

Benchmarking and Improving Generative Diversity in Language Models via Diverse Preference Optimization

Stanford CS224N Custom Project

Shyam Sai Bethina
Department of Computer Science
Stanford University
sbethina@stanford.edu

Annika Singh
Department of Computer Science
Stanford University
annikaks@stanford.edu

Abstract

The primary goal of this project is to understand and improve generative diversity in aligned large language models. We investigate the scientific question:

How does model reasoning configuration (“thinking” vs standard generation) affect response diversity, and can diversity-aware preference optimization (DivPO) improve diversity without degrading output quality?

To explore this question, we measure the accuracy and diversity of baseline open-source Qwen models and then apply DivPO to encourage more diverse outputs while preserving prompt adherence. Comparing baseline and post-training results, we find that reasoning configuration alone does not improve the diversity–quality trade-off, and that in our setting DivPO improves judged prompt adherence but does not increase embedding-based diversity.

1 Key Information to include

- Mentor: Chenglei Si
- External Collaborators (if you have any): N/A
- Sharing project: N/A
- Late Days (if you are using any): N/A

2 Introduction

Large language models (LLMs) are widely used for open-ended generation tasks such as storytelling and creative writing. In these settings, it is not enough for responses to be correct or coherent; users often want multiple outputs that are meaningfully different from one another. However, aligned language models frequently produce responses that are semantically similar across repeated generations, suggesting a form of diversity collapse. Improving generative diversity while maintaining output quality is therefore an important challenge.

Recent frontier models also expose reasoning or “thinking” modes that allocate additional computation before producing a final response. While these modes are designed to improve performance on complex reasoning tasks, their effect on open-ended generation is less clear. Additional reasoning may increase diversity by exploring a broader space of possible outputs, or it may reduce diversity by concentrating probability mass on a smaller set of high-confidence responses.

In this work, we study generative diversity across multiple model families under both reasoning and non-reasoning configurations using prompts from the Creative Writing v3 benchmark. We generate

multiple responses per prompt and measure diversity using embedding-based semantic deviation metrics. To evaluate output quality, we use an LLM-as-judge setup that scores prompt adherence and relevance.

Our baseline experiments reveal a clear trade-off: configurations that produce more diverse outputs often exhibit lower judged accuracy. To address this, we investigate Diverse Preference Optimization (DivPO), which constructs preference pairs that favor responses that are both high-quality and less redundant relative to other generations. Our goal is to determine whether diversity-aware preference optimization can improve generative diversity without significantly degrading output quality.

3 Related Work

Post-training techniques such as supervised fine-tuning, reinforcement learning from human feedback (RLHF) [1], and preference optimization have become standard approaches for aligning large language models with human expectations. While these methods improve instruction following and helpfulness, they often concentrate probability mass around high-reward responses, which can reduce generative diversity. This issue is particularly noticeable in open-ended tasks such as creative writing, where varied outputs are desirable.

Direct Preference Optimization (DPO) [2] is a widely used alignment method that trains models using pairs of preferred and dispreferred responses. In typical pipelines, these pairs are constructed based primarily on quality signals from reward models or human annotations. As a result, models tend to learn to reproduce common high-quality responses rather than exploring less frequent but still valid alternatives.

Diverse Preference Optimization (DivPO) [3] addresses this limitation by incorporating diversity into the construction of preference pairs. Instead of selecting responses solely by reward ranking, DivPO samples a pool of candidate generations for each prompt and evaluates them using both quality and diversity signals. Responses that are high-quality but uncommon relative to the pool are selected as preferred examples, while more typical or lower-quality responses are treated as rejected examples. Because DivPO modifies only how preference pairs are constructed, it can be integrated into existing DPO-style training pipelines without changing the underlying loss function.

Empirical results from the DivPO paper show that this approach can significantly increase diversity in creative generation tasks while maintaining comparable quality metrics. These findings suggest that diversity-aware preference construction provides a practical mechanism for encouraging varied outputs without degrading overall model performance. Our work builds on this idea by studying diversity in reasoning and non-reasoning LLM configurations and evaluating whether diversity-aware preference optimization can improve the diversity–quality trade-off in creative writing tasks.

4 Approach

We have two stages to our experiment (1) benchmarking diversity and accuracy for the baseline model for both reasoning and non reasoning modes and (2) applying Diverse Preference Optimization (DivPO) [3] to mitigate diversity collapse and encourage the model to have increased diversity while maintaining accuracy. The following describes our experimental setup:

1. Extract prompts that specifically inspire creativity from models (from creative writing bench [4])
2. Generate 8 distinct responses from the base model. Because post-training requires access to model weights, we use the open-source Qwen3-8B model family.
3. Utilize *intfloat/e5-large* to retrieve embedding representation of all 8 responses for each of the prompts for each of the models tested.
4. Calculate two diversity metrics: (1) average generation diversity (deviation) and (2) largest diversity between any two responses. This is more precisely described in Section 5.2
5. Use *claude-opus-4-6* as an LLM judge to score each response for prompt adherence and relevance on a 0–100 scale. Calculate the average score across all responses for overall accuracy.

This helps us establish our baselines after this we still need to train and retest the newly trained model. Thus we need to do the following:

1. Use non reasoning and reasoning models to generate 8 responses per prompts 0 - 26 from the dataset
2. Calculate the diversity and accuracy component for each prompt and give each prompt a score based on a combination of the two metrics.
3. Create pairwise training data that can be used by Tinker [5] for finetuning
4. Finetune model, experimenting with hyperparameter to best minimize loss and maximize reward.
5. Use the new model and test as we did with the baselines (above description).

5 Experiments

5.1 Data

For this task we use the Creative Writing v3 dataset [4]. This consists of 32 distinct prompts each with 10 different potential modifications. We designated prompts 0 - 26 as the training split and prompts 28 - 33 as the held-out evaluation split. In practice, prompts 0 and 27 were missing from the dataset file, so the actual training set consisted of prompts 1 - 26 and the held-out evaluation set consisted of prompts 28 - 33.

5.2 Evaluation Method

For evaluation, we used prompts 28-33, again using all 10 seed modifiers together with the unmodified base prompt. This yields 11 prompt conditions per prompt and 66 total evaluation conditions. This scale is comparable to other diversity-focused benchmarks that use on the order of 10^3 prompts to measure variability in generative behavior [6, 7]. For each condition, we sampled 8 responses from the model, both before and after DivPO fine-tuning, so that baseline and post-training comparisons were performed under the same generation setup.

We evaluate models along two dimensions: **generative diversity** and **output quality**. Diversity is measured using embedding-based semantic distance across multiple responses generated for the same prompt. For each prompt condition, we sample $n = 8$ responses and embed them using the *intfloat/e5-large* embedding model. We compute two diversity metrics.

The first metric is **Average Deviation**, defined as the average cosine distance between each response embedding and the centroid of the response set:

$$\text{AvgDev} = \frac{1}{n} \sum_{i=1}^n \left(1 - \frac{x_i^\top c}{\|x_i\| \|c\|} \right), \quad c = \frac{1}{n} \sum_{i=1}^n x_i.$$

The second metric is **Max Deviation**, which captures the most semantically distinct response in the set:

$$\text{MaxDev} = \max_i \left(1 - \frac{x_i^\top c}{\|x_i\| \|c\|} \right).$$

These metrics quantify how spread out the responses are in semantic embedding space and therefore provide a proxy for generative diversity.

To evaluate output quality, we use an **LLM-as-judge** setup. Each response is scored by *claude-opus-4-6* on a 0–100 scale based on prompt adherence and relevance. We report the mean judge score across all scored responses as the overall quality metric for each model configuration. This allows us to analyze the trade-off between diversity and prompt adherence across different model configurations and optimization strategies.

5.3 Experimental details - DivPO Training

In practice, prompt 0 was missing from the dataset file, so the actual training set consisted of prompts 1-26. For each training prompt, we used the base prompt together with all 10 seed modifiers, yielding 11 prompt conditions per prompt and 286 total training prompt-conditions.

For each prompt-condition, we sampled 8 candidate responses from the target Qwen model. First, we scored each candidate using a quality score as we did in the baseline ($Q(y)$ = a score between 0 - 100 representing the accuracy of the LLM’s response to the prompt).

Second, we computed a marginal diversity contribution in embedding space relative to a greedily selected set S of already chosen responses. If $e(y)$ denotes the embedding of y , then

$$D(y | S) = \begin{cases} 0, & S = \emptyset, \\ \frac{1}{|S|} \sum_{y' \in S} d_{\cos}(e(y), e(y')), & \text{otherwise,} \end{cases}$$

where $d_{\cos}$ is cosine distance.

Within each prompt-condition, we retained only responses above the above a accuracy score equal to the average baseline scores (33.22 for non-reasoning and 4.19 for reasoning, see Table 2). For the remaining candidates, we defined a combined reward

$$R(y | S) = \lambda_q Q(y) + \lambda_d D(y | S),$$

with $\lambda_q = 1.0$ and $\lambda_d = 1.5$ to signify the promotion of diversity while maintaining quality. We then constructed pairwise preferences greedily: at each step, we selected the highest-reward remaining candidate as the *chosen* response and paired it against each lower-reward remaining candidate as a *rejected* response. Each resulting preference pair was annotated with `chosen_quality`, `rejected_quality`, `chosen_marginal_diversity`, `rejected_marginal_diversity`, and `reward_gap`. This was done for both the reasoning and non-reasoning models and their responses were used to train their respective models.

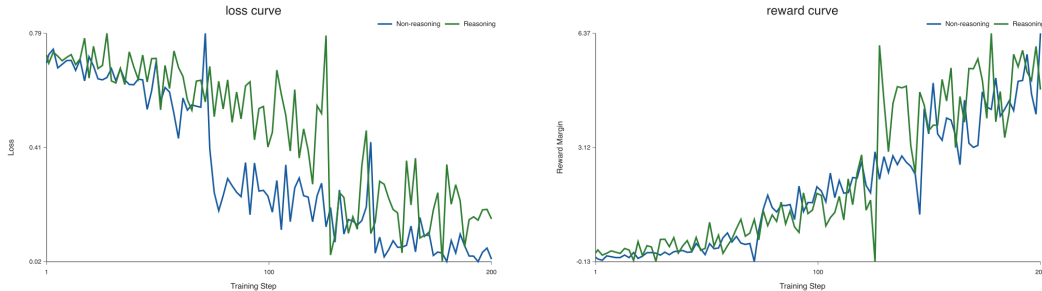
5.4 Results

In order to see a reasonable loss curve we training our model with the generated pairs a couple times. The results we describe in this section are from our best training attempt, specifically with a larger pair batch size and a smaller learning rate:

$$\text{batch_size_pairs} = 16, \quad \text{learning_rate} = 2 \times 10^{-5}, \quad \text{num_steps} = 200.$$

We focus on this run because earlier settings, particularly smaller batch sizes and a larger learning rate, produced substantially noisier optimization curves and less stable reward-margin dynamics. We found that this is a common trend with a DPO/DivPO objective. Because preference signal is already noise, smaller batches can make the loss trajectory hard to interpret [2, 8].

5.4.1 Training Results



(a) Loss curve for reasoning and non-reasoning

(b) Reward curve for reasoning and non-reasoning

Figure 1: Loss and Reward curve from best training run

We first analyze the optimization behavior of the final run configuration. Figure 1 shows the loss and reward curves for the final DivPO training run, with non-reasoning shown in blue and reasoning shown in green. In both cases, lower loss and larger reward margin indicate that the model is increasingly assigning higher preference to the chosen responses than to the rejected responses. Table 1 shows how overtime the loss decrease and the reward increases. However, these trends did not translate into strong improvements in diversity, as discussed in Section 5.4.2.

Mode	Δ Loss (last20 – first20)	Δ Reward (last20 – first20)
Non-reasoning	-0.6341	+4.8399
Reasoning	-0.4402	+2.2018

Table 1: Change in mean loss and mean reward from the first 20 to the last 20 steps of DivPO

5.4.2 Evaluation Results

We evaluated both the baseline and post-trained models on prompts 28-33, using the base prompt and all 10 seed modifiers, with 8 samples per prompt-condition. Table 2 reports the held-out embedding-based diversity metrics and LLM-as-judge accuracy before and after DivPO fine-tuning.

Mode	Model	Avg Dev.	Max Dev.	LLM Acc.
Non-reasoning	Baseline	0.02808	0.03738	33.22
Non-reasoning	Post-train	0.02614	0.03325	42.02
Non-reasoning	Δ (Post – Base)	-0.00194	-0.00413	+8.79
Reasoning	Baseline	0.01058	0.01470	4.19
Reasoning	Post-train	0.01101	0.01565	27.85
Reasoning	Δ (Post – Base)	+0.00044	+0.00094	+23.66

Table 2: Avg Dev. and Max Dev. are cosine-deviation-to-centroid and LLM Acc. is the mean prompt-adherence score. Bolded Δ rows show the change from baseline to post-training.

6 Analysis

There are several things that stood out to us in our experiments in both the baseline as well as the finetuned models.

6.1 Baseline Analysis

Interestingly, when testing on the baseline Qwen3-8B model we noticed something unexpected. The *non-reasoning model did better than the reasoning model in both diversity and accuracy*. The reasoning model achieved mean average deviation 0.01058 and mean LLM-judge accuracy 4.19, while the non-reasoning model achieved mean average deviation 0.02808 and mean LLM-judge accuracy 33.22. This suggests that simply enabling reasoning in Qwen3-8B does not improve the diversity–quality trade-off for this creative writing task but rather substantially worsened prompt adherence. However, this conclusion does not necessarily generalize across model families. We conducted the exact same baseline experiment described in Section 4 on other state of the art LLMs, specifically: *gpt-5.2* (with and without thinking enabled), *claude-opus-4-6* (with and without thinking enabled), *grok-4-1-fast-non-reasoning*, and *grok-4-1-fast-reasoning*.

Model Family	Avg Dev.	Max Dev.	LLM Acc.	Avg Dev.	Max Dev.	LLM Acc.
GPT	0.0953	0.1908	31.20	0.0597	0.0816	52.13
Claude	0.0644	0.0929	57.31	0.0631	0.0923	61.33
Grok	0.0520	0.0736	62.29	0.0538	0.0770	58.91

Table 3: Baseline diversity and LLM-as-judge accuracy by model family. **Reasoning** metrics are shown in green and **Non-reasoning** metrics are shown in blue.

This table is more consistent with the expected diversity–accuracy trade-off: for both GPT and Claude, the reasoning configuration is more diverse but less accurate than the non-reasoning configuration. Grok more closely resembles Qwen in that its reasoning configuration does not improve both objectives simultaneously. Across frontier models, increasing diversity often corresponds to reduced prompt adherence as seen in Figure B. By contrast, Qwen behaves differently: its reasoning baseline is substantially worse than its non-reasoning baseline on both diversity and accuracy, and post-training improves prompt adherence more strongly than diversity.

6.2 Training Analysis

DivPO substantially improved prompt adherence for both models, but it did not materially improve semantic diversity. In the non-reasoning case, post-training increased judged accuracy from 33.22 $\rightarrow$ 42.02, but average deviation fell from 0.02808 $\rightarrow$ 0.02614 and maximum deviation fell from 0.03738 $\rightarrow$ 0.03325. In other words, the model became better at following the prompt, but slightly less semantically diverse. This is the opposite of the intended DivPO outcome.

However for the reasoning model, its prompt-adherence score increased sharply from 4.19 $\rightarrow$ 27.85, and its diversity metrics also increased slightly (average deviation 0.01058 $\rightarrow$ 0.01101, maximum deviation 0.01470 $\rightarrow$ 0.01565). However, even after post-training, the reasoning model still underperformed the non-reasoning model on both accuracy and diversity. Thus, although DivPO substantially repaired the reasoning model’s baseline failure mode, it did not make reasoning superior to non-reasoning for this task.

6.3 Recommended Approach

Overall, our results suggest that in this setting DivPO is more effective as a **quality-recovery method** than as a mechanism for substantially increasing held-out semantic diversity. The strongest evidence for this is the non-reasoning run: it optimized successfully, improved judged accuracy, but still became slightly less diverse on the baseline.

We believe the most important next step is to improve the pair-construction stage. In our current setup, the combined reward uses $\lambda_q = 1.0$ and $\lambda_d = 1.5$, so diversity is weighted more heavily than the quality proxy. Yet, the learned behavior still appears to favor prompt adherence over held-out diversity, suggesting that the issue is not only the reward weights themselves, but also the quality score, the candidate pool size, and the structure of the preference pairs. Future work should therefore explore stronger diversity-only comparisons among quality-matched candidates and larger training pools.

7 Conclusion

We first evaluated multiple model families under reasoning and non-reasoning generation settings and found a consistent diversity–accuracy trade-off: configurations that produced more varied outputs often exhibited lower prompt-adherence scores. We then implemented a DivPO-style preference construction pipeline and fine-tuned a Qwen3-8B model to investigate whether diversity-aware preference optimization could improve this trade-off.

Our results show that DivPO training significantly improved prompt adherence for both reasoning and non-reasoning configurations, but did not consistently increase held-out semantic diversity. In particular, the non-reasoning model became more accurate while slightly less diverse, suggesting that the optimization process prioritized quality signals over diversity signals during training. The reasoning configuration showed modest diversity gains alongside large improvements in prompt adherence, though it still underperformed the non-reasoning model overall.

These findings suggest that in our current setup DivPO acts more reliably as a prompt-adherence stabilization mechanism than as a strong driver of held-out diversity improvements. We hypothesize that this behavior is largely determined by the preference pair construction stage, where the balance between quality and diversity signals directly shapes the optimization objective. Future work should therefore explore alternative pair construction strategies, larger candidate pools, and stronger diversity weighting in the reward formulation. More broadly, our experiments highlight the importance of

carefully designing diversity-aware training signals when attempting to improve generative diversity without sacrificing output quality.

Team contributions (Required for multi-person team)

Annika: Entire non-reasoning model (baseline, pairwise construction post training), Writeup

Sai: Entire reasoning model (baseline, pairwise construction post training), Writeup

References

- [1] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.
- [2] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2024.
- [3] Jack Lanchantin et al. Diverse preference optimization. *arXiv preprint arXiv:2501.18101*, 2025.
- [4] EQ-Bench. Creative writing benchmark v3. <https://github.com/EQ-bench/creative-writing-bench>, 2024.
- [5] Thinking Machines Lab. Tinker: Tools for preference optimization and alignment experiments. <https://github.com/thinkingmachines/tinker>, 2025. Accessed: 2026.
- [6] Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. Noveltybench: Evaluating creativity and diversity in language models, 2025.
- [7] Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. Writingbench: A comprehensive benchmark for generative writing, 2025.
- [8] Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is dpo superior to ppo for llm alignment? a comprehensive study. *ICML*, 2024.

A Appendix

A.1 LLM-as-Judge Prompt

You are grading prompt adherence for creative-writing outputs.
For EACH response, assign a score from 0 to 100 for how well
it answers the prompt and follows constraints.
Scale: 0=unrelated/fails prompt, 50=partially follows prompt,
100=fully addresses prompt constraints.

Return STRICT JSON ONLY with this exact schema:

```
{
  "scores": [
    {"sample_index": <int>, "record_id": "<string>", "score":
    <int 0-100>}
  ]
}
```

No markdown. No extra keys.

Prompt:

```
{prompt_text}
```

```
Responses:
sample_index: {sample_index}
record_id: {record_id}
response:
{response_text}
```

A.2 Testing Prompts

1. Teen Horror (Goosebumps style)

Write a chapter where a middle school student discovers their new smart watch has the ability to rewind time, but doesn't realize that each use is aging them until they notice physical changes (like accelerating through puberty) and that people are treating them differently. They inadvertently cause a timeline shift with disastrous consequences — and it can only be rectified (maybe?) by reversing the clock once more.

Length: 1000 words.

2. Biopunk — Polymerase Chain Reaction

Write a chapter set in a near-future post-apocalyptic world where corporate genetic engineering was designed with built-in obsolescence—engineered DNA lasting only one generation to maximize profits before societal collapse. Follow a biohacker working with salvaged equipment to splice viable food crops for their settlement using rapidly degrading genetic material. As world building / window dressing, include details about jury-rigged lab equipment, and the physical toll (and unexpected upsides) of using untested genetic modifications as medicine. The plot of this chapter is centred around a black market contract.

Length: 1000 words.

3. Folk Horror: Útiseta

Write a chapter set in rural northern Sweden in the mid-1990s. Linnea, a teenage girl from Stockholm, has reluctantly moved to a small village with her parents after her father's job transfer. She's befriended Astrid, a local girl who claims her grandmother taught her ancient practices passed down through generations. Astrid has been introducing Linnea to "midnight games" that involve old rituals, convincing her they can use them to learn classmates' secrets or influence their behavior.

Tonight, Astrid has convinced Linnea to participate in Útiseta—sitting out on an ancient burial mound at midnight to "receive knowledge from beyond." The chapter focuses on this single midnight excursion, the ritualistic preparations, Astrid's matter-of-fact instructions that mix teenage rebellion with eerily specific traditional elements, and the mounting tension as Linnea becomes increasingly uncertain whether they're simply scaring themselves or if something is actually responding to their presence.

The narrative should maintain ambiguity about whether the strange occurrences are supernatural or psychological. First person, past tense, told as Linnea's recollection of events in the present day.

Length: 1000 words.

4. Young Adult Drama: Gone South

Taylor is a 16 year old teen with a history of "concerning behavioural patterns" (or that's how the stupid smug school counsellor worded it on the phone call to Mom). When you live in a backwater town in Tennessee, your choices are: a. be mind-fuckingly bored, or b. invent some drama. Two days ago, Taylor came out to their family as trans. It didn't go well.

In a moment of desperate inspiration, Taylor decided to get the hell out of dodge. Now Taylor is two states away, having been unceremoniously ejected from their hitched ride in downtown New Orleans. Write the next chapter in Taylor's story as they find their feet. Tell the chapter over the course of their first day in New Orleans, in which Taylor meets another runaway with issues and secrets. Write in Taylor's voice, incorporating their internal monologue as they grapple with intense & conflicting emotions. First person, mixed tense.

Length: 1000 words.

5. Whalepunk Fantasy: The Morska Conspiracy

The Morska thieves guild experiences a sudden power vacuum. Aya, a young woman born into aristocracy, has been moonlighting with the guild using her latent ability to summon

dark entities from the immateria. Aya and her companions, Katra and Simeon, make a play for an ancient scythan artifact in the depths of the guild HQ.

Write the scene as the trio enter the underhold and find themselves confronted by the guild leader, decidedly *un*-incarcerated. Aya is captured and unmasked. She attempts to stall their demise by negotiating on her value as a hostage. Lean into the gritty Victorian aesthetic with gothic horror elements. In this world, men trust brass and oil and steel; magicka is feared and barely understood. Third person past tense, heavy dialogue component.

Length: 1000 words.

6. **Cozy Sci-Fi: The Last Outpost**

Nettle (or Nets to her friends) bartends at The Last Outpost — a lazy hub for miners and merchants in an asteroid belt on the Andromeda outer rim. Escaping from her past life, Nets found a home here. Tonight, a frazzled looking stranger, tentacled dreads askew, quietly plops down at the bar: “Vega whisky, neat. Better make that a double.”

Write a chapter in this story. Establish conflict for narrative interest, but don’t resolve it yet. Aim for about 60% dialogue. First person, present tense.

Length: 1000 words.

B Accuracy vs Diversity

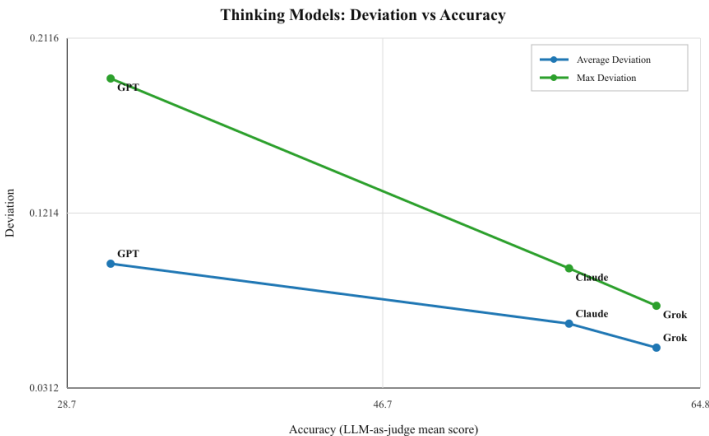


Figure 2: Accuracy (LLM-as-judge) vs Max Deviation