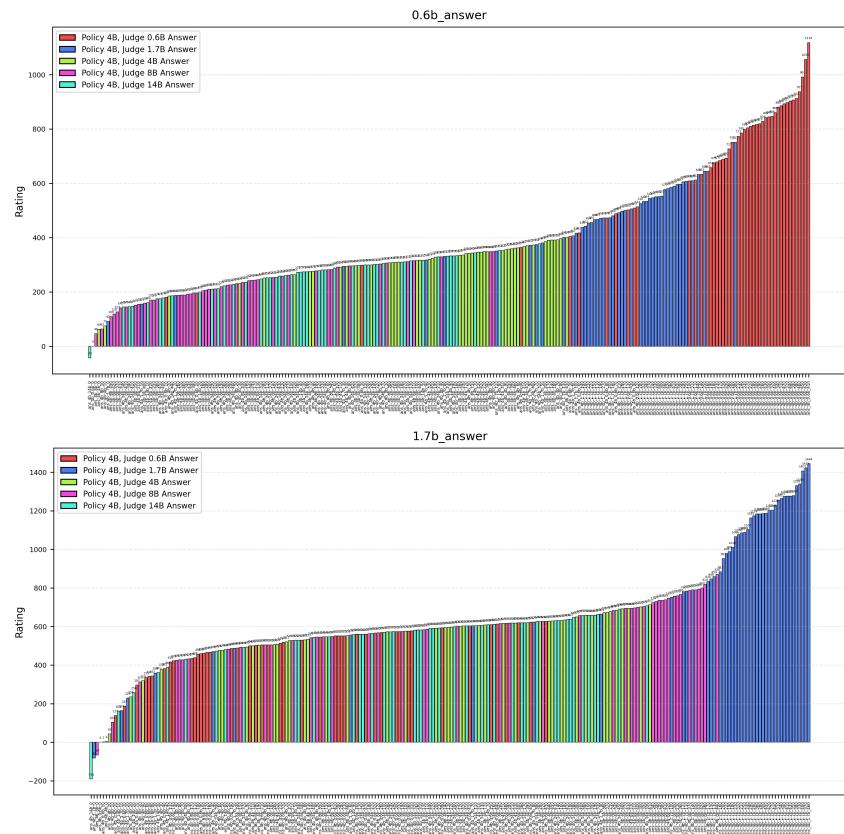
696 B.2 GENRM PAIRWISE EVALUATION PROMPT

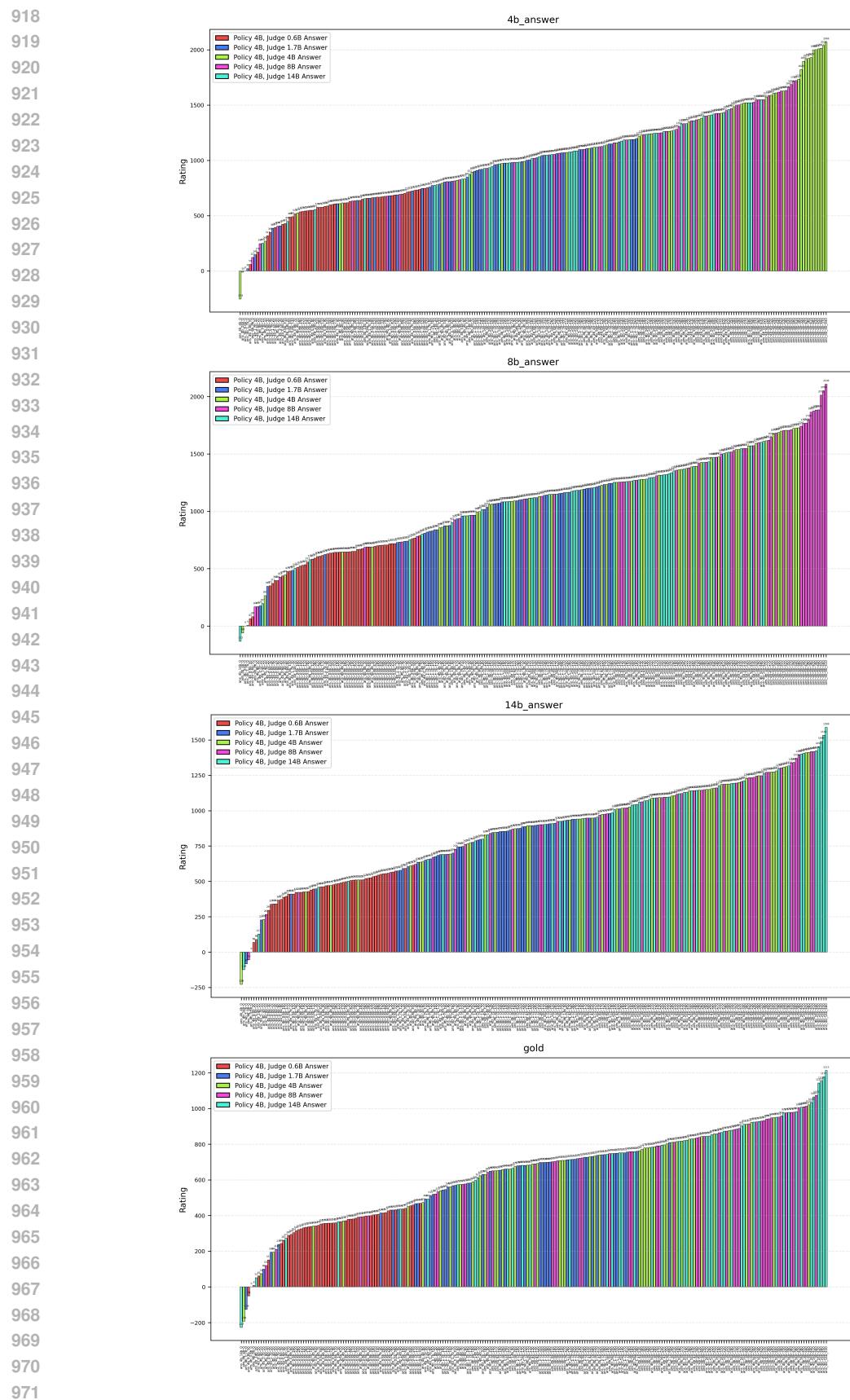
698 The generative reward model receives the following prompt template for pairwise comparison:

699
 700 Two AI models were given the same instruction and each produced a reply.
 701 Your task is to judge which reply better fulfills the instruction.

```
702 [BEGIN INSTRUCTIONS GIVEN TO BOTH MODELS]  
703 {POLICY_PROMPT}  
704 [END INSTRUCTIONS]  
705  
706 [BEGIN REPLY A]  
707 {RESPONSE_A}  
708 [END REPLY A]  
709  
710 [BEGIN REPLY B]  
711 {RESPONSE_B}  
712 [END REPLY B]  
713 Output ONLY one of:  
714 <answer>[ [A] ]</answer>  
715 <answer>[ [B] ]</answer>  
716  
717 where {POLICY_PROMPT} contains the full policy prompt with the specific question, and  
718 {RESPONSE_A} and {RESPONSE_B} contain the two candidate responses to be compared. The  
719 model outputs a structured verdict indicating which response better satisfies the original instruction.  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755
```

864 C.2 GENRM EVALUATION
865
866
867
868880
881
882
883
884
885
886
887 C.3 GENRM SIZES (ANSWER)
888
889
890
891
892



972

973

974

975

976

977

978

979

980

981

982

983

984

985

986

987

988

989

990

991

992

993

994

995

996

997

998

999

1000

1001

1002

1003

1004

1005

1006

1007

1008

1009

1010

1011

1012

1013

1014

1015

1016

1017

1018

1019

1020

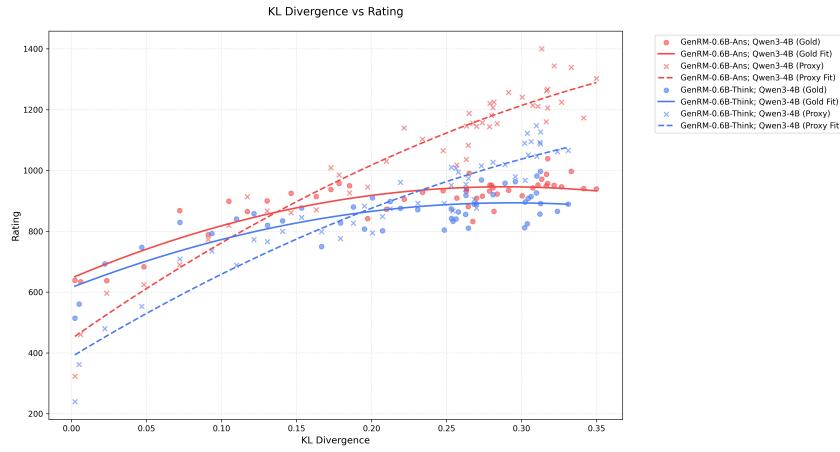
1021

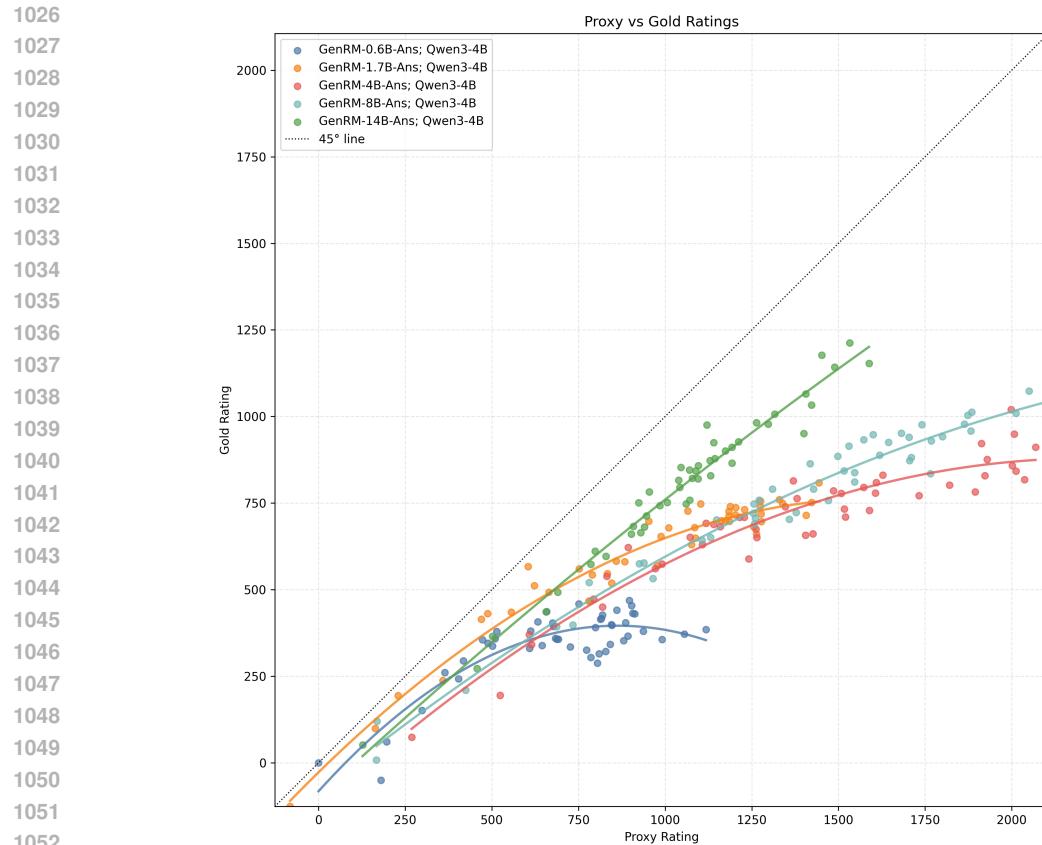
1022

1023

1024

1025





1080
1081

C.4 GENRM SIZES (THINKING)

1082
1083

1084

1085

1086

1087

1088

1089

1090

1091

1092

1093

1094

1095

1096

1097

1098

1099

1100

1101

1102

1103

1104

1105

1106

1107

1108

1109

1110

1111

1112

1113

1114

1115

1116

1117

1118

1119

1120

1121

1122

1123

1124

1125

1126

1127

1128

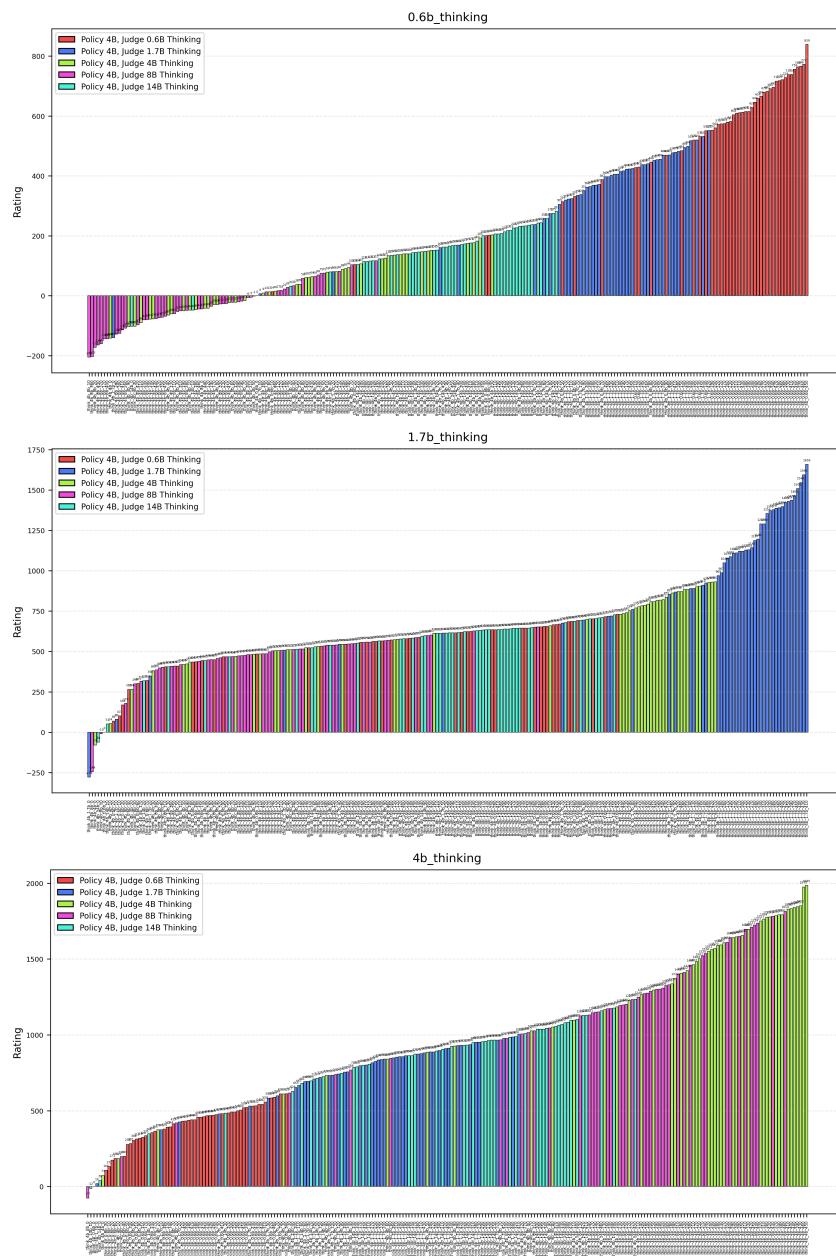
1129

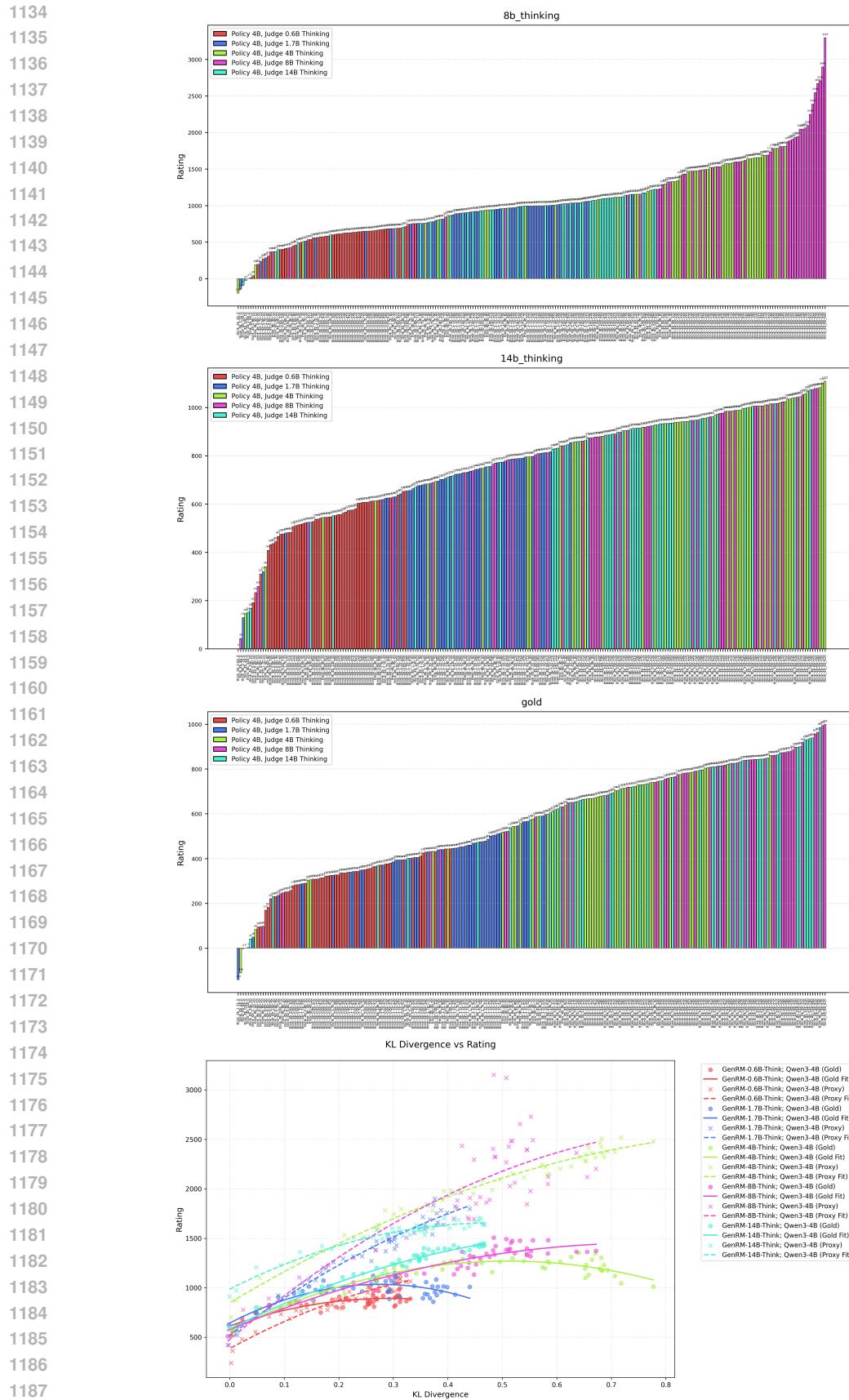
1130

1131

1132

1133





1188

1189

1190

1191

1192

1193

1194

1195

1196

1197

1198

1199

1200

1201

1202

1203

1204

1205

1206

1207

1208

1209

1210

1211

1212

1213

1214

1215

1216

1217

1218

1219

1220

1221

1222

1223

1224

1225

1226

1227

1228

1229

1230

1231

1232

1233

1234

1235

1236

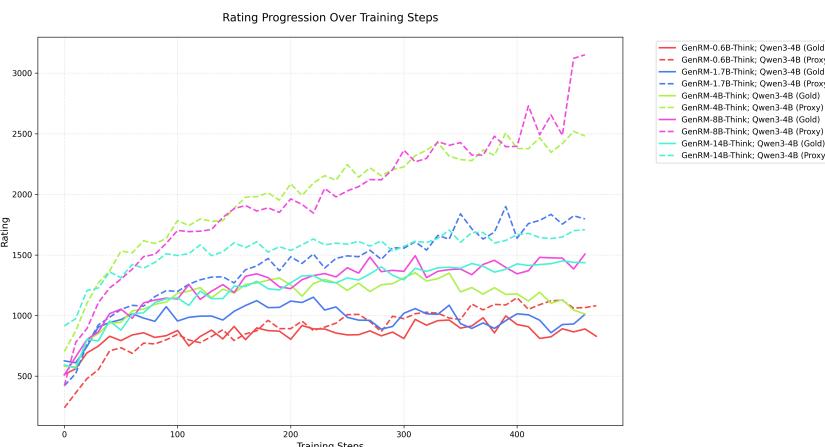
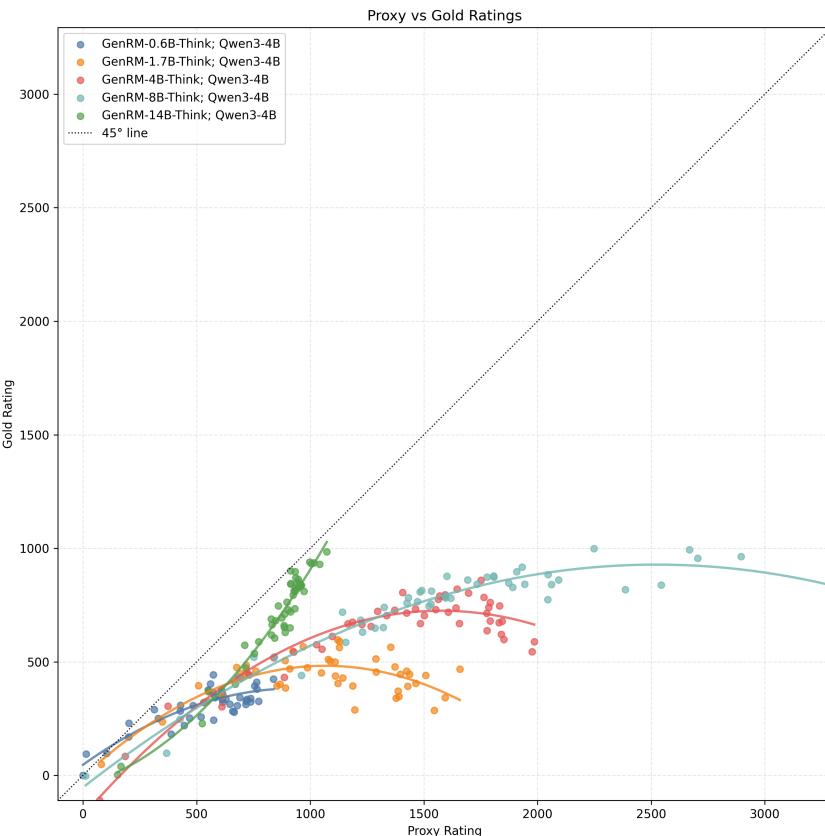
1237

1238

1239

1240

1241



1242 C.5 PARAM BALANCE (ANSWER)
1243

1244

1245

1246

1247

1248

1249

1250

1251

1252

1253

1254

1255

1256

1257

1258

1259

1260

1261

1262

1263

1264

1265

1266

1267

1268

1269

1270

1271

1272

1273

1274

1275

1276

1277

1278

1279

1280

1281

1282

1283

1284

1285

1286

1287

1288

1289

1290

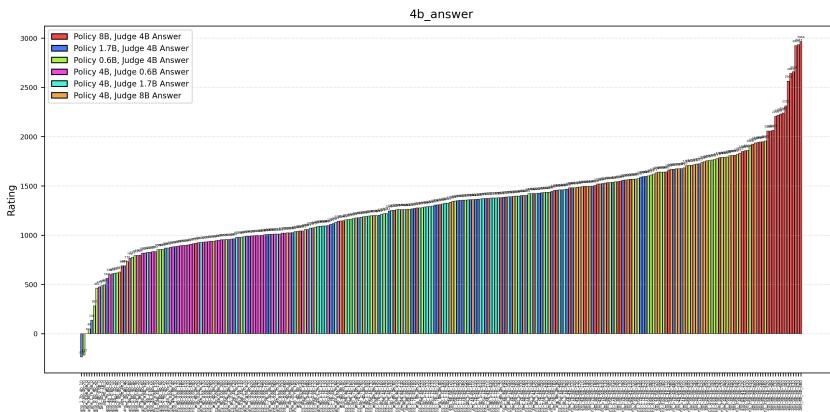
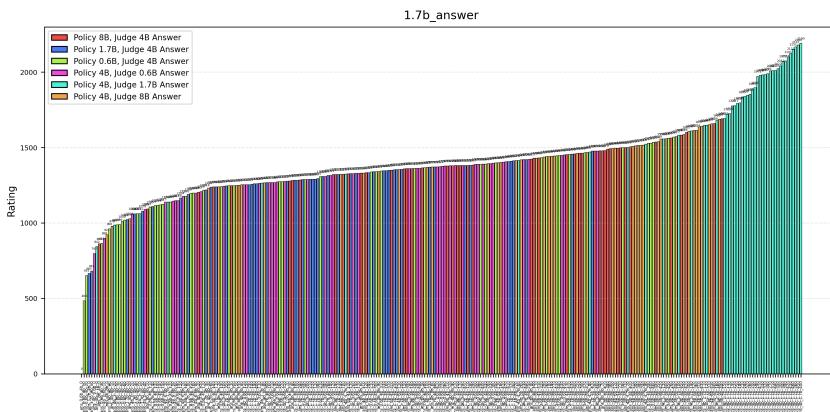
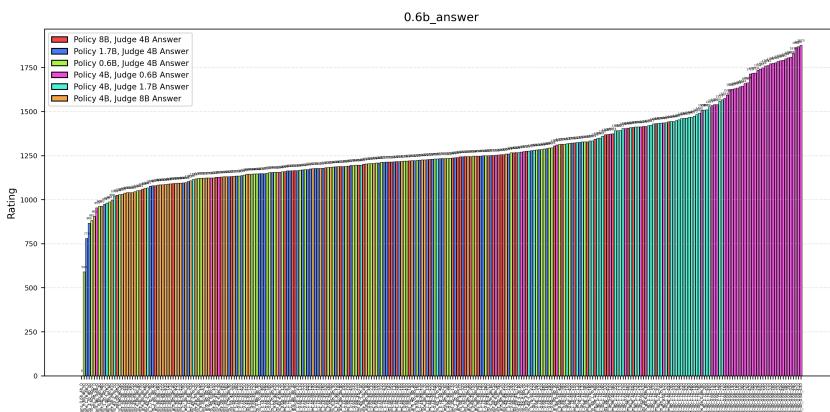
1291

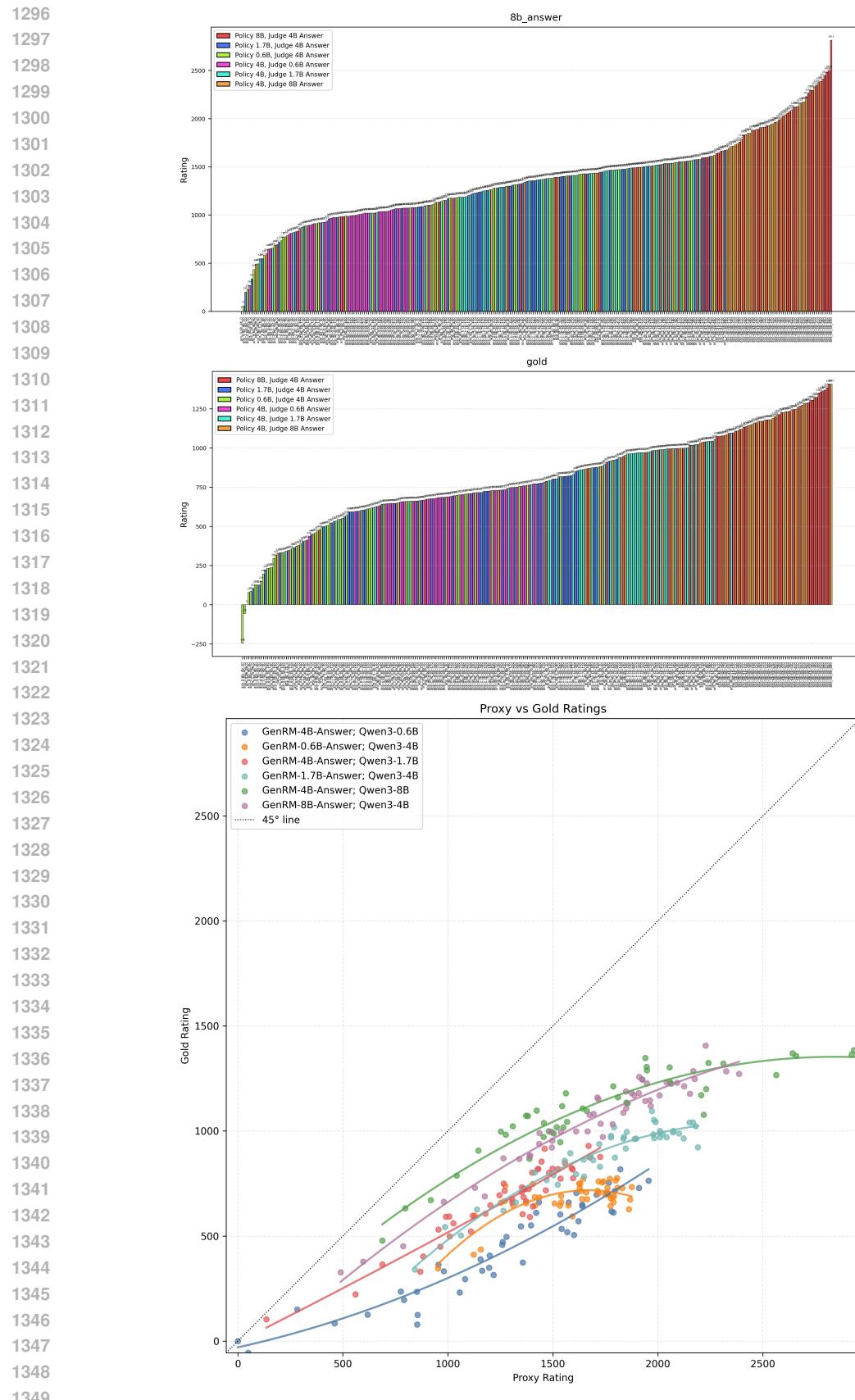
1292

1293

1294

1295





1350 C.6 PARAM BALANCE (THINKING)
1351

1352

1353

1354

1355

1356

1357

1358

1359

1360

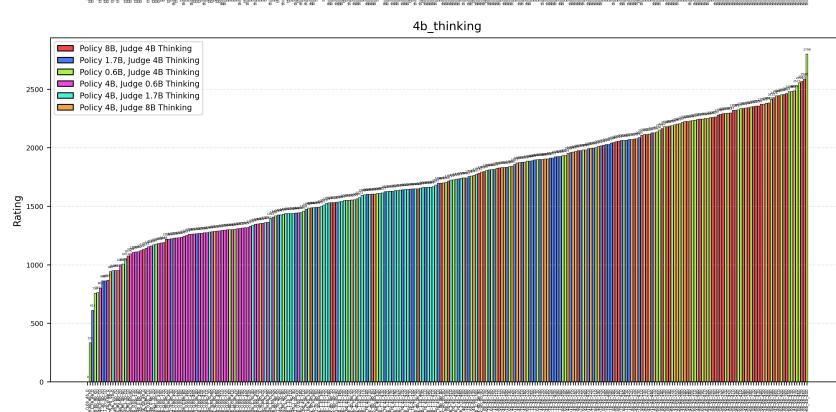
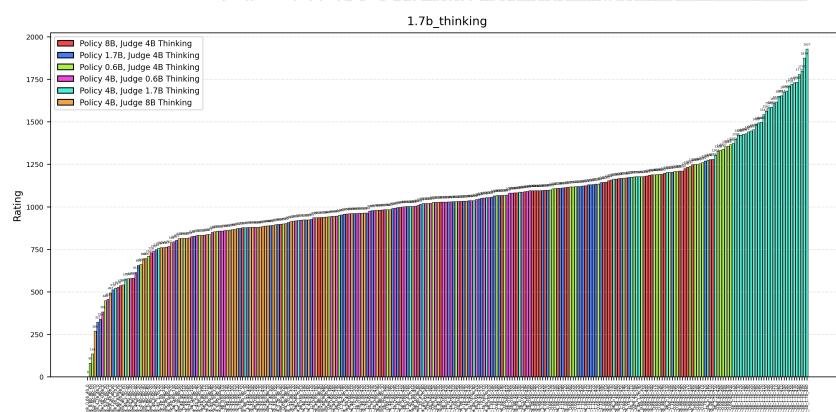
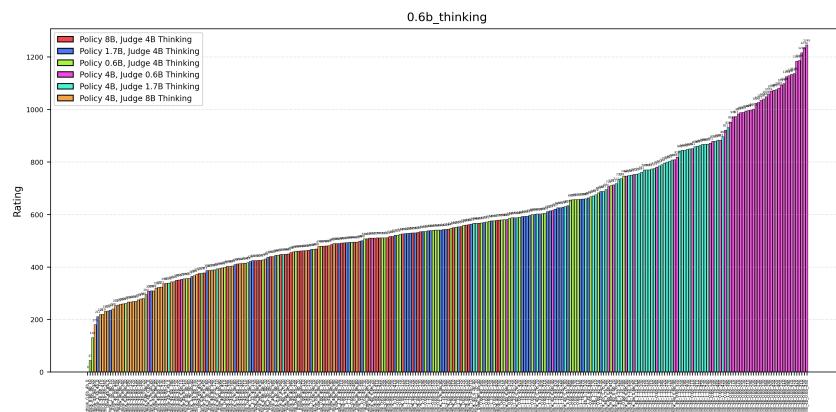
1361

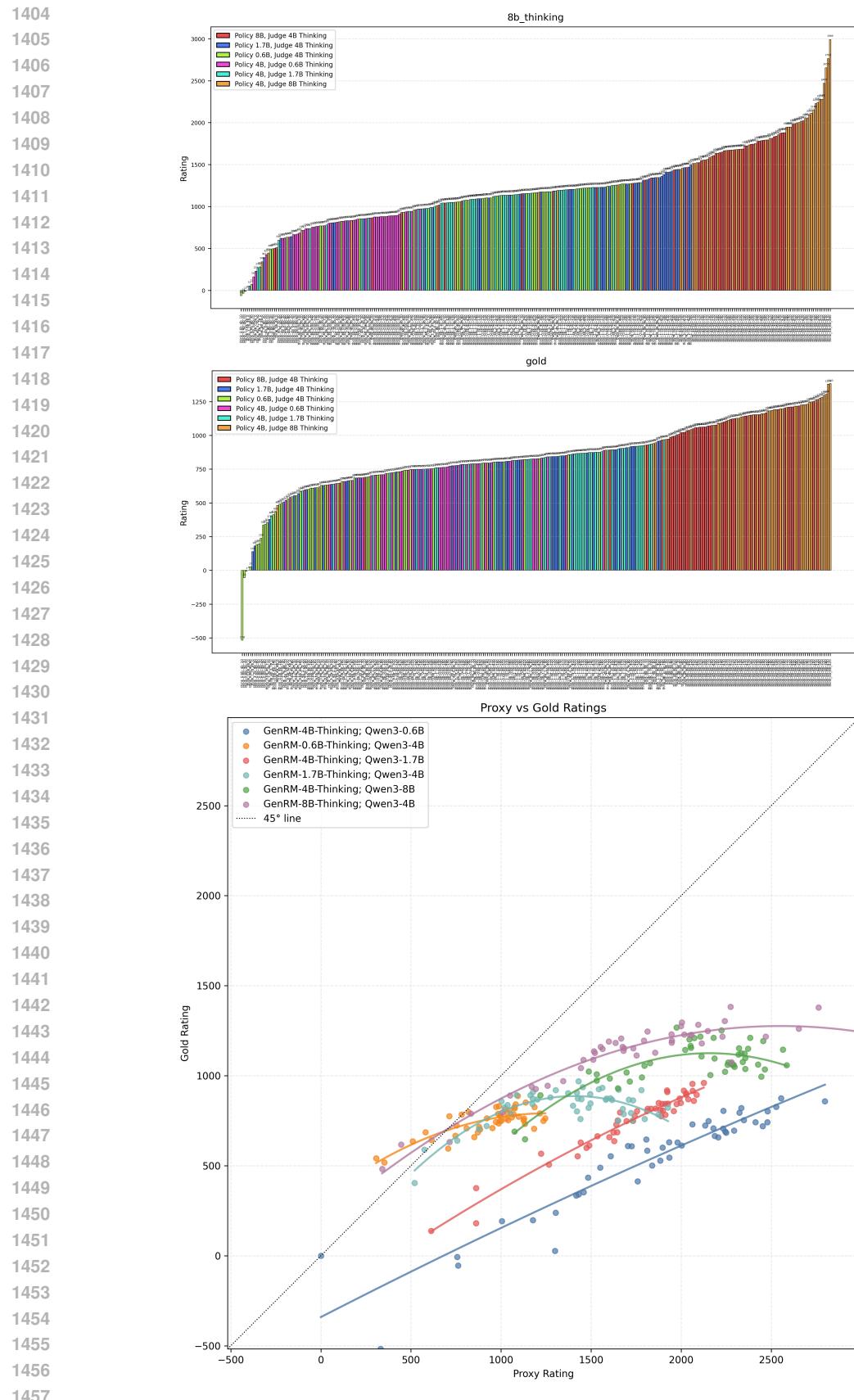
1362

1363

1364

1365





1458 C.7 POLICY SIZES (ANSWER)
1459

1460

1461

1462

1463

1464

1465

1466

1467

1468

1469

1470

1471

1472

1473

1474

1475

1476

1477

1478

1479

1480

1481

1482

1483

1484

1485

1486

1487

1488

1489

1490

1491

1492

1493

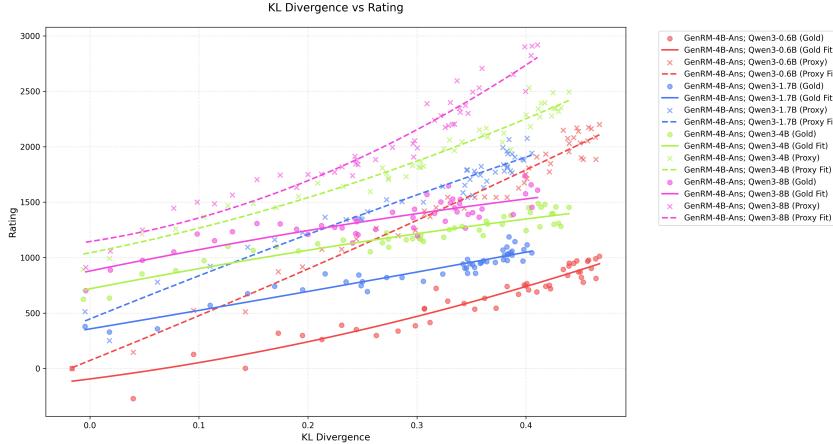
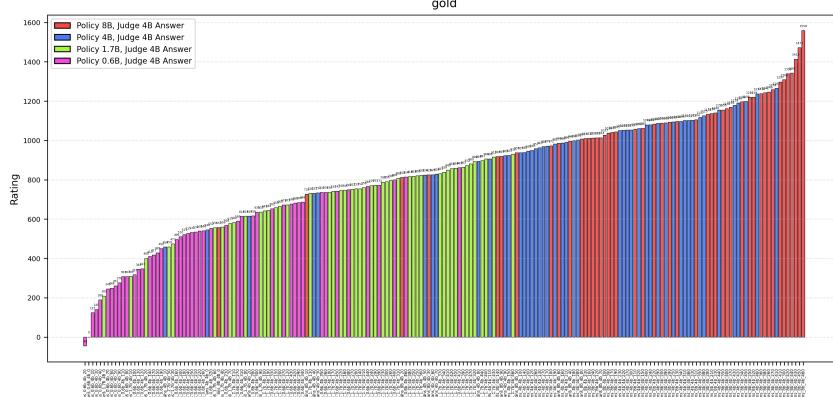
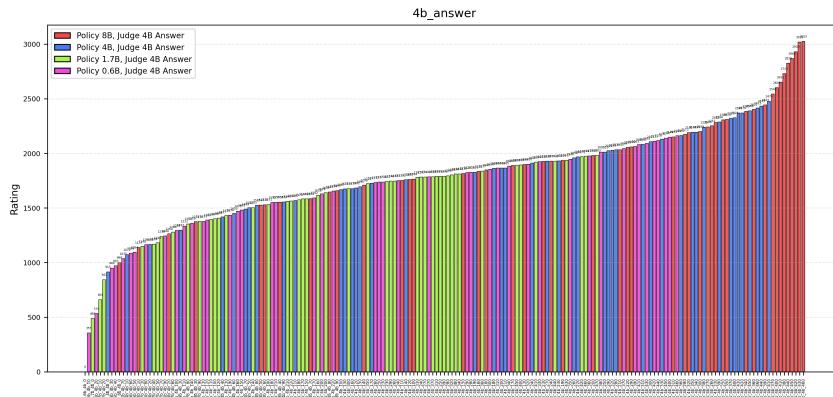
1494

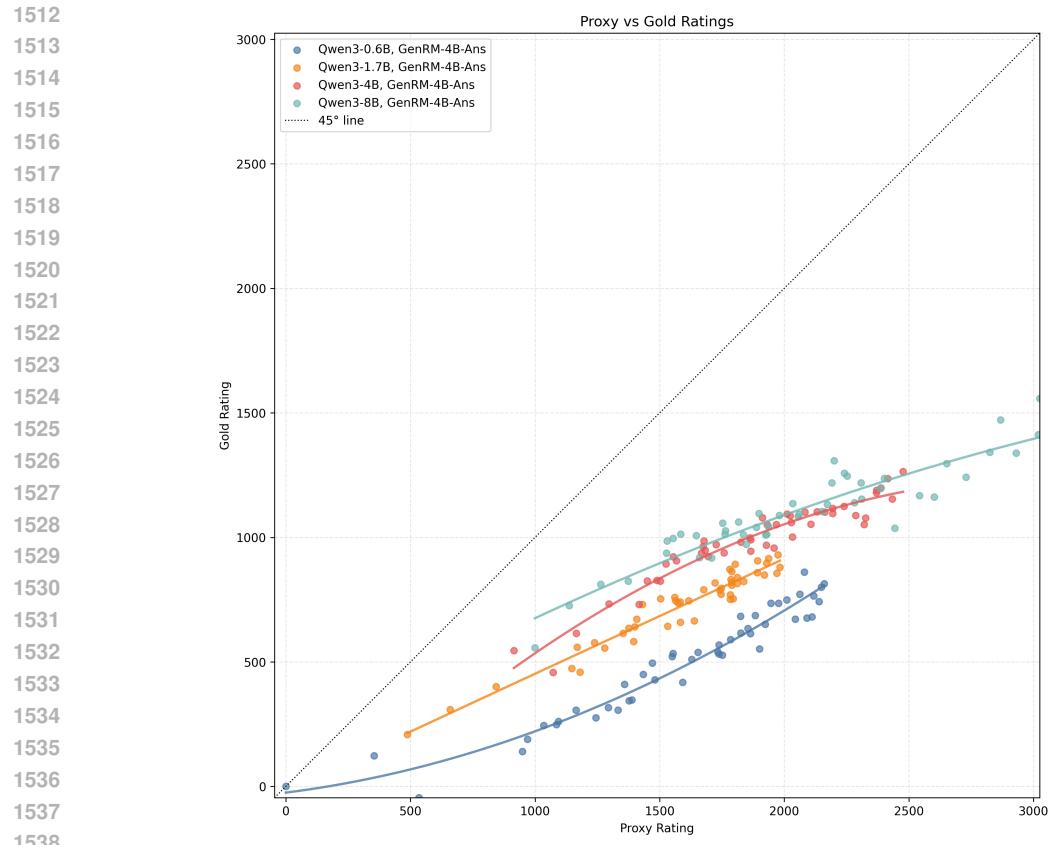
1495

1496

1497

1498





1566 C.8 POLICY SIZES (THINKING)
1567

1568

1569

1570

1571

1572

1573

1574

1575

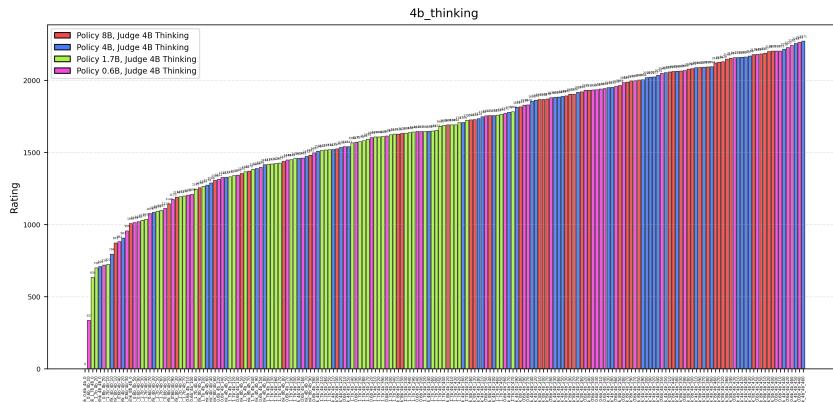
1576

1577

1578

1579

1580



1581

1582

1583

1584

1585

1586

1587

1588

1589

1590

1591

1592

1593



1594

1595

1596

1597

1598

1599

1600

1601

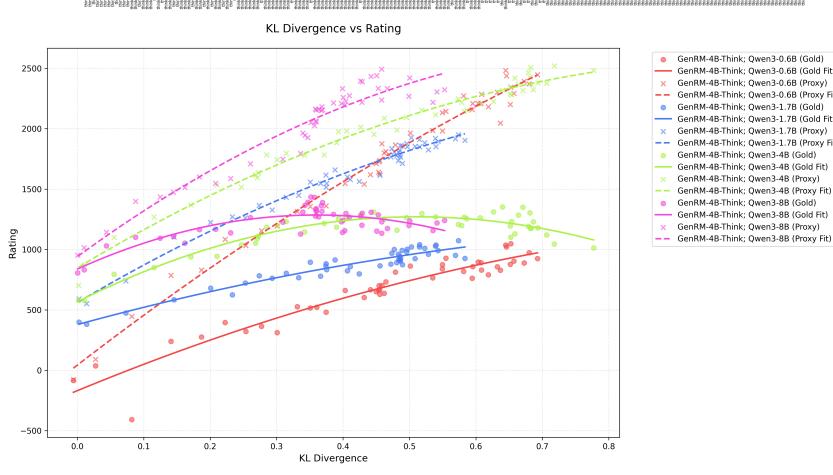
1602

1603

1604

1605

1606



1607

1608

1609

1610

1611

1612

1613

1614

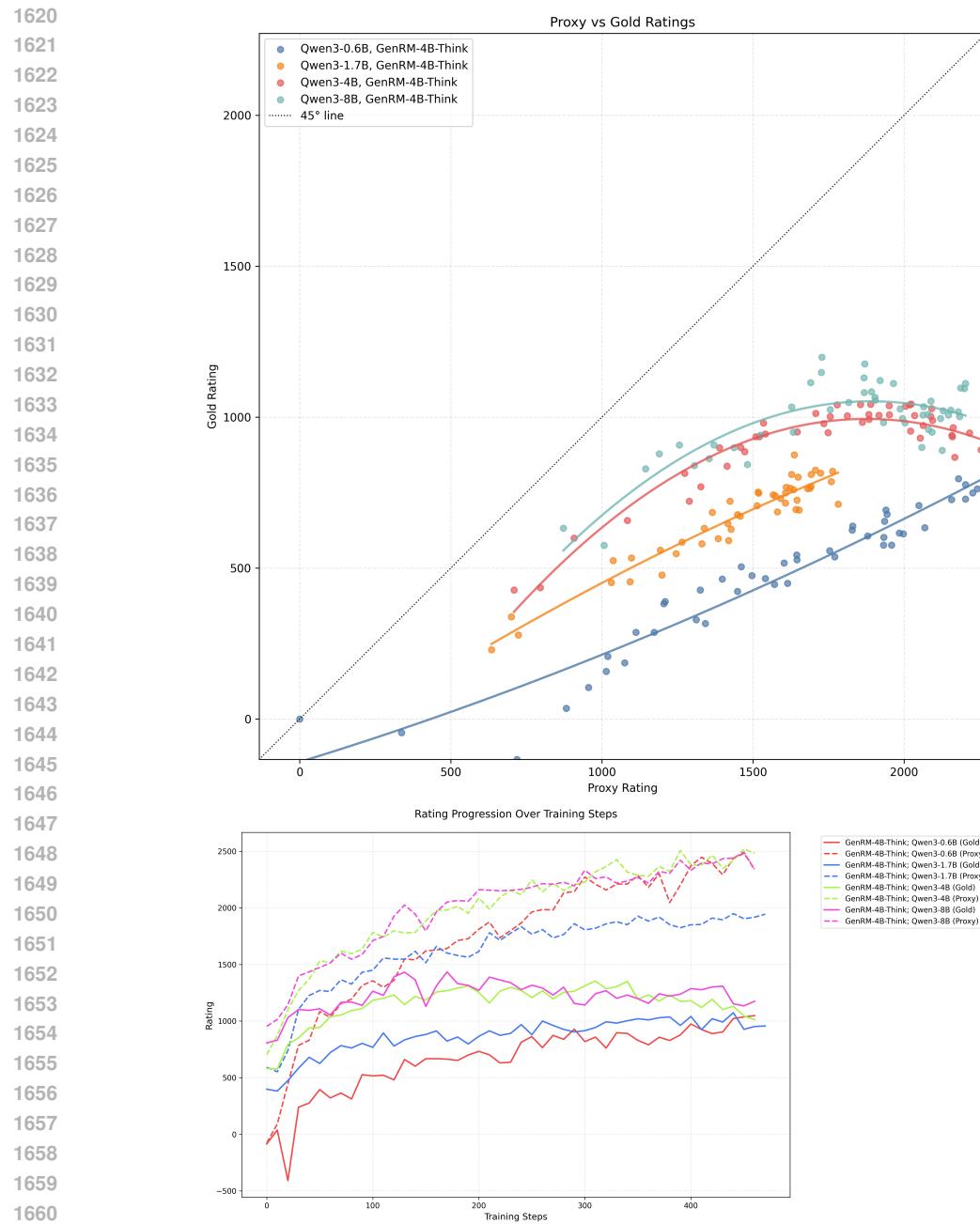
1615

1616

1617

1618

1619



1674 C.9 THINKING VS ANSWER (0.6B VS 4B)
1675

1676

1677

1678

1679

1680

1681

1682

1683

1684

1685

1686

1687

1688

1689

1690

1691

1692

1693

1694

1695

1696

1697

1698

1699

1700

1701

1702

1703

1704

1705

1706

1707

1708

1709

1710

1711

1712

1713

1714

1715

1716

1717

1718

1719

1720

1721

1722

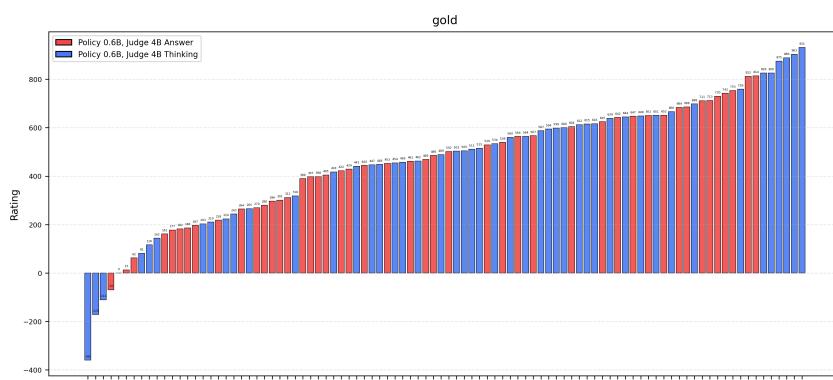
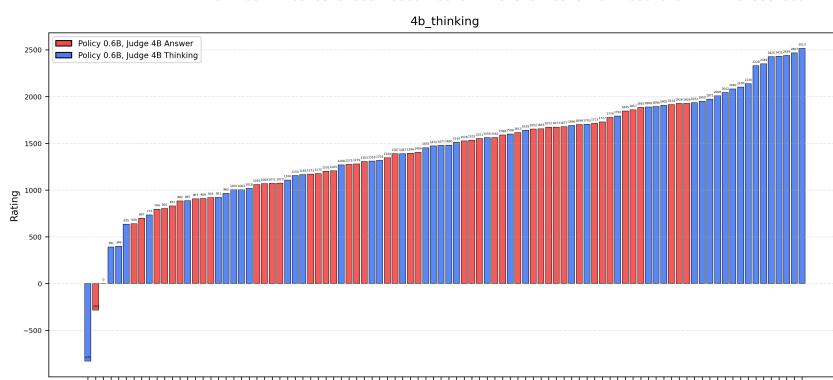
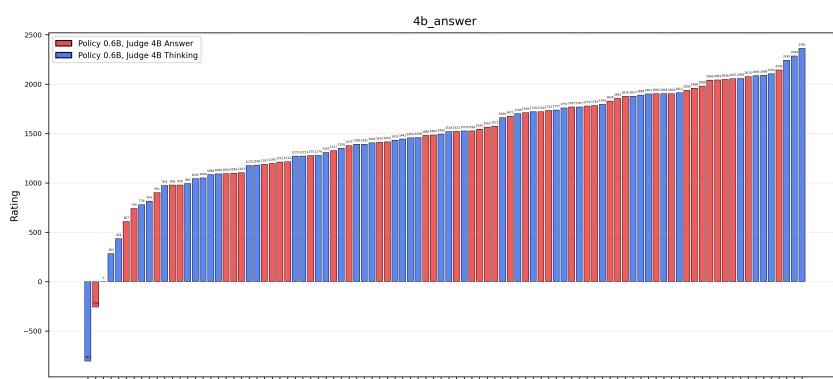
1723

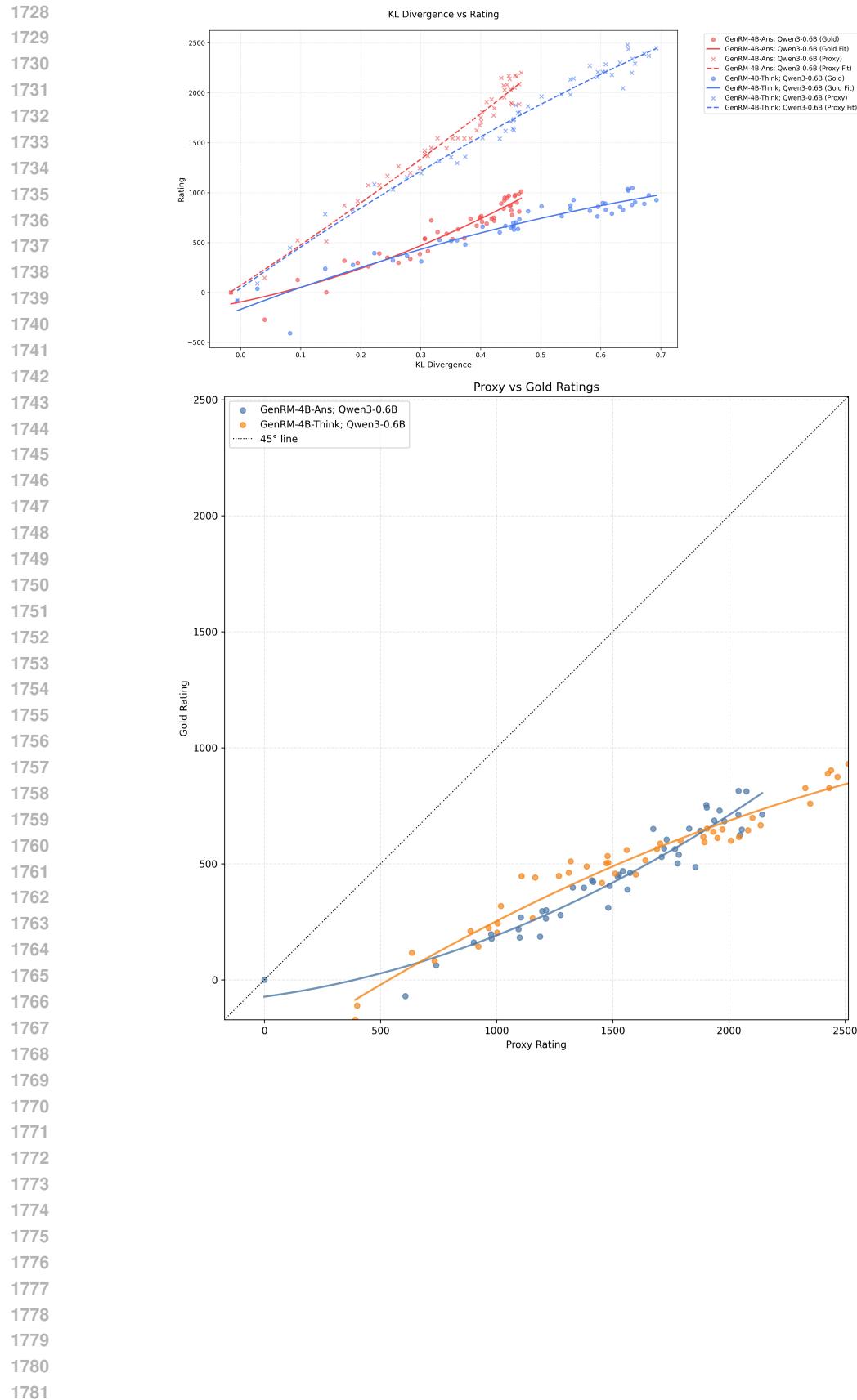
1724

1725

1726

1727





1782
1783

C.10 THINKING VS ANSWER (1.7B VS 4B)

1784
1785

1786

1787

1788

1789

1790

1791

1792

1793

1794

1795

1796

1797

1798

1799

1800

1801

1802

1803

1804

1805

1806

1807

1808

1809

1810

1811

1812

1813

1814

1815

1816

1817

1818

1819

1820

1821

1822

1823

1824

1825

1826

1827

1828

1829

1830

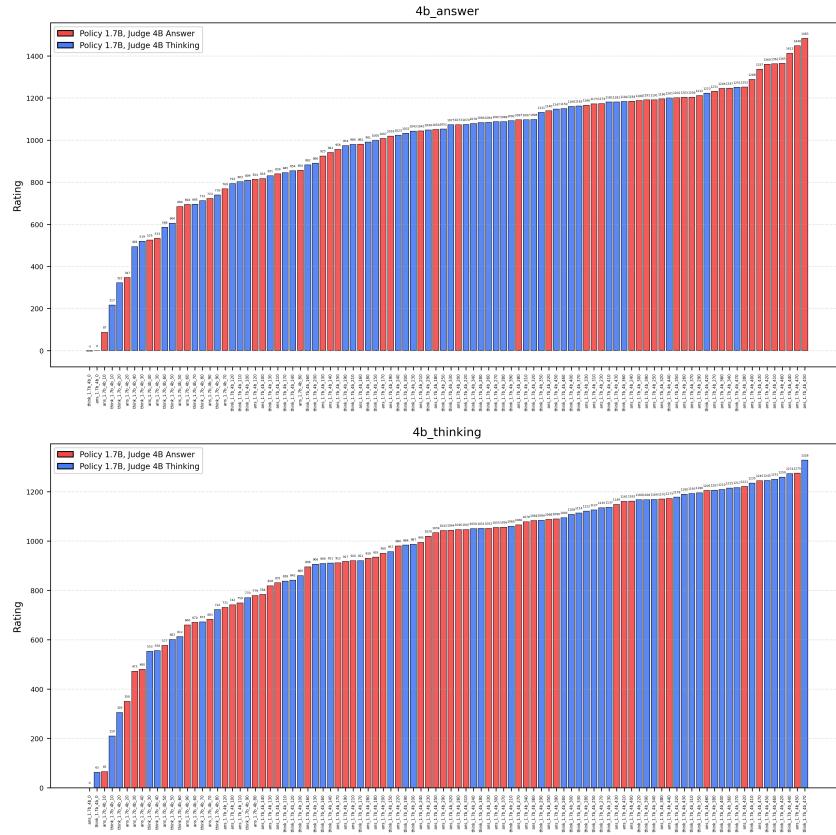
1831

1832

1833

1834

1835



1836 C.11 THINKING VS ANSWER (4B vs 0.6B)
1837

1838

1839

1840

1841

1842

1843

1844

1845

1846

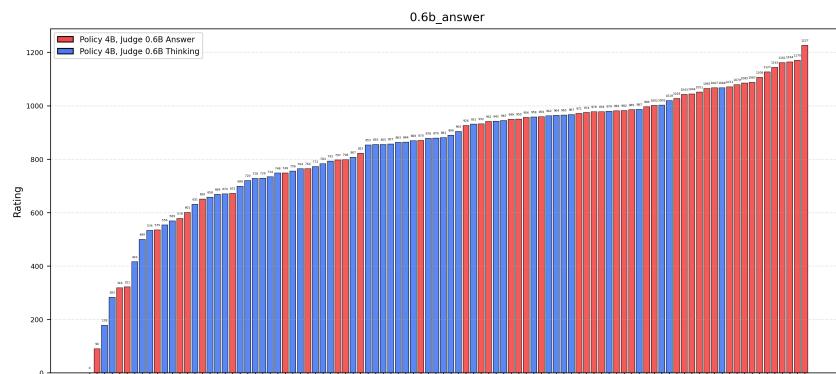
1847

1848

1849

1850

1851



1852

1853

1854

1855

1856

1857

1858

1859

1860

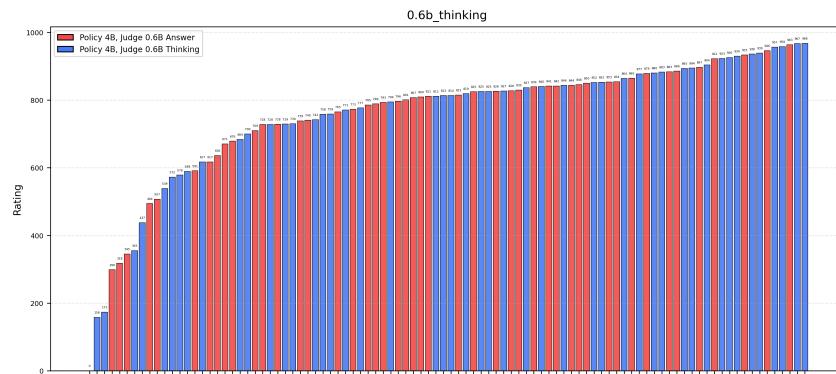
1861

1862

1863

1864

0.6b_thinking



1865

1866

1867

1868

1869

1870

1871

1872

1873

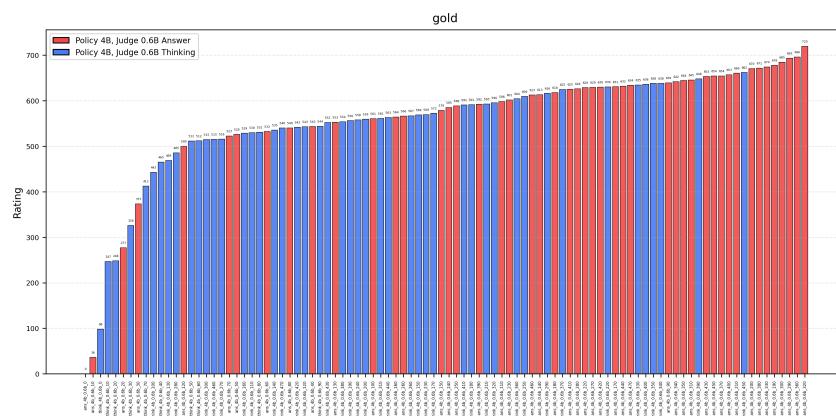
1874

1875

1876

1877

gold



1878

1879

1880

1881

1882

1883

1884

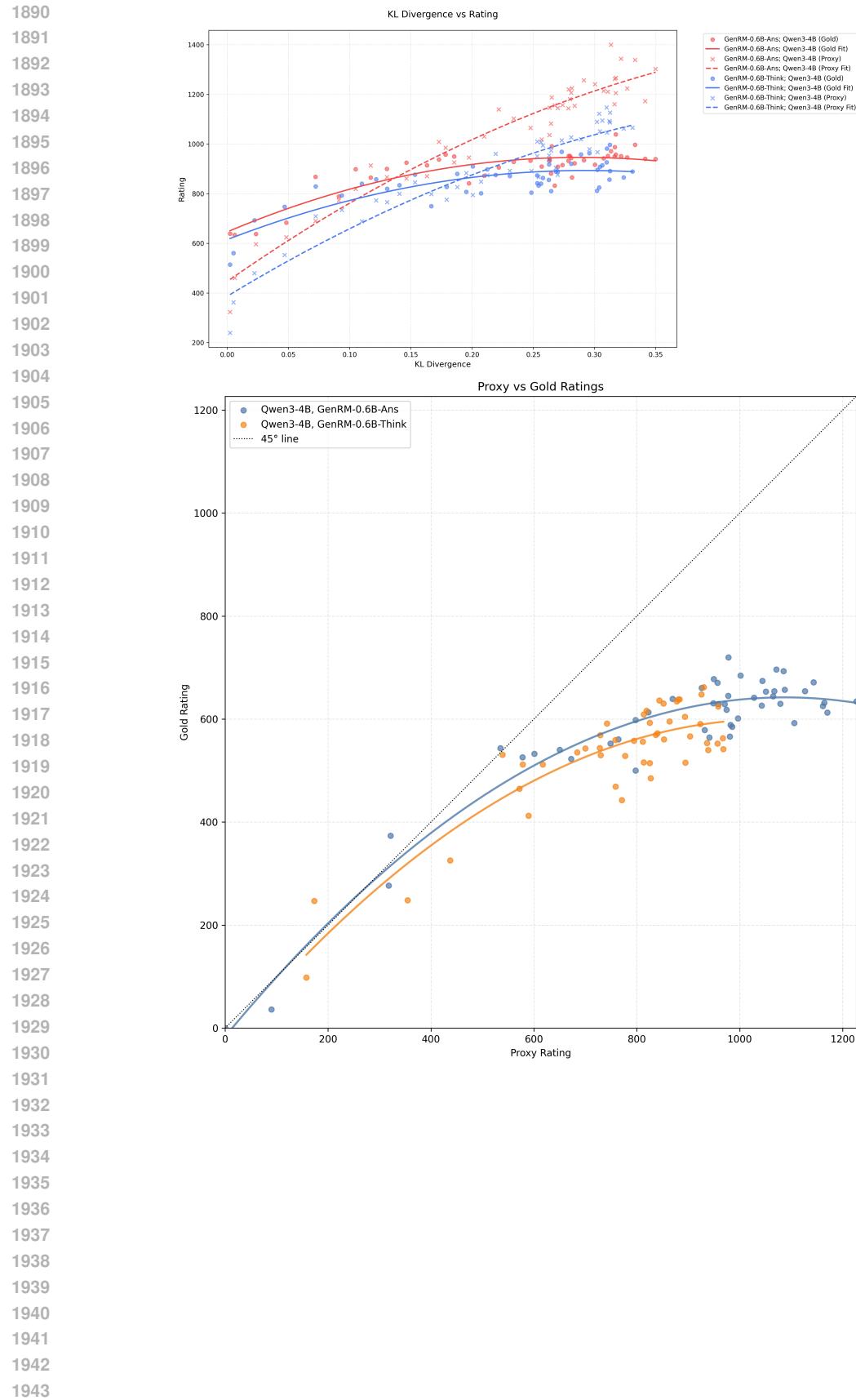
1885

1886

1887

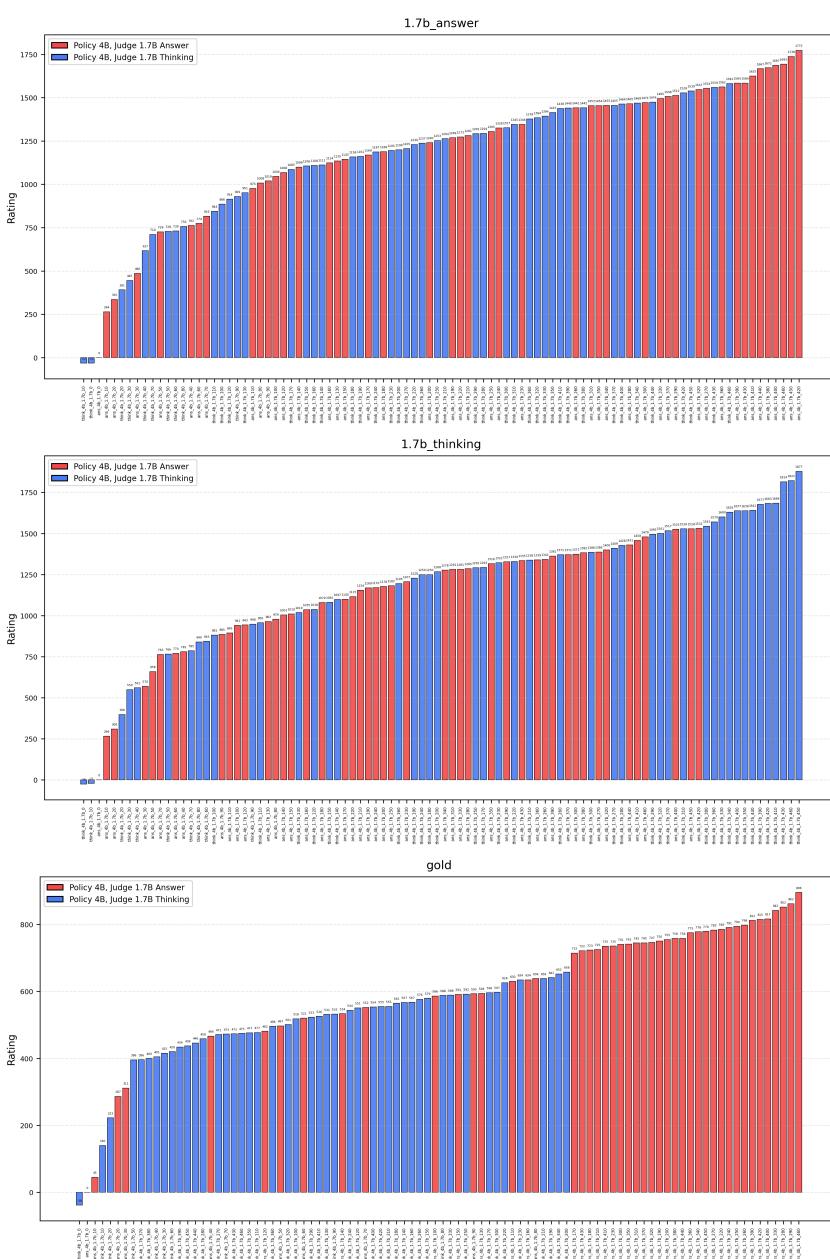
1888

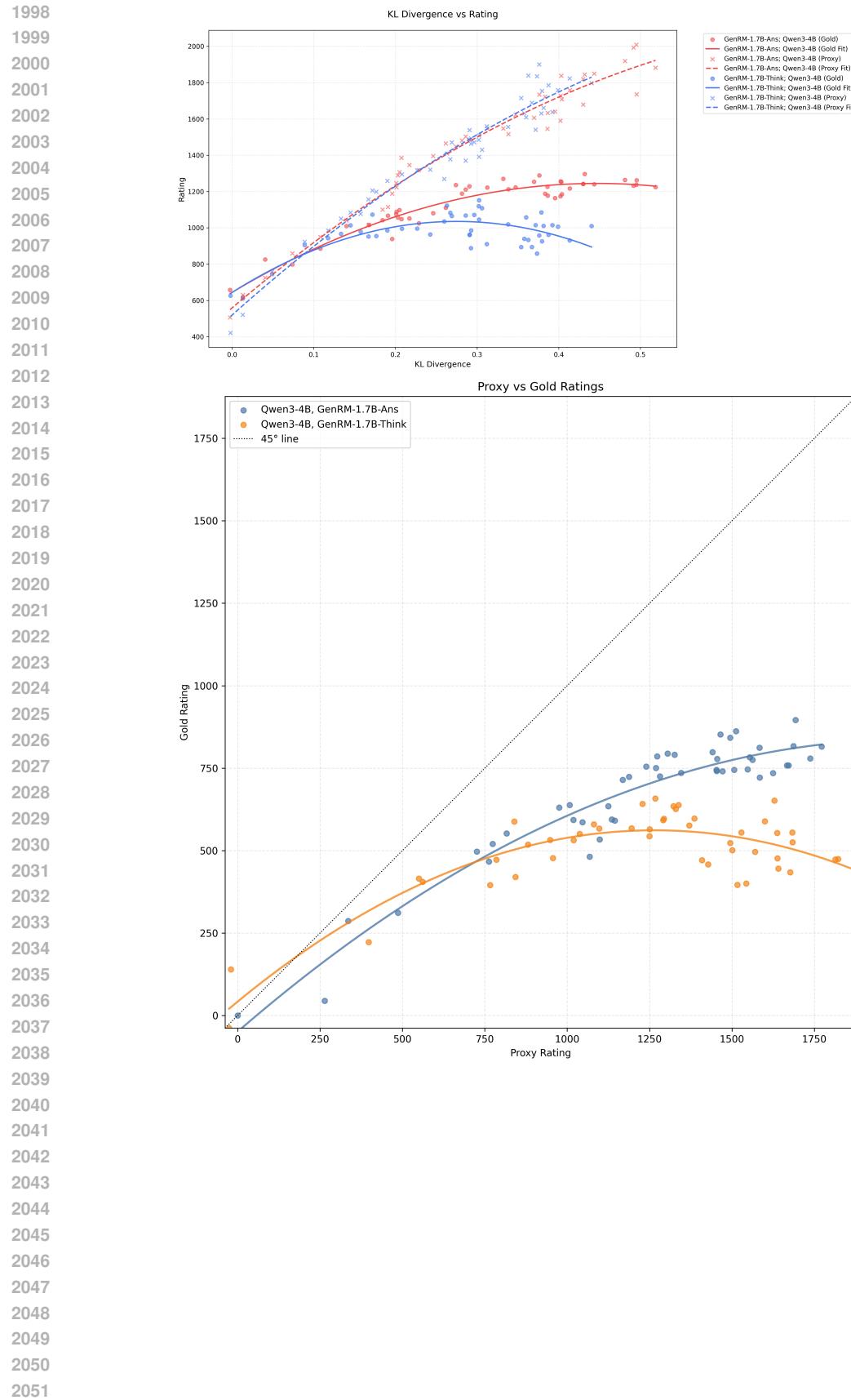
1889



1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997

C.12 THINKING VS ANSWER (4B vs 1.7B)





2052
2053

C.13 THINKING VS ANSWER (4B VS 4B)

2054

2055

2056

2057

2058

2059

2060

2061

2062

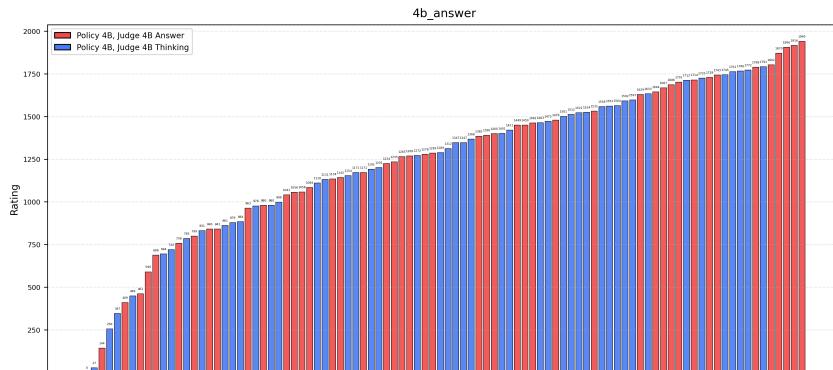
2063

2064

2065

2066

2067



2068

2069

2070

2071

2072

2073

2074

2075

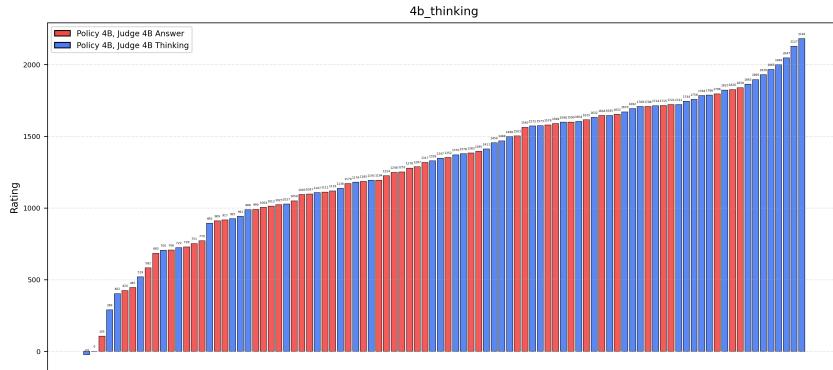
2076

2077

2078

2079

2080



2081

2082

2083

2084

2085

2086

2087

2088

2089

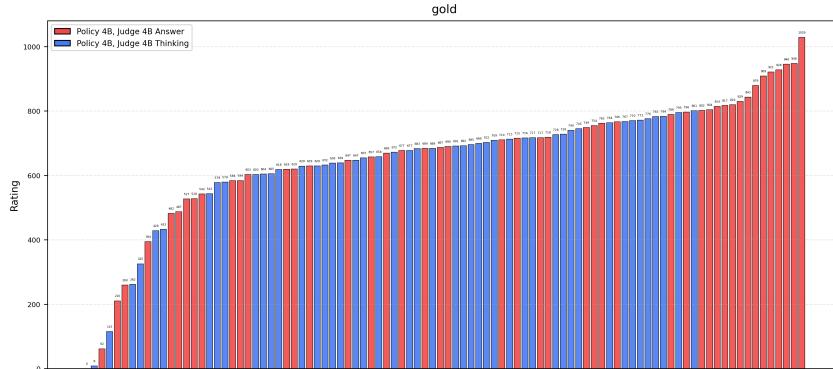
2090

2091

2092

2093

2094



2095

2096

2097

2098

2099

2100

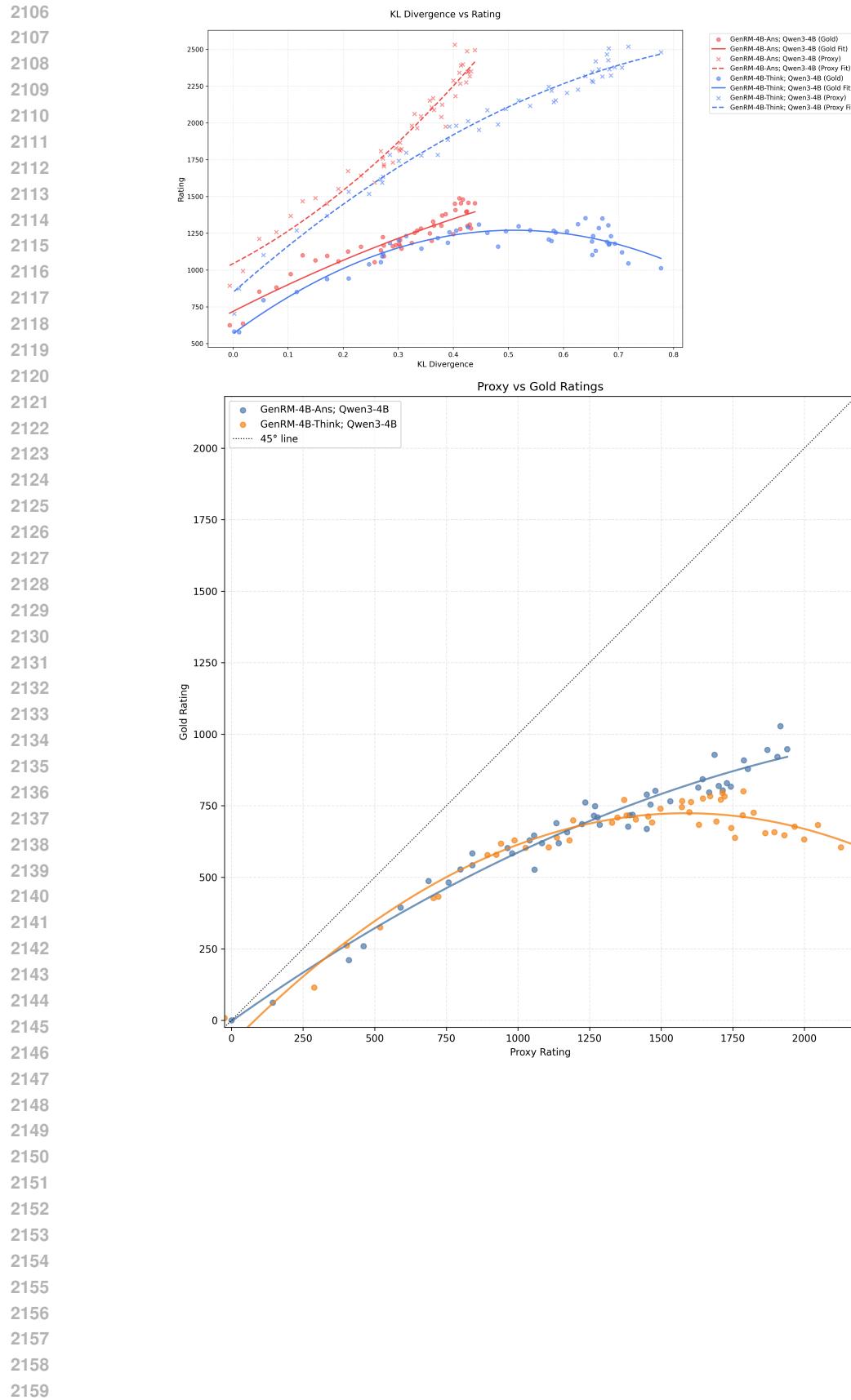
2101

2102

2103

2104

2105



2160 C.14 THINKING VS ANSWER (4B VS 8B)
2161

2162

2163

2164

2165

2166

2167

2168

2169

2170

2171

2172

2173

2174

2175

2176

2177

2178

2179

2180

2181

2182

2183

2184

2185

2186

2187

2188

2189

2190

2191

2192

2193

2194

2195

2196

2197

2198

2199

2200

2201

2202

2203

2204

2205

2206

2207

2208

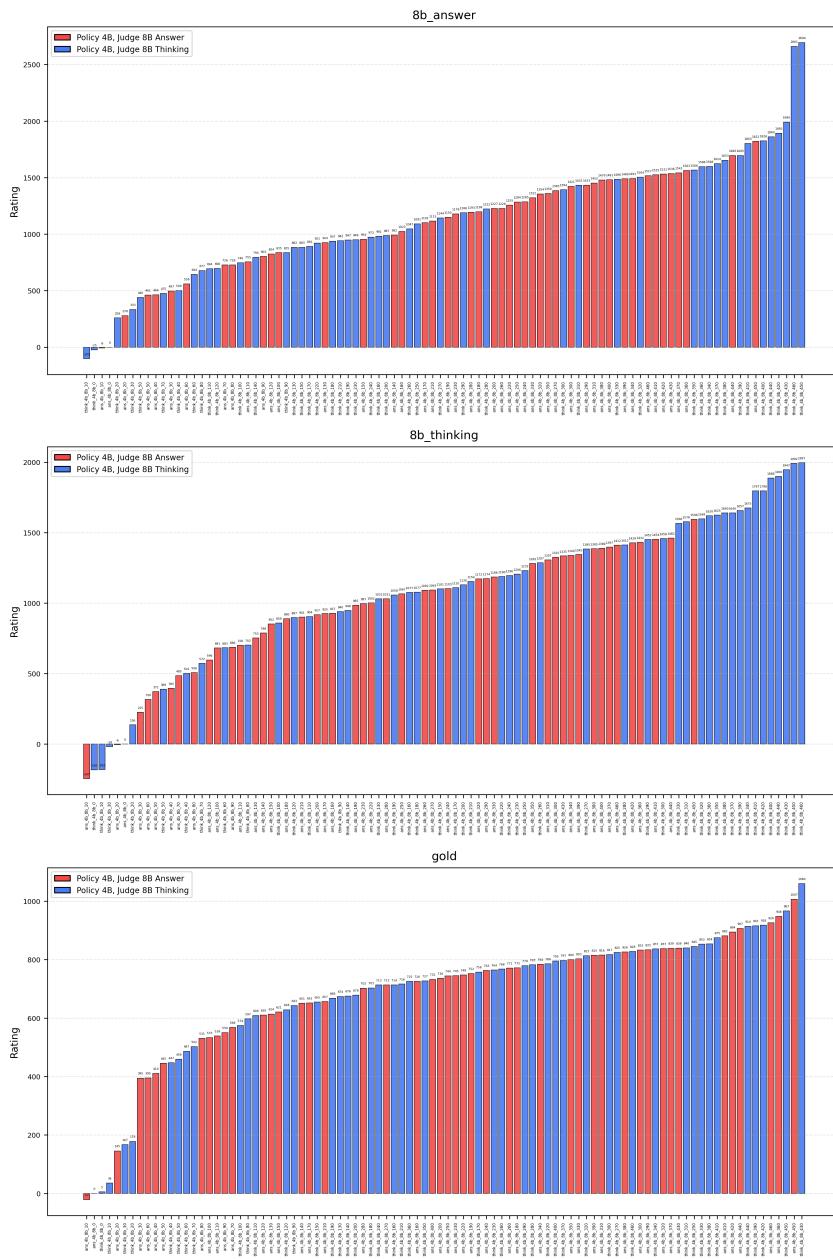
2209

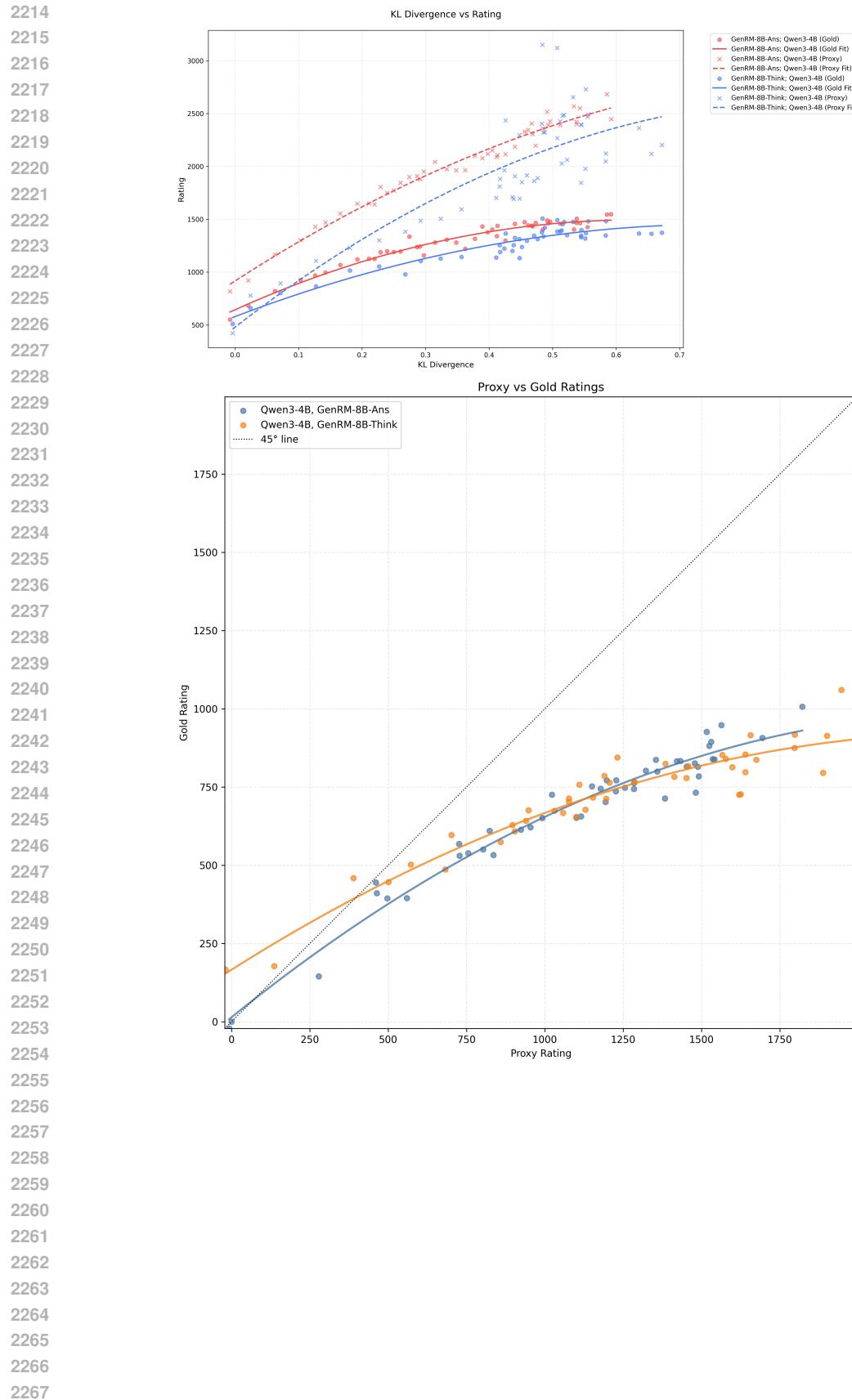
2210

2211

2212

2213





2268 C.15 THINKING VS ANSWER (4B VS 14B)
2269

2270

2271

2272

2273

2274

2275

2276

2277

2278

2279

2280

2281

2282

2283

2284

2285

2286

2287

2288

2289

2290

2291

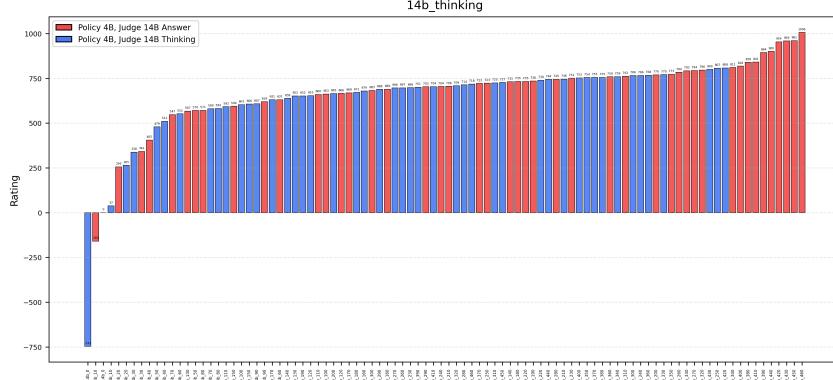
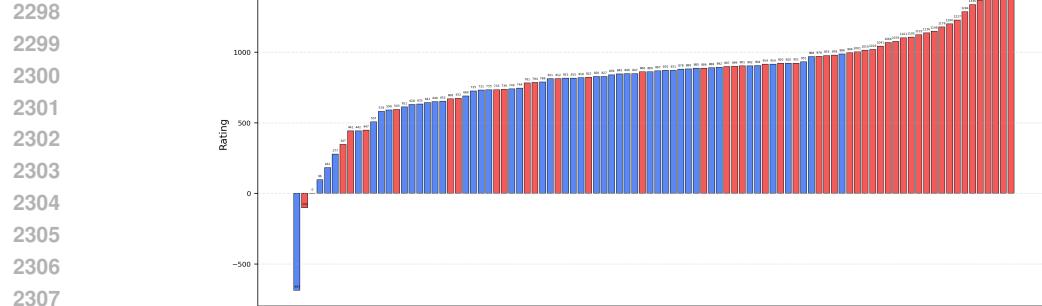
2292

2293

2294

2295

2296



2322 C.16 THINKING VS ANSWER (8B VS 4B)
2323

2324

2325

2326

2327

2328

2329

2330

2331

2332

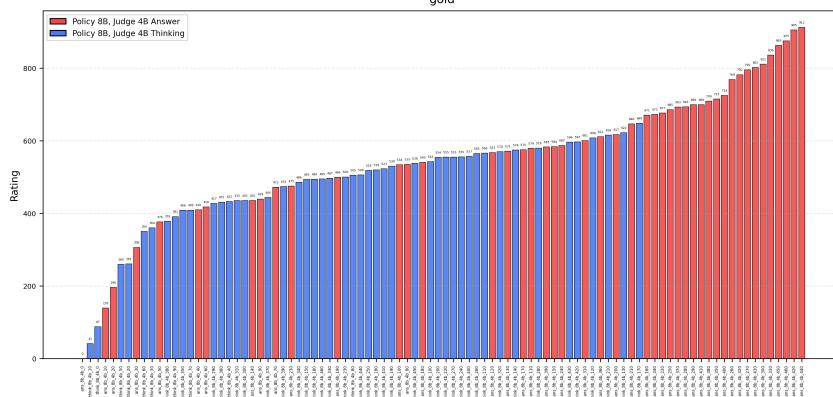
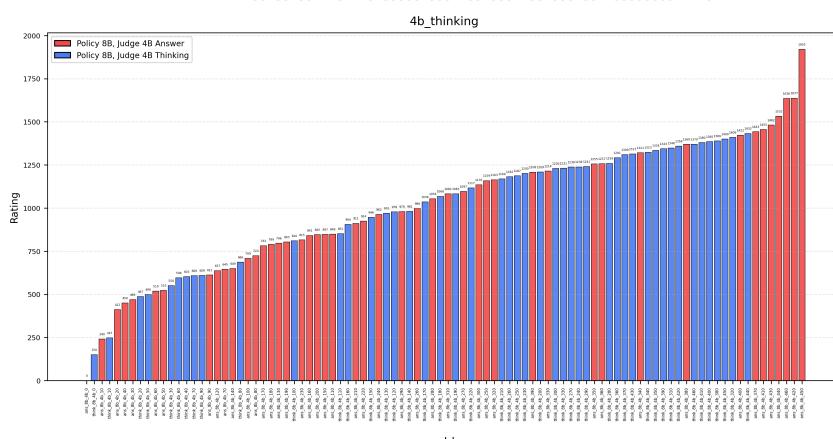
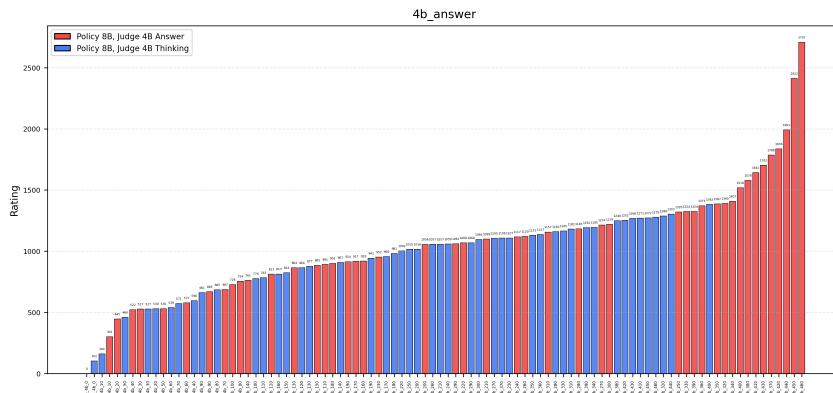
2333

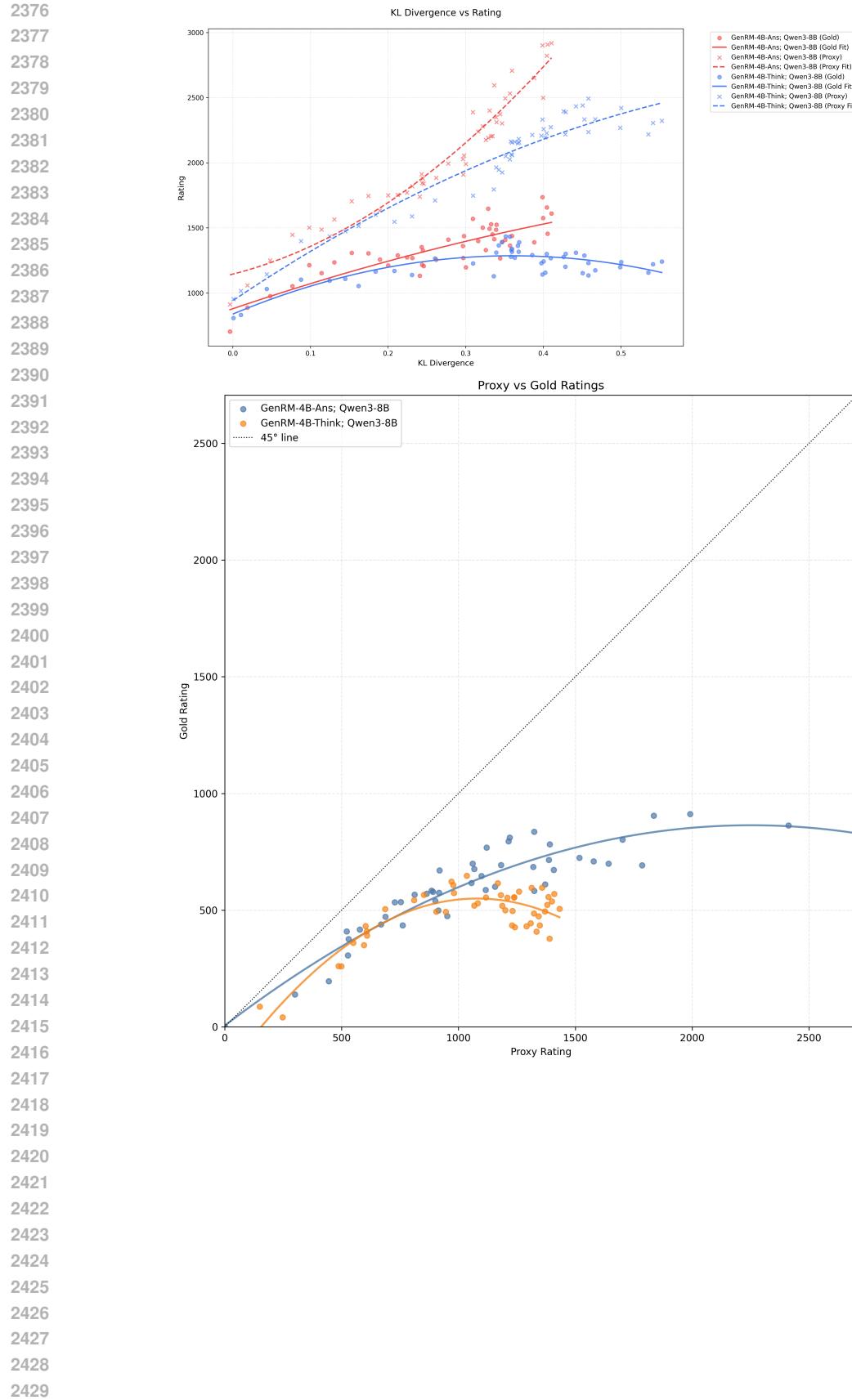
2334

2335

2336

2337





2430 C.17 TRAINED VS BASELINE
2431

2432

2433

2434

2435

2436

2437

2438

2439

2440

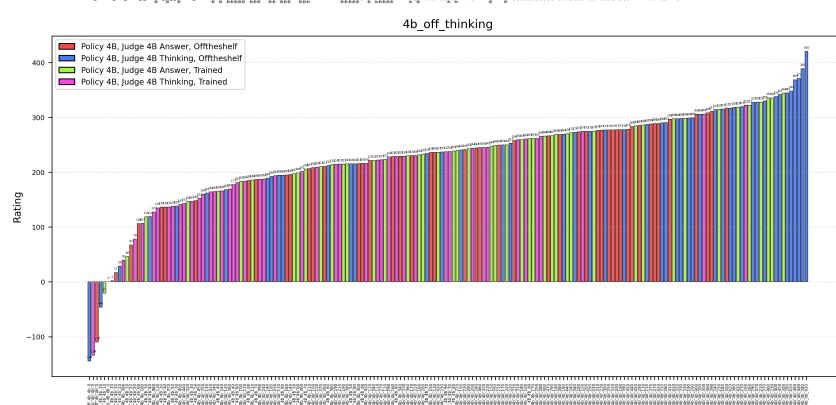
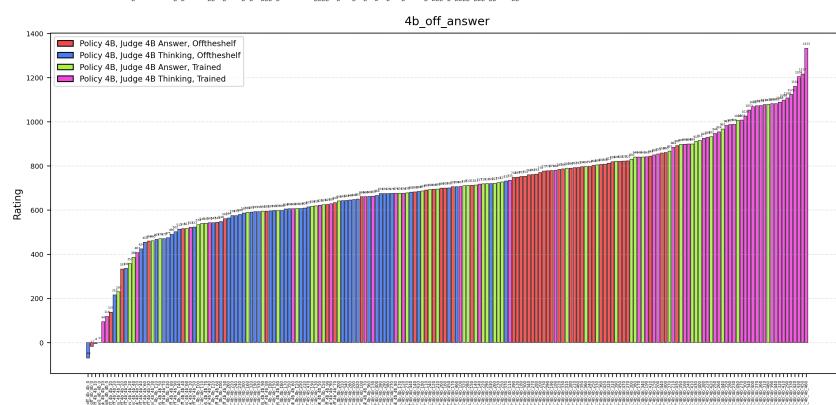
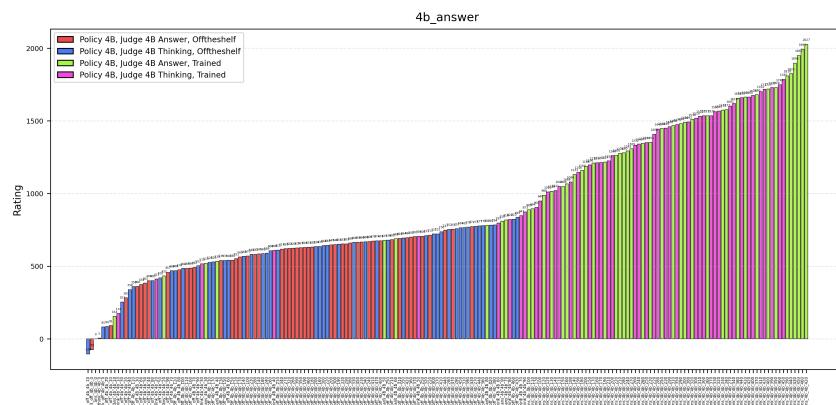
2441

2442

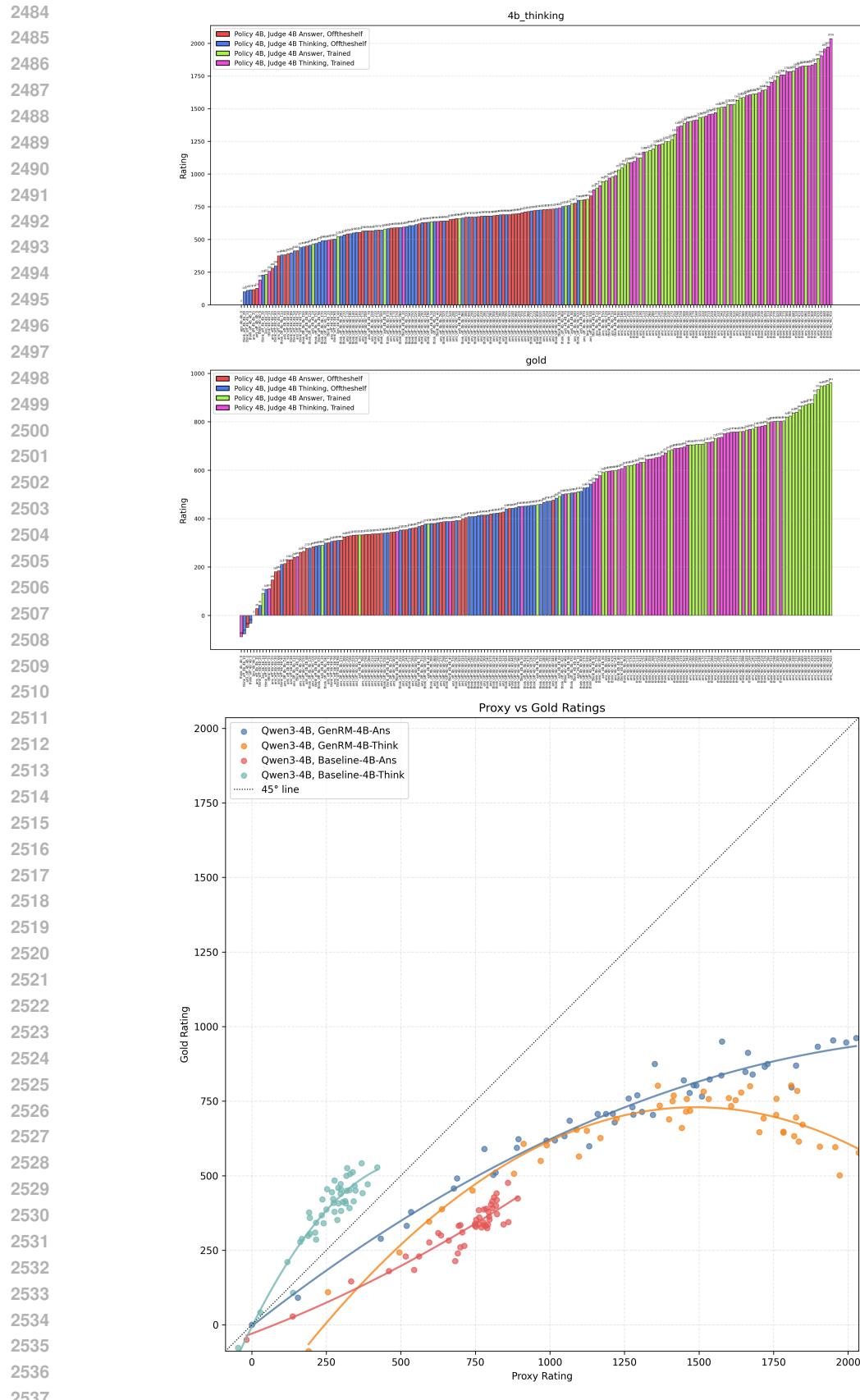
2443

2444

2445



2483



2538 C.18 ALL PARAMS

2539

2540

2541

2542

2543

2544

2545

2546

2547

2548

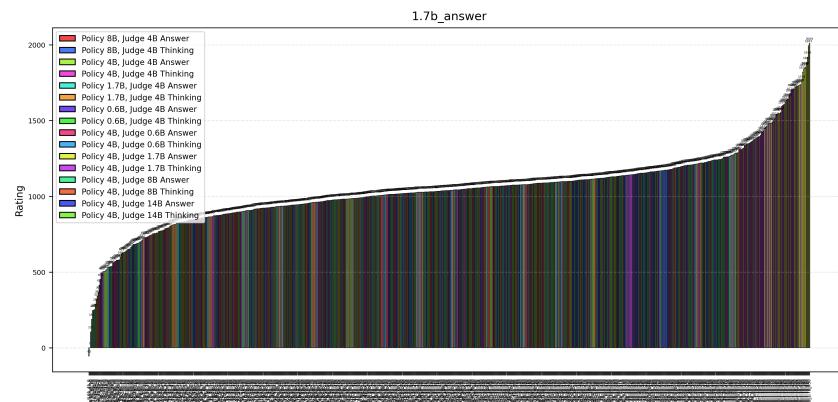
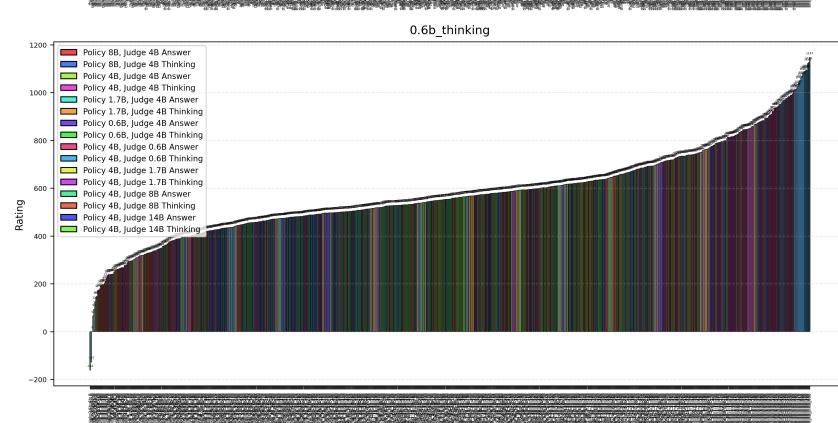
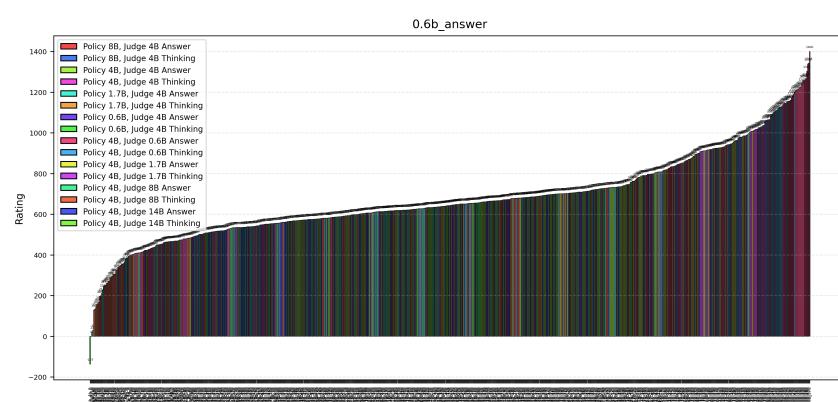
2549

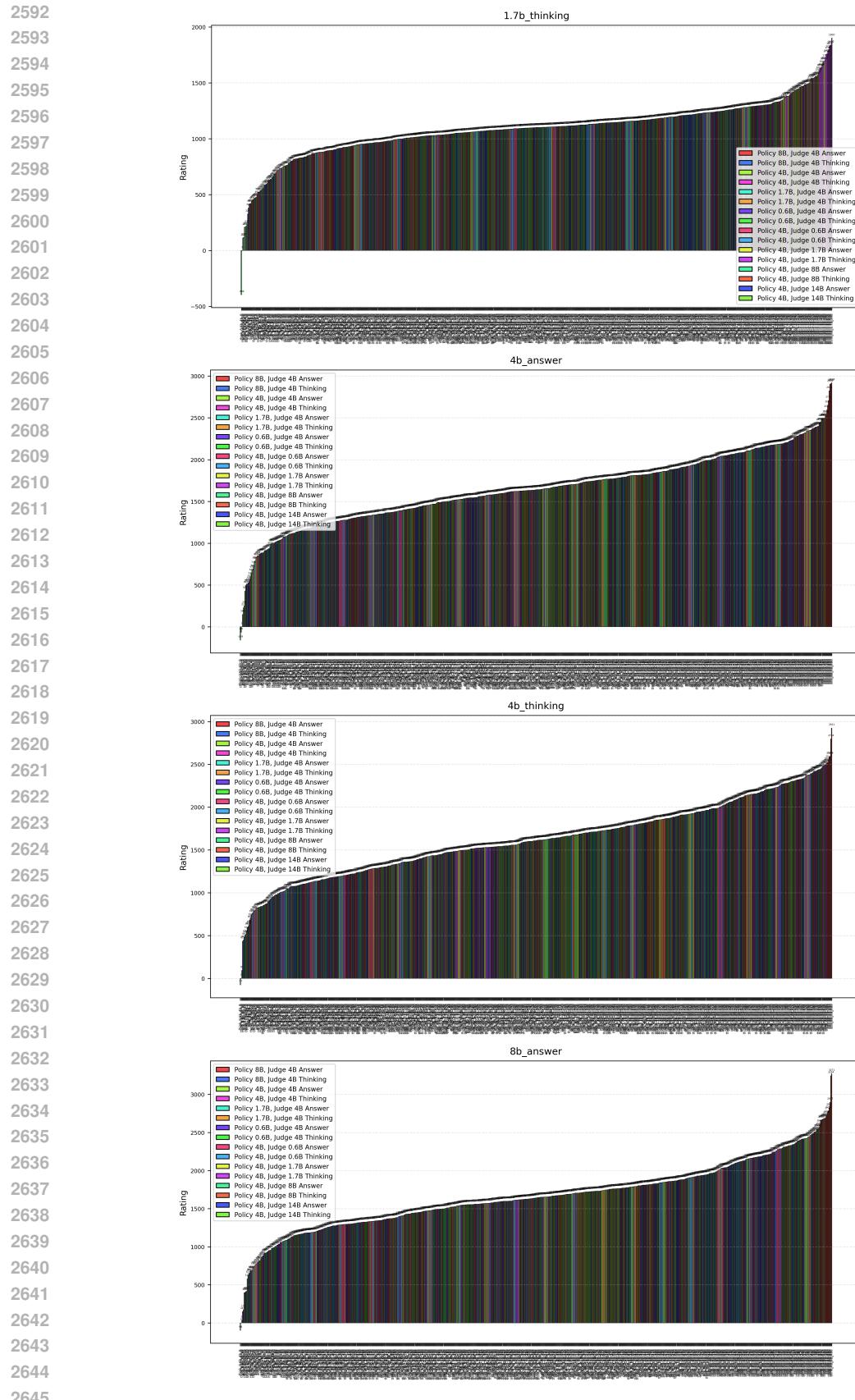
2550

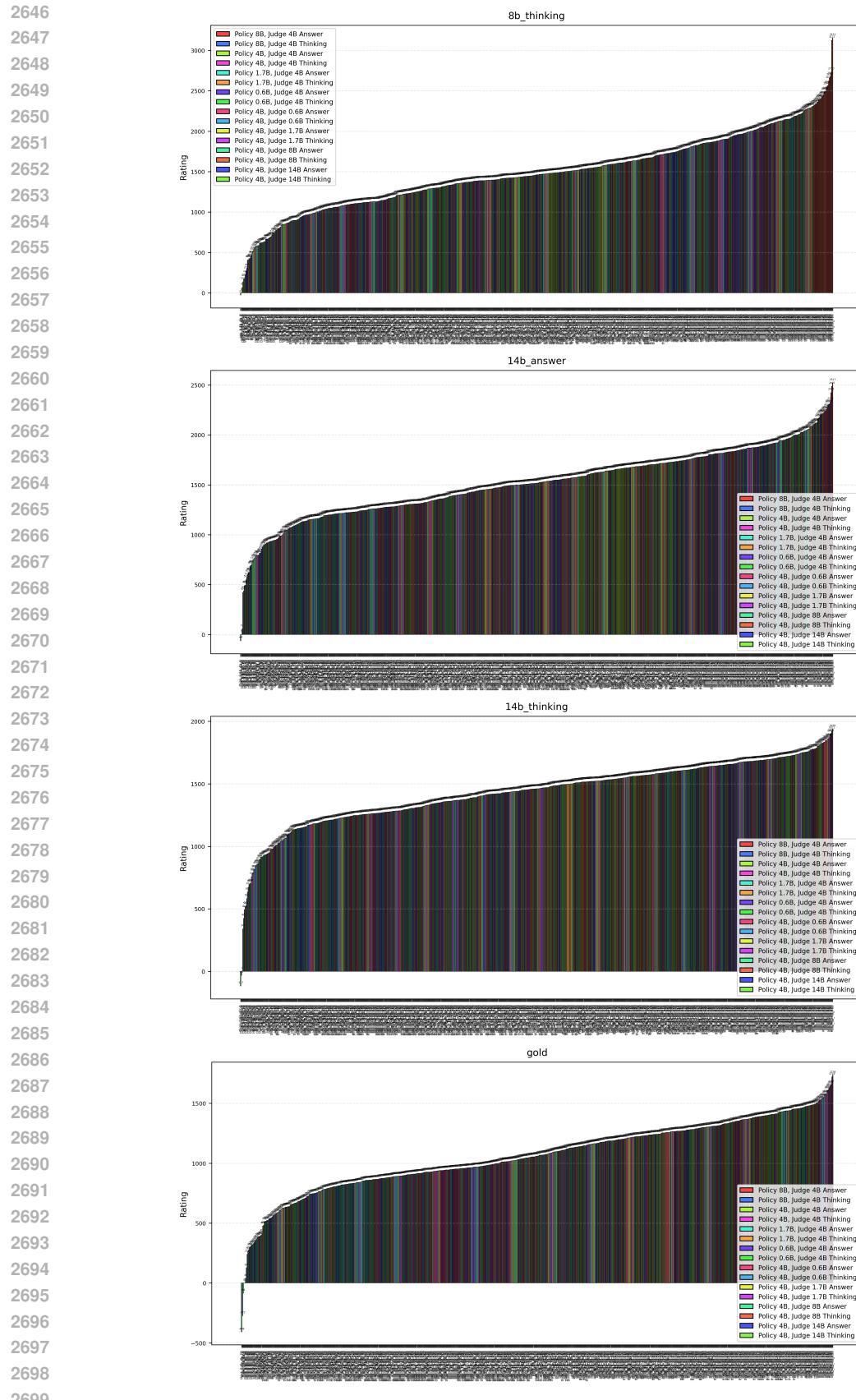
2551

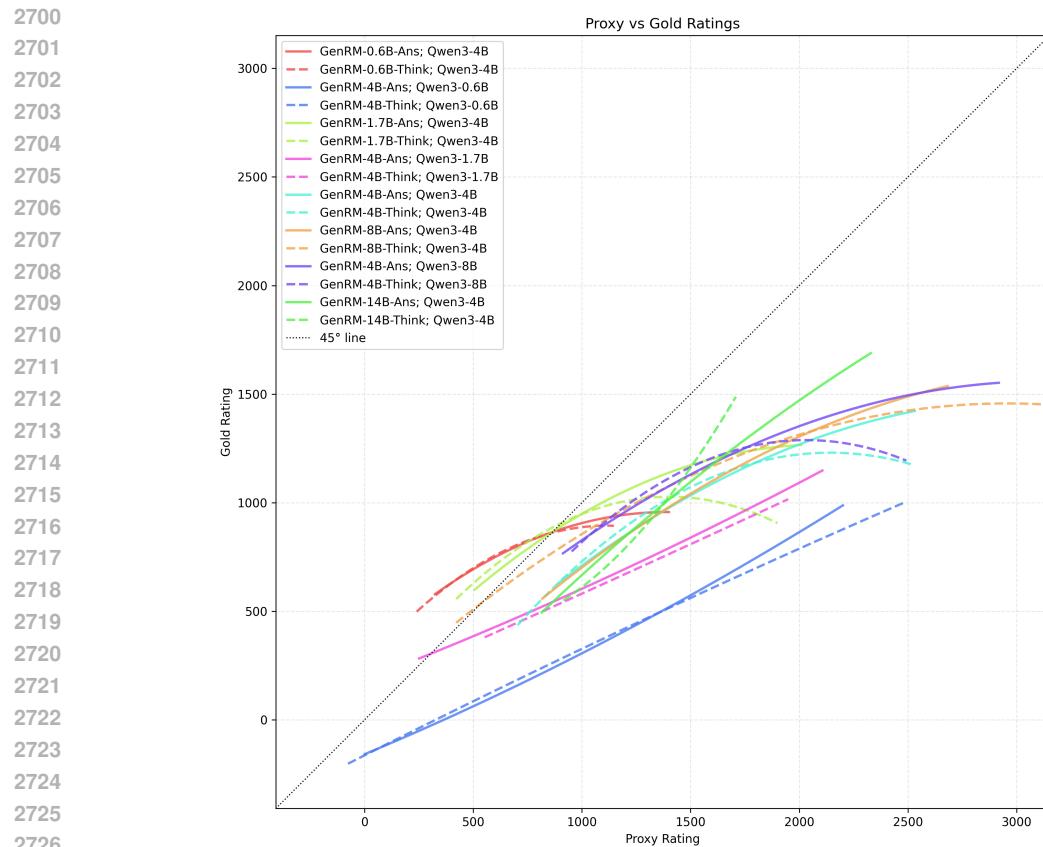
2552

2553

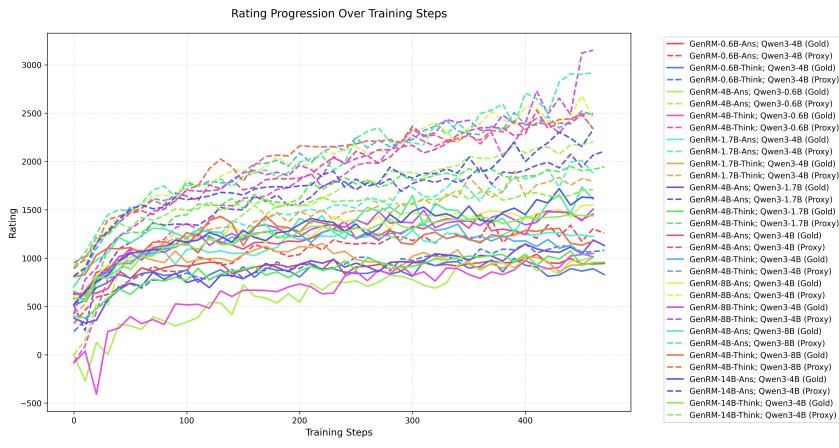








2754



2755

2756

2757

2758

2759

2760

2761

2762

2763

2764

2765

2766

2767

2768

2769

2770

2771

2772

2773

2774

2775

2776

2777

2778

2779

2780

2781

2782

2783

2784

2785

2786

2787

2788

2789

2790

2791

2792

2793

2794

2795

2796

2797

2798

2799

2800

2801

2802

2803

2804

2805

2806

2807

