

# Budget-Aware Anytime Reasoning with LLM-Synthesized Preference Data

Xuanming Zhang<sup>1\*</sup>, Shwan Ashrafi<sup>2</sup>, Aziza Mirsaidova<sup>2</sup>, Amir Rezaeian<sup>2</sup>, Miguel Ballesteros<sup>2</sup>, Lydia B. Chilton<sup>1</sup>, Zhou Yu<sup>1</sup>, Dan Roth<sup>2</sup>

<sup>1</sup>Columbia University, <sup>2</sup>Oracle AI

xz2995@columbia.edu, {shwan.ashrafi,dan.roth}@oracle.com

## Abstract

We study the reasoning behavior of large language models (LLMs) under limited computation budgets. In such settings, producing useful partial solutions quickly is often more practical than exhaustive reasoning, which incurs high inference costs. Many real-world tasks, such as trip planning, require models to deliver the best possible output within a fixed reasoning budget. We introduce an anytime reasoning framework and the Anytime Index, a metric that quantifies how effectively solution quality improves as reasoning tokens increase. To further enhance efficiency, we propose an inference-time self-improvement method using LLM-synthesized preference data, where models learn from their own reasoning comparisons to produce better intermediate solutions. Experiments on NaturalPlan (Trip), AIME, and GPQA datasets show consistent gains across Grok-3, GPT-oss, GPT-4.1/4o, and LLaMA models, improving both reasoning quality and efficiency under budget constraints.

## 1 Introduction

Many real-world planning and decision-making tasks often face strict compute or latency budgets due to practical deployment constraints. In interactive systems, models must respond within seconds; in resource-limited environments, they must operate under bounded token or cost budgets. In these settings, even partial solutions can provide immediate utility (e.g., a feasible but incomplete trip plan), while additional computation can further refine them. This motivates the need for anytime reasoning, a process in which intermediate outputs improve in quality as more reasoning tokens are generated (Qi et al., 2025). Classic AI formalized this concept through anytime algorithms (Zilberstein, 1996; Dean and Boddy, 1988; Hansen and Zilberstein, 2001), emphasizing measurable solu-

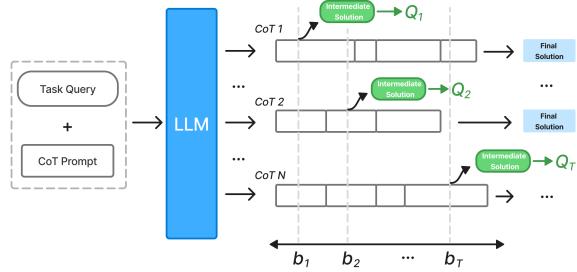


Figure 1: Overview of our anytime reasoning evaluation framework. The model generates  $N$  CoT traces, each truncated at token budgets  $b_i$  to evaluate intermediate solution quality  $Q_i$ . Final solutions are derived from full reasoning traces.

tion quality and monotonic improvement under resource constraints.

Modern large language models (LLMs) have shown strong reasoning capabilities through step-by-step methods such as Chain-of-Thought (CoT) (Wei et al., 2022; Chowdhery et al., 2023), Tree-of-Thoughts (Yao et al., 2023), and Self-Refine (Madaan et al., 2023). However, these techniques typically assume unrestricted computation and focus solely on final answer quality. In real-world use, reasoning can be interrupted and models should still produce the best possible answer within a fixed token or time budget. Yet current approaches provide no principled way to measure how reasoning quality evolves over time or how efficiently a model converts additional tokens into better solutions. This motivates two central research questions: (1) *Are current LLMs effective anytime reasoners?* and (2) *Can we improve their intermediate reasoning quality by restricting the amount of reasoning tokens being used?*

Recent work has started to explore reasoning efficiency via test-time scaling (Muennighoff et al., 2025), dynamic early exit strategies (Yang et al., 2025), and token-length control (Han et al., 2024). Others have warned of performance degradation

\*Work done during an internship at Oracle.

from excessive reasoning (Ma et al., 2025; Chen et al., 2024b). However, these approaches typically target final performance, and offer no systematic way to evaluate or improve the trajectory of reasoning quality. Budget-aware techniques like BRPO (Qi et al., 2025) and mode selectors (Fang et al., 2025) optimize when to stop thinking, but do not address how to think better under constraints.

To fill this gap, we propose a framework for evaluating anytime reasoning in LLMs. Our framework introduces the Anytime Index, a metric that quantifies a model’s quality-per-token tradeoff across multiple reasoning budgets. This allows us to measure how solution quality evolves as reasoning progresses, capturing not just final accuracy, but the efficiency and consistency of reasoning over time. We apply this framework across diverse model families, including Grok-3, GPT-oss, GPT-4.1, GPT-4o, and LLaMA-3.3, to systematically compare their reasoning efficiency under budget constraints.

To address the question of how LLMs can become better anytime reasoners, we further propose an inference-time self-improvement method using LLM-generated preference data. Without human supervision, the model generates alternative reasoning traces at fixed token budgets and identifies those that lead to higher-quality intermediate solutions. These preference pairs are then reused as in-context examples to guide future reasoning. As an inference-time strategy, the approach is lightweight and broadly applicable across tasks and models. We evaluate this method across a diversity of tasks, including NaturalPlan (Zheng et al., 2024) for structured planning, AIME2024 for math reasoning, and GPQA-Diamond (Rein et al., 2023) for expert-level scientific reasoning. Overall, our approach consistently improves performance under fixed token budgets and outperforms strong CoT baselines in terms of Anytime Index and overall task performance.

Our contributions are three-fold:

- An extensible evaluation framework for anytime reasoning in LLMs, using the Anytime Index to quantify reasoning efficiency under token budgets.
- A scalable inference-time self-improvement method that leverages LLM-generated preference data without human supervision.
- Empirical evidence that this approach improves both intermediate and final solution

quality, outperforming strong CoT baselines on structured planning tasks.

## 2 Anytime Reasoning Evaluation Framework

### 2.1 Evaluation Pipeline

To evaluate whether current LLMs exhibit anytime reasoning capabilities, we propose a budget-aware evaluation framework that measures solution quality across varying reasoning lengths. The goal is to assess not just whether a model eventually arrives at a correct answer, but how efficiently it reaches high-quality intermediate solutions under a token budget.

The framework is illustrated in Figure 1. For each task input, we use Chain-of-Thought (CoT) prompting (Wei et al., 2022) to sample multiple full reasoning traces per model, up to a global maximum of 4,096 tokens for NaturalPlan and 16,384 tokens for AIME and GPQA. Each trace includes the model’s intermediate reasoning steps and final answer. We then define a series of token budget checkpoints  $b_1, b_2, \dots, b_n$ , where each  $b_i$  corresponds to the number of reasoning tokens observed. For NaturalPlan, we use budgets  $b_i \in 100, 200, \dots, 800$ ; for AIME and GPQA, we use  $b_i \in 200, 300, \dots, 1600$ .<sup>1</sup>

For each checkpoint  $b_i$ , we truncate the CoT trace to retain only the first  $b_i$  reasoning tokens, discarding the remainder. To simulate inference-time constraints, we re-prompt the model to generate a final answer (e.g., a trip itinerary or math solution), using only the truncated reasoning prefix as context. This setup emulates real-world scenarios where a system may be interrupted early and must produce an answer based on partial computation. Our evaluation closely mirrors this process by freezing the reasoning at  $b_i$  and measuring the quality of the resulting answer. To ensure robust evaluation, we sample  $N$  CoT traces per input, capturing variance from stochastic decoding. For each trace and budget  $b_i$ , we compute a task-specific quality score  $Q_i$ : for instance, Constraint Satisfaction Rate (CSR) for planning tasks, and accuracy for math and scientific QA tasks. Full metric definitions are provided in Section 4.3. We report final performance at each budget  $b_i$  as the average quality score across all sampled traces, yielding a stable and low-variance

---

<sup>1</sup>Budget ranges are adapted to each domain’s typical reasoning length.

estimate of model performance under varying computational budgets.

## 2.2 Anytime Index

To summarize a model’s overall efficiency under budget constraints, we propose the Anytime Index, inspired by area-under-the-curve (AUC) metrics in optimization (Qi et al., 2025). It is defined

$$\text{AnytimeIndex} = \frac{\sum_{t=1}^{T-1} \frac{(Q_t^* + Q_{t+1}^*)}{2} (b_{t+1} - b_t)}{(b_T - b_1) Q_{\max}},$$

where  $Q_t^* = \max_{i \leq t} Q_i$  denotes the best score achieved up to budget  $b_t$ , and  $Q_{\max}$  is the global upper bound across all methods. This metric rewards models that reach high-quality solutions quickly, rather than only at the end of a long reasoning trace.

Conceptually, the Anytime Index captures how quickly a model approaches high-quality solutions as the token budget increases. Two methods may achieve similar final performance at the largest budget, yet differ substantially in the trajectory by which they reach that level. A model that steadily improves and reaches high quality early will receive a higher Anytime Index than one that lags at smaller budgets and only improves sharply near the end. In this sense, the metric distinguishes “fast-thinking” models from “slow-thinking” ones, and is used for evaluating methods under budget-aware anytime reasoning. A detailed discussion regarding the novel insights from the Anytime Index is illustrated in Section A.

## 3 LLM-Generated Preference Data Prompting

To improve anytime reasoning under limited token budgets, we introduce a lightweight, inference-time self-improvement method that leverages LLM-generated preference data. This method teaches the model to produce higher-quality intermediate reasoning traces by contrasting good versus poor reasoning samples, all generated and evaluated by the model itself. No additional supervision or fine-tuning is required.

Our approach is grounded in the principle that a strong anytime reasoner should maximize the Anytime Index (Section 2), achieving high-quality solutions as early as possible within a token-constrained setting. Inspired by recent work on the self-improvement capabilities of LLMs (Huang et al., 2022; Madaan et al., 2023; Yao, 2024), we aim to guide reasoning generation through exposure to contrastive examples within the input prompt. We

describe the preference data generation pipeline and our prompting strategy below.

### 3.1 Preference Data Generation

Our method builds on the evaluation setup from Section 2.1. For each task input, we sample a large pool of  $N = 64$  CoT reasoning traces using the base model. Each trace is truncated at predefined token budget checkpoints  $b_i$ , and the corresponding output (e.g., a trip plan or math solution) is scored using task-specific metrics such as CSR or accuracy. This yields a set of quality scores  $Q_i$  for each trace at each budget.

At each budget  $b_i$ , we then rank all traces by their  $Q_i$  scores and construct preference pairs, each consisting of a preferred trace (higher  $Q_i$ ) and a rejected trace (lower  $Q_i$ ) of the same length. By fixing the budget, we isolate differences in reasoning quality, not length, ensuring that the contrastive supervision targets how the model reasons, not just how much it reasons. These budget-specific preference pairs form the core of our self-generated preference data, which we later use as in-context examples to guide reasoning during inference (Section 3). Example preference pairs are shown in Appendix B.1.

### 3.2 Preference Data Prompting

After generating preference pairs, we incorporate them as in-context examples to guide the model’s reasoning during inference. For each dataset, from the pool of model-generated preference pairs, we select the examples with the largest quality gap at each token budget  $b_i$ . This yields eight few-shot examples, one per budget checkpoint, for NaturalPlan, AIME, and GPQA.

Each example includes a preferred reasoning trace and its corresponding quality evaluation, contrasted with a rejected trace that yields a lower-quality solution under the same budget. Importantly, we omit the intermediate solutions from the prompt, encouraging the model to learn solely from contrastive reasoning patterns rather than imitating surface-level outputs. An example prompt containing the preference data is illustrated in Appendix B. Furthermore, we argue that our Preference Data Prompting method incurs minimal computational cost. See Section B.5 for a detailed analysis. In addition, we also describe the explicit narrative connection between the Anytime Index and Preference Data Prompting in Section B.6.

Model	Method	NaturalPlan (Trip)			AIME 2024			GPQA			Overall		
		Final	Avg	Anytime	Final	Avg	Anytime	Final	Avg	Anytime	Final	Avg	Anytime
<b>Grok-3</b>	Base	74.7	66.8	68.4	24.0	11.1	11.0	69.8	63.5	63.5	56.2	47.1	47.6
	LEAP	87.9	<b>76.8</b>	<b>79.1</b>	22.8	12.0	11.9	69.3	63.4	63.4	60.0	50.7	51.5
	PDP(+)	89.8	76.6	78.9	<b>25.0</b>	11.8	11.5	<b>70.3</b>	63.7	63.8	<b>61.7</b>	50.7	51.4
	PDP	<b>90.2</b>	76.0	78.1	24.9	<b>12.6</b>	<b>12.3</b>	69.7	<b>64.3</b>	<b>64.4</b>	61.6	<b>51.0</b>	<b>51.6</b>
<b>Grok-3-mini</b>	Base	81.5	76.8	84.7	80.6	75.2	80.9	<b>99.3</b>	<b>90.4</b>	<b>90.6</b>	87.1	80.8	85.4
	LEAP	90.2	84.7	85.4	86.7	77.9	81.7	95.7	90.0	<b>90.6</b>	90.9	84.2	85.9
	PDP(+)	85.7	85.4	87.6	83.3	79.0	82.0	96.9	85.9	85.8	88.6	83.4	85.1
	PDP	<b>90.7</b>	<b>85.6</b>	<b>89.7</b>	<b>100.0</b>	<b>86.0</b>	<b>87.1</b>	98.9	89.2	89.3	<b>96.5</b>	<b>86.9</b>	<b>88.7</b>
<b>GPT-oss-120B</b>	Base	36.7	37.9	45.9	32.0	<b>43.9</b>	<b>54.4</b>	44.3	51.5	63.8	37.7	44.4	54.7
	LEAP	46.1	38.3	45.2	30.2	40.5	49.8	36.8	47.9	63.6	37.7	42.2	52.9
	PDP(+)	<b>80.3</b>	<b>76.8</b>	<b>78.4</b>	30.2	38.8	46.5	50.6	54.4	62.5	53.7	56.7	62.5
	PDP	79.5	76.7	78.3	<b>38.9</b>	41.7	52.9	<b>69.4</b>	<b>66.2</b>	<b>66.2</b>	<b>62.6</b>	<b>61.5</b>	<b>65.8</b>
<b>GPT-oss-20B</b>	Base	51.5	46.4	47.7	16.5	22.7	40.4	28.4	43.5	58.2	32.1	37.5	48.8
	LEAP	36.1	33.6	37.8	9.8	25.5	33.7	21.0	36.8	56.7	22.3	32.0	42.7
	PDP(+)	<b>58.6</b>	45.3	47.3	13.9	28.9	38.6	34.2	47.3	58.2	35.6	40.5	48.0
	PDP	55.6	<b>47.5</b>	<b>50.8</b>	<b>17.7</b>	<b>29.1</b>	<b>40.6</b>	<b>60.7</b>	<b>56.4</b>	<b>60.7</b>	<b>44.7</b>	<b>44.3</b>	<b>50.7</b>
<b>GPT-4.1</b>	Base	69.4	68.2	69.6	2.8	4.3	7.2	50.9	57.7	61.7	41.0	43.4	46.2
	LEAP	74.6	<b>70.8</b>	<b>72.2</b>	<b>10.2</b>	<b>7.5</b>	<b>9.5</b>	47.7	58.2	63.7	44.2	45.5	<b>48.5</b>
	PDP(+)	<b>76.6</b>	70.7	71.9	1.2	3.4	6.6	52.9	57.2	62.3	43.6	43.8	46.9
	PDP	76.3	70.2	71.4	7.7	3.9	5.2	<b>69.4</b>	<b>66.2</b>	<b>67.1</b>	<b>51.1</b>	<b>46.8</b>	47.9
<b>GPT-4o</b>	Base	55.7	55.1	57.0	3.8	2.9	4.3	52.8	53.0	54.7	37.4	37.0	38.7
	LEAP	44.1	49.0	52.4	3.7	3.0	<b>5.0</b>	56.4	52.0	53.1	34.7	34.7	36.8
	PDP(+)	<b>66.2</b>	<b>60.9</b>	<b>62.3</b>	2.4	2.8	4.4	59.0	52.4	53.0	42.5	38.7	<b>39.9</b>
	PDP	62.8	58.7	60.2	<b>5.1</b>	<b>3.1</b>	4.3	<b>65.3</b>	<b>54.8</b>	<b>54.9</b>	<b>44.4</b>	<b>38.9</b>	39.8
<b>Llama-3.3-70B</b>	Base	71.5	73.6	74.1	17.1	23.6	28.7	39.3	46.4	52.5	42.6	47.9	51.8
	LEAP	76.2	74.6	78.4	<b>33.7</b>	21.7	26.3	42.8	44.7	52.0	<b>50.9</b>	47.0	52.2
	PDP(+)	81.0	76.6	79.1	24.1	24.6	28.6	45.4	47.3	<b>52.8</b>	50.2	49.5	53.3
	PDP	<b>82.0</b>	<b>78.2</b>	<b>80.2</b>	22.3	<b>25.3</b>	<b>29.0</b>	<b>48.1</b>	<b>47.9</b>	52.4	50.8	<b>50.5</b>	<b>53.9</b>

Table 1: Experiment results across three benchmarks: trip planning (NaturalPlan), math reasoning (AIME 2024), and scientific QA (GPQA). For each method, we report the Final/Avg CSR for NaturalPlan, and Accuracy for AIME and GPQA, labeled as "Final" and "Avg" for each corresponding column. The "Anytime" column represents the *Anytime Index* for each method. The "Overall" shows the macro-average across the three datasets (e.g., Overall Avg is the average of Trip Avg CSR, AIME Avg Accuracy, and GPQA Avg Accuracy). The best result in each column for each backbone model is highlighted in bold.

## 4 Experiments

### 4.1 Dataset

We evaluate anytime reasoning across three domains where solution quality can evolve with additional computation. We use **NaturalPlan** (Zheng et al., 2024) for structured trip planning, where intermediate solutions correspond to partially feasible itineraries evaluated via Constraint Satisfaction Rate (CSR); **AIME 2024** for multi-step mathematical reasoning, where accuracy improves as reasoning unfolds; and **GPQA-Diamond** (Rein et al., 2023) for expert-level scientific question answering that requires sustained, knowledge-intensive reasoning. Together, these datasets span planning,

mathematical, and scientific reasoning, enabling a comprehensive evaluation of budget-aware anytime behavior across domains. More dataset details are provided in Appendix E.

### 4.2 Models

We evaluate both reasoning-specialized and general-purpose LLMs. For reasoning models, we include Grok-3 and Grok-3-mini (xAI, 2025), GPT-oss-120B and GPT-oss-20B (OpenAI, 2025). For non-reasoning models, we evaluate GPT-4.1, GPT-4o, and Llama-3.3-70B (Meta, 2024). Each model is evaluated with four prompting strategies: (1) **Base**: Standard Chain-of-Thought (CoT) prompting following Wei et al. (2022), (2) **LEAP**:

context principle learning from mistakes (Zhang et al., 2024), which learns from explicitly abduced principles, a strategy that first induces and diagnoses mistakes on a few examples, then distills explicit reasoning principles from these mistakes, and finally applies both examples and principles during inference; (3) **PDP(+)**: An ablated variant of Preference Data Prompting, where only high-quality reasoning traces are used as in-context examples, omitting lower-quality alternatives from the preference pairs; (4) **PDP**: Preference Data Prompting is our full method, where models leverage self-generated preference-labeled CoT pairs to guide inference-time reasoning, as described in Section 3.

### 4.3 Evaluation Metrics

We evaluate model performance using the *Constraint Satisfaction Rate* (CSR) for NaturalPlan<sup>2</sup>, and standard *accuracy* for AIME and GPQA.

For NaturalPlan, CSR measures the extent to which a model-generated itinerary satisfies the structured constraints specified in the query. A rule-based constraint checker extracts constraints (e.g., number of cities, total trip length, and duration per city) using regular expressions and verifies whether each constraint is satisfied in the generated solution. CSR is computed as the fraction of satisfied constraints out of the total number of constraints for a given instance. We validate the correctness of this checker in Section D.

For AIME and GPQA, we adopt the standard accuracy metric, which evaluates whether the answer exactly matches the gold solution. At each token budget  $b_i$ , the corresponding CSR or accuracy score is used as  $Q_i$  when computing the Anytime Index (Section 2.2). In addition to the Anytime Index, we report the final CSR/accuracy at the maximum budget, as well as the average CSR/accuracy across all evaluated budgets.

### 4.4 Results

We evaluate whether LLMs exhibit effective anytime reasoning behavior and whether *Preference Data Prompting* (PDP) improves reasoning efficiency under token constraints. Table 1 reports final and average performance, and Anytime Index scores across NaturalPlan (Trip), AIME 2024, and GPQA for seven models and four prompting strategies.

<sup>2</sup>See Section C for a discussion of why we do not use the exact-match metric proposed in Zheng et al. (2024).

**Overall performance and model-specific trends.** Across all three benchmarks, Preference Data Prompting (PDP) consistently improves anytime reasoning performance relative to standard chain-of-thought (CoT) prompting. When averaged across datasets (*Overall* columns in Table 1), PDP achieves the highest Anytime Index for both reasoning-specialized models and Llama-3.3-70B. These improvements over the baselines are primarily driven by more effective intermediate reasoning traces, as evidenced by higher intermediate solution CSR/accuracy across varying token budgets. In most cases, these improvements also lead to better final performance at the maximum budget. For general-purpose non-reasoning models (GPT-4.1 and GPT-4o), PDP remains competitive and consistently improves over CoT, though it does not always attain the best Anytime Index among all prompting strategies. Example performance curves are shown in Figure 3. Overall, these trends suggest that while PDP is broadly applicable across model families, it is especially effective when models can reliably distinguish higher- and lower-quality reasoning traces. This observation is consistent with prior work showing that stronger reasoning models benefit more from self-generated preference signals (Pang et al., 2024; Chen et al., 2024a).

**The role of contrastive preference supervision.** To isolate the effect of preference contrast, we compare PDP against PDP(+), an ablation that performs in-context learning using only high-quality reasoning trajectories. PDP(+) almost always improves over CoT, confirming that prioritizing strong trajectories is beneficial. However, full PDP yields stronger anytime behavior overall. Across the 21 model–dataset settings, PDP achieves higher Anytime Index and higher average CSR/accuracy than PDP(+) in the majority of cases, and higher final performance in a substantial fraction of settings. This pattern is especially clear for reasoning-specialized models, where PDP substantially improves average and Anytime Index scores over PDP(+). These results indicate that explicitly contrasting higher- and lower-quality reasoning traces provides an additional signal beyond exposure to good examples alone.

**Comparison with learning-from-mistakes baselines.** We also compare PDP against LEAP (Zhang et al., 2024), a strong baseline that learns explicit principles from model-generated mistakes. While LEAP often outperforms vanilla

CoT, particularly on NaturalPlan and AIME, PDP consistently matches or surpasses LEAP in terms of the Anytime Index. As shown in Table 1, PDP achieves a higher *overall* Anytime Index than LEAP across six out of seven models, and it also demonstrates superior average CSR/accuracy in all cases. This advantage is most notable for models with strong reasoning capabilities, where PDP improves *overall* final CSR/accuracy by 17.6%, average CSR/accuracy by 11.4%, and the Anytime Index by 7.9. These results highlight that preference-based contrastive supervision more effectively guides intermediate reasoning under limited token budgets, outperforming in-context learning from principles derived from model-generated mistakes.

## 5 Conclusion

We presented an evaluation framework for anytime reasoning in LLMs, introducing the Anytime Index to quantify quality-per-token efficiency under budget constraints. Building on this, we proposed an inference-time self-improvement method that uses LLM-synthesized preference data to guide reasoning toward higher-quality intermediate outputs without human supervision. Experiments on trip planning show consistent gains in both task performance and Anytime Index, demonstrating that LLMs can learn to reason more efficiently through self-generated contrastive examples. In future work, we plan to extend this framework to additional domains and investigate how anytime reasoning can be internalized during training.

## 6 Limitations

While our work introduces a novel framework for evaluating and improving anytime reasoning in LLMs, several limitations remain. First, our experiments focus primarily on the trip planning domain using the NaturalPlan dataset. Although this domain is well-suited for evaluating structured, constraint-based reasoning under budget constraints, future work should validate the framework on a broader set of tasks, including open-ended domains such as code generation.

Second, we primarily compare standard Chain-of-Thought prompting with our proposed preference data prompting. While sufficient for demonstrating the utility of the Anytime Index, we do not exhaustively benchmark against other prompting strategies (e.g., Tree-of-Thoughts, Self-

Consistency, etc.) Expanding our evaluation to include a wider range of reasoning methods would provide deeper insights into the relative strengths of different approaches under budget constraints.

Finally, our self-improvement method is limited to inference-time prompting with LLM-generated preference pairs. While effective, this leaves open the opportunity to train models explicitly for better anytime behavior using preference-driven fine-tuning methods such as Direct Preference Optimization (DPO). If prompting already yields meaningful gains, integrating preference supervision into the training loop could further improve budget-aware reasoning and close the self-improvement cycle.

## References

- Kaiyuan Chen, Jin Wang, and Xuejie Zhang. 2024a. Learning to reason via self-iterative process feedback for small language models. *arXiv preprint arXiv:2412.08393*.
- Xingyu Chen, Jiahao Xu, Tian Liang, Zhiwei He, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, et al. 2024b. Do not think that much for  $2+3=?$  on the overthinking of o1-like llms. *arXiv preprint arXiv:2412.21187*.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.
- Thomas L Dean and Mark S Boddy. 1988. An analysis of time-dependent planning. In *AAAI*, volume 88, pages 49–54.
- Gongfan Fang, Xinyin Ma, and Xinchao Wang. 2025. Thinkless: Llm learns when to think. *arXiv preprint arXiv:2505.13379*.
- Tingxu Han, Zhenting Wang, Chunrong Fang, Shiyu Zhao, Shiqing Ma, and Zhenyu Chen. 2024. Token-budget-aware llm reasoning. *arXiv preprint arXiv:2412.18547*.
- Eric A Hansen and Shlomo Zilberstein. 2001. Monitoring and control of anytime algorithms: A dynamic programming approach. *Artificial Intelligence*, 126(1-2):139–157.
- Jiaxin Huang, Shixiang Shane Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. 2022. Large language models can self-improve. *arXiv preprint arXiv:2210.11610*.
- Wenjie Ma, Jingxuan He, Charlie Snell, Tyler Griggs, Sewon Min, and Matei Zaharia. 2025. Reasoning

- models can be effective without thinking. *arXiv preprint arXiv:2504.09858*.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.
- Meta. 2024. Llama 3.3 70b instruct: Text-only instruction-tuned model. <https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct>. Released December 6, 2024; supports 128 k context, improved reasoning and multilingual capabilities.
- Niklas Muennighoff, Zitong Yang, Weijia Shi, Xi-ang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.
- OpenAI. 2025. Gpt-oss-20b: Open-weight reasoning model. <https://openai.com/index/gpt-oss-model-card/>. Released under Apache 2.0; open-weight model capable of running on 16 GB hardware.
- Richard Yuanzhe Pang, Weizhe Yuan, He He, Kyunghyun Cho, Sainbayar Sukhbaatar, and Jason Weston. 2024. Iterative reasoning preference optimization. *Advances in Neural Information Processing Systems*, 37:116617–116637.
- Penghui Qi, Zichen Liu, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. 2025. Optimizing anytime reasoning via budget relative policy optimization. *arXiv preprint arXiv:2505.13438*.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. In *arXiv preprint arXiv:2311.12022*. GPQA-Diamond subset (198 expert-validated graduate-level science questions).
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- xAI. 2025. **Grok 3: The age of reasoning agents.** Model release / technical report, xAI. Beta introduction; reasoning & pretraining model.
- Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Zheng Lin, Li Cao, and Weiping Wang. 2025. Dynamic early exit in reasoning models. *arXiv preprint arXiv:2504.15895*.
- Liang Yao. 2024. Large language models are contrastive reasoners. *arXiv preprint arXiv:2403.08211*.
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.
- Tianjun Zhang, Aman Madaan, Luyu Gao, Steven Zheng, Swaroop Mishra, Yiming Yang, Niket Tandon, and Uri Alon. 2024. In-context principle learning from mistakes. *arXiv preprint arXiv:2402.05403*.
- Huaixiu Steven Zheng, Swaroop Mishra, Hugh Zhang, Xinyun Chen, Minmin Chen, Azade Nova, Le Hou, Heng-Tze Cheng, Quoc V Le, Ed H Chi, et al. 2024. Natural plan: Benchmarking llms on natural language planning. *arXiv preprint arXiv:2406.04520*.
- Shlomo Zilberstein. 1996. Using anytime algorithms in intelligent systems. *AI magazine*, 17(3):73–73.

## A Novel Insights for Anytime Index

### A.1 What does Anytime Index measure conceptually?

The Anytime Index is a summary of a model’s entire anytime performance profile, i.e., how its expected solution quality changes as the token budget increases. This follows the classical view of anytime algorithms (Zilberstein, 1996), where one characterizes an algorithm by its performance curve (quality vs. time) and often summarizes that curve via an expectation or utility over time. In our context, the Anytime Index makes explicit how quickly a method approaches high quality: whether it is “fast-thinking” (reaches high CSR early and then plateaus) or “slow-thinking” (stays weak until very large budgets). Two methods can have very similar Final CSR but quite different Anytime Index values if one achieves good quality much earlier; this is exactly the distinction we care about for judging good anytime reasoners.

In our setting: 1) The curve is CSR (or task accuracy) as a function of token budget ( $b$ ) (e.g., 100–800, 200–1600, depending on the task and model); 2) Anytime Index is essentially the area under this curve, normalized to  $[0,1]$ . This is analogous to how AUC–ROC summarizes classifier performance across all thresholds rather than focusing on a single operating point. Intuitively, Anytime Index answers the question: “If I do not know in advance exactly how many tokens I will be allowed (e.g., interruption, latency budget, dynamic stopping), how good is this model on average over the whole budget range?” By contrast, Final CSR only reports quality at one point (the largest budget) and cannot distinguish methods that reach that quality quickly from those that only get there at the very end.

### A.2 Why can rankings coincide, but the metric is still useful?

It is true that, for the experiments reported in the main table, the ordering of prompting methods and models by Anytime Index and by Final CSR is largely consistent. This outcome is not surprising: 1) The methods that perform better at the largest budget also tend to be better (or at least not much worse) at intermediate budgets; 2) The budget span on NaturalPlan is relatively narrow (e.g., 100–800 tokens), so methods that are clearly better at 800 tokens generally perform better across the entire small range as well.

### A.3 Scenario where Anytime Index provides strictly more nuanced insights than other metrics.

While our main results emphasize that PDP improves both Final CSR and Anytime Index, the metric is particularly informative in the following concrete scenario: **Similar final CSR, different “speed” of reasoning.** Consider Figure 2a, both baseline CoT and PDP achieve Final CSR  $\approx 0.95$  at 800 tokens. PDP reaches 0.9 CSR already at 350 tokens and then slowly improves, while baseline CoT stays near 0.83–0.85 CSR until it jumps up at 500–600 tokens, then: Final CSR would say PDP  $\approx$  baseline CoT; however, Anytime Index is higher for PDP, indicating that PDP makes Grok3-mini a better anytime reasoner under variable budgets.

We will highlight specific pairs of methods where AnyIndex reveals a larger gap than Final CSR suggests (e.g., methods whose Final CSR is close but whose early- and mid-budget behavior differs markedly). Hence, Anytime Index is not introduced to produce a different ranking in our particular tables, but to (i) align LLM-based reasoning with the classical analysis of anytime algorithms, (ii) provide a robust, budget-agnostic summary of performance over the whole budget range, and (iii) enable future work to distinguish methods that have similar final quality but very different anytime behavior.

## B Preference Data Prompting Details

For NaturalPlan, we synthesize preference pairs for each model using the five few-shot examples provided by the dataset, which we use to construct in-context learning (ICL) prompts. Specifically, we sample  $N = 64$  CoT traces per example and apply the pipeline described in Section 3.1 to generate preference pairs used in Preference Data Prompting. For AIME and GPQA, we similarly create preference data by randomly sampling 5 and 30 validation examples, respectively. For each sampled example, we generate  $N = 32$  CoT traces and construct preference pairs using the same method. An illustrative preference pair from each dataset and the corresponding prompting template are provided below.

### B.1 Example Preference Pair from NaturalPlan

We present example preference pairs generated by GPT-OSS-20B at the 100-token budget checkpoint.

The preferred reasoning trace yields a solution with 100% constraint satisfaction, while the rejected trace achieves 0%, failing to account for overlapping days between travel and stay.

### B.1.1 Example task in NatualPlan

You plan to visit 6 European cities for 18 days in total. You only take direct flights to commute between cities. You plan to stay in Bucharest for 5 days. From day 6 to day 10, there is an annual show you want to attend in Bucharest. You would like to visit Prague for 2 days. You want to spend 2 days in Budapest. You plan to stay in Split for 5 days. You plan to stay in Tallinn for 4 days. You have to attend a workshop in Tallinn between day 1 and day 4. You plan to stay in Dublin for 5 days. During day 10 and day 14, you have to attend a conference in Dublin.  
Here are the cities that have direct flights:  
Prague and Budapest, Tallinn and Prague, Bucharest and Dublin, Budapest and Bucharest, Prague and Bucharest, Prague and Split, Dublin and Split, Tallinn and Dublin, Budapest and Dublin, Prague and Dublin.  
Find a trip plan of visiting the cities for 18 days by taking direct flights to commute between them.

### B.1.2 Preferred reasoning trace at the 100-token budget checkpoint

Sentences contributing to effective reasoning are highlighted in green.

analysisWe need schedule 18 days total. Cities: Bucharest 5 days (including show day6-10). Tallinn 4 days workshop day1-4. Dublin 5 days (conference day10-14). Prague 2 days. Budapest 2 days. Split 5 days. **Total days sum =5+4+5+2+2+5=23 exceeds 18, so must overlap some stays? Actually you can be in city and attend events overlapping with stay counts.** Must allocate days exactly as durations but maybe some cities visited partially? In examples **durations equal stay days. Here total exceeds so need ordering with overlapping constraints?** maybe some cities can be visited same day as flight? In examples they

### B.1.3 Rejected reasoning trace at the 100-token budget checkpoint

Sentences contributing to poor reasoning are highlighted in red.

analysisWe need schedule 18 days total. Constraints: Bucharest 5 days total, with show day6-10 (so must be in Bucharest those days). Prague 2 days, Budapest 2 days, Split 5 days, Tallinn 4 days (workshop day1-4 must be in Tallinn). Dublin 5 days, with conference day10-14 must be in Dublin.

We need to fit timeline. Start day1 in Tallinn for 4 days (day1-4). Then need to be in Bucharest day6-10, so must travel before day6. Also Dublin conference day10-14, overlapping with Bucharest show day6-10. Day10 both places can't. Need transition: maybe leave Bucharest day10 to Dublin arriving day10, but conference starts day10, so need be in Dublin day10. So Bucharest stay must be day5-9? Show is day6-10 inclusive, but if we leave day10 early maybe still attend? Show day6-10 inclusive, need be there through day10. Conflict. So maybe order different: go to Dublin first, conference day10-14, then Bucharest day6-10 can't after. **Overlap impossible.** Need overlapping days? maybe we can split

### B.2.1 Example question in AIME

Let  $b \geq 2$  be an integer. Call a positive integer  $n$   $b$ -eautiful if it has exactly two digits when expressed in base  $b$ , and these two digits sum to  $\sqrt{n}$ . For example, 81 is 13-beautiful because

$$81 = 6 \cdot 3_{13} \quad \text{and} \quad 6 + 3 = \sqrt{81}.$$

Find the least integer  $b \geq 2$  for which there are more than ten  $b$ -beautiful integers.

### B.2.2 Preferred reasoning trace at the 1200-token budget checkpoint

Sentences contributing to effective reasoning are highlighted in blue.

analysisWe need count of  $n$  with representation two digits in base  $b$ :  $n = a \cdot b + c$  with  $0 \leq c < b$ ,  $1 \leq a \leq b - 1$  (since two digits, leading digit nonzero). Condition:  $a + c = \sqrt{n}$ . So  $\sqrt{n}$  must be an integer; let  $\sqrt{n} = k$ . Then  $n = k^2$ . Also  $n = ab + c$  with  $a + c = k$ . So we need to find  $k$  such that  $k^2$  expressed in base  $b$  has digits  $a, c$  satisfying  $a + c = k$ . We have  $a = \lfloor k^2/b \rfloor$  and  $c = k^2 \bmod b$  with  $1 \leq a \leq b - 1$ ,  $0 \leq c < b$ . Thus we need  $k$  such that  $\lfloor k^2/b \rfloor + (k^2 \bmod b) = k$ . Let  $k^2 = ab + c$ , so  $a = (k^2 - c)/b$ . Equation:  $(k^2 - c)/b + c = k \implies (k^2 + c(b - 1))/b = k \implies k^2 + c(b - 1) = kb$ . Thus  $c(b - 1) = k(b - k)$  and so  $c = \frac{k(b-k)}{b-1}$  with  $0 \leq c < b$ . So for each  $k$  we require  $(b - 1) \mid k(b - k)$ . Let  $d = b - 1$ . Then  $c = \frac{k(b-k)}{d}$  must be an integer, and  $a = k - c$  must satisfy  $1 \leq a \leq b - 1$ . Since  $n = k^2 < b^2$ , we have  $1 \leq k \leq d$ . Thus we need  $k$  such that  $(b - 1) \mid k(b - k)$ , i.e.,  $d \mid k(b - k)$ . Since  $b = d + 1$ ,  $k(b - k) \equiv k(1 - k) \pmod{d}$ . Therefore  $d \mid k(1 - k) \iff k(k - 1) \equiv 0 \pmod{d}$ . So the condition simplifies to  $d \mid k(k - 1)$ . Hence we count  $k$  for which  $k(k - 1)$  is a multiple of  $d$ . For each prime power  $p^e \mid d$ ,  $k \equiv 0 \text{ or } 1 \pmod{p^e}$ . By CRT, the total number of solutions is  $2^{\omega(d)}$ , where  $\omega(d)$  is the number of distinct prime factors of  $d$ . We need  $2^r > 10$ , so  $r \geq 4$ . The smallest  $d = b - 1$  with four distinct primes is  $2 \cdot 3 \cdot 5 \cdot 7 = 210$ , so  $d = 210$  and  $b = 211$ .

### B.2 Example Preference Pair from AIME

We present example preference pairs generated by GPT-OSS-120B at the 1200-token budget checkpoint. The preferred reasoning trace generates a correct solution, while the rejected trace generates an incorrect solution, imposed by a subtle but fatal logic error.

### B.2.3 Rejected reasoning trace at the 1200-token budget checkpoint

Sentences contributing to poor reasoning are highlighted in red.

analysis We need count numbers  $n$  with two digits in base  $b$ :  $n = a*b+c$  where  $a, c$  are digits  $0 \leq c < b$ ,  $1 \leq a < b$  (since two-digit, leading digit nonzero). Sum  $a + c = \sqrt{n}$ . So  $\sqrt{n}$  integer  $\Rightarrow n$  is perfect square. Let  $n = k^2$ . Then  $k = a + c$ . Also  $n = ab + c = k^2$ . So we have  $ab + c = (a + c)^2$ .  
 We have  $0 \leq c \leq b - 1$ ,  $1 \leq a \leq b - 1$ .  
 Let's rearrange:  $ab + c = a^2 + 2ac + c^2 \Rightarrow$  bring terms:  $ab = a^2 + 2ac + c^2 - c = a^2 + 2ac + c(c - 1)$ . So  
 $ab = a^2 + 2ac + c(c - 1)$ .  
 Solve for  $b$ :  $b = (a^2 + 2ac + c(c - 1))/a = a + 2c + c(c - 1)/a$ . Since  $b$  integer, **a must divide  $c(c - 1)$** . Let's denote  $d = c(c - 1)$ . Then  $b = a + 2c + d/a$ .  
 We need  $b \geq \max(a, c) + ?$  Actually  $b$  is base, must be  $> \max(a, c)$ , i.e.,  $b > \max(a, c)$ . Since digits less than  $b$ , condition satisfied automatically if  $b > a, c$ .  
 We need count of  $(a, c)$  producing same  $b$ ? For a given  $b$ , we count solutions  $(a, c)$  with  $1 \leq a \leq b - 1$ ,  $0 \leq c \leq b - 1$ , and  $b = a + 2c + d/a$  (integer). Equivalent to  $a$  divides  $d$ .  
 Let's brute think: For each  $a$ , choose  $c$  where  $m = c(c - 1)/a$ . Then  $b = a + 2c + m$ . This determines  $b$ . So for each  $(a, c)$  we get a  $b$ . We want  $b$  such that many  $(a, c)$  map to same  $b$ .  
 So need collisions. Let's compute small  $b$  maybe 2, 3... but likely need larger.  
 We could think of equation:  $b = a + 2c + c(c - 1)/a$ . Define  $t = c$ . Then  $b = a + 2t + (t(t - 1))/a$ . For given  $b$ , we look for integer  $a$  dividing  $t(t - 1)$  and satisfying this equality.  
 We can attempt to find many solutions by choosing  $t$  such that many divisors  $a$  produce same  $b$ ? But  $b$  depends on  $a$ . Probably the minimal  $b$  is 7? Let's try quick enumeration mental maybe  $b = 5$ ? ... (lengthy enumeration attempt) ... So only one for  $b = 5$ . Need  $> 10 \Rightarrow$  larger. **Probably  $b = 10$ ?** Let's approximate count roughly equals number of divisors? Might be around  $b$ ? Need  $> 10$ ,

analysis We need identify  $X$ : equimolar mixture of two liquids that decolorize bromine water  $\rightarrow$  they are alkenes (or phenols). No conjugated double bonds. So likely two alkenes reacting with Br<sub>2</sub>. Heated with Pt causes disproportionation to  $Y$ : mixture of two other liquids not decolorizing bromine  $\Rightarrow$  alkanes. Disproportionation of alkenes over Pt gives alkanes (hydrogenation) but also maybe cracking? Actually Pt at high temp causes hydrogen transfer: 2 alkenes  $\rightarrow$  one alkane more substituted, one alkene less? But result  $Y$  does not decolorize, so both are alkanes. Hydrogenation of  $X$  and  $Y$  gives same hydrocarbon  $Z$ , hydrogen mass fraction 14.28%  $\rightarrow$  compute formula. Let  $Z$  be  $C_nH_m$ . Mass% H =  $m/(12n + m) \cdot 100 = 14.28$ . Solve approx: try  $n = 6, m = 14$ ? mol weight = 86, H% = 16.3 too high.  $n = 8, m = 18$ : 15.8.  $n = 10, m = 22$ : 15.5. Need 14.28  $\rightarrow$  maybe  $n = 12, m = 26$ : 15.3. lower. try  $n = 14, m = 30$ : 15.15. Need lower. maybe  $n = 4, m = 10$ : 17.2. Hmm. Could be benzene?  $C_6H_6$ : 7.7% no. Maybe cyclohexane  $C_6H_{12}$ : 12/84 = 14.29%. **Yes! So  $Z =$  cyclohexane.**

So  $Y$  mixture contains cyclohexane as constituent. The other liquid perhaps also?  $Y$  is two liquids, both alkanes, one is cyclohexane, the other likely something else from disproportionation of  $X$ . Starting  $X$  are two alkenes equimolar that hydrogenate to cyclohexane  $\rightarrow$  **both must be  $C_6H_{10}$  isomers (cyclohexene-type)**. Two alkenes that hydrogenate to cyclohexane: could be 1-hexene and cyclohexene? But 1-hexene gives hexane, not cyclohexane. Need both give cyclohexane after Pt disproportionation? Actually Pt heated causes alkene to undergo isomerization and disproportionation: two alkenes give one alkane with longer chain and one alkene shorter. Eventually both become cyclohexane after hydrogenation. Likely  $X$  are cyclohexene ( $C_6H_{10}$ ) and methylcyclopentene? Both  $C_6H_{10}$ . Two isomeric cyclohexenes? They are liquids, decolorize Br<sub>2</sub>, no conjugated. **Both  $C_6H_{10}$ , so total H = 10 + 10 = 20. But options include 18, 22, 16, 12.** Maybe  $X$  are cyclopentene  $C_5H_8$  and 1-hexene  $C_6H_{12}$  (both decolorize?). But total H = 20 not listed. Consider that after disproportionation,  $Y$  are cyclohexane ( $C_6H_{12}$ ) and another alkane not decolorizing, maybe benzene? No. Maybe  $X$  are two alkenes  $C_3H_6$  and  $C_5H_8$  totals H=14 not option. Check options, likely 22? maybe

### B.3 Example Preference Pair from GPQA

We present example preference pairs generated by GPT-OSS-120B at the 800-token budget checkpoint. The preferred reasoning trace selects the correct option, while the rejected trace opts for an incorrect option, resulting from incorrect assumptions and key constraint violations.

#### B.3.1 Example question in GPQA

If an equimolar mixture  $X$  of two liquids, which decolorizes bromine water, is treated with platinum when heated, then an equimolar mixture  $Y$  of two other liquids is formed as a result of disproportionation, which does not decolorize bromine water. Hydrogenation of both mixture  $X$  and mixture  $Y$  in the presence of platinum under rather severe conditions gives only one substance, a certain hydrocarbon  $Z$  (mass fraction of hydrogen is 14.28%), which is a constituent of mixture  $Y$  and widely used as a solvent. Substance  $Z$  does not react further with hydrogen. There are no conjugated multiple bonds in the molecules of the compounds of mixture  $X$ .

Indicate the total number of hydrogen atoms in two liquids of mixture  $X$ .

- 18
- 22
- 16
- 12

#### B.3.3 Rejected reasoning trace at the 800-token budget checkpoint

Sentences contributing to poor reasoning are highlighted in red.

analysis We need identify  $X$ : equimolar mixture of two liquids that decolorize bromine water (so they have C=C double bonds). No conjugated multiple bonds. They undergo disproportionation on Pt heated giving  $Y$ : mixture of two other liquids that does NOT decolorize bromine water (so saturated). Hydrogenation of both  $X$  and  $Y$  under severe conditions gives only one substance  $Z$ , hydrocarbon with H mass fraction 14.28% = H mass/M = 0.1428. Let M = molecular weight, H mass = (n)/M = 0.1428  $\Rightarrow$  n = 0.1428M. For hydrocarbon  $C_xH_y$ , M = 12x + y. So y/(12x + y) = 0.1428. Solve approximate: try x = 3, y ≈ 6 giving  $C_3H_6$ . **But  $C_3H_6$  is unsaturated and  $Z$  "does not react further with hydrogen," so this violates the condition.** Next try x = 4, giving y ≈ 8  $\rightarrow$   $C_4H_8$ . **But  $C_4H_8$  is also an alkene and cannot be  $Z$ .**

Concluding  $Z$  might be  $C_3H_6$  or  $C_4H_8$  contradicts the requirement that  $Z$  is a fully saturated solvent.

Then " $Z$  does not react further with hydrogen"  $\rightarrow$  saturated  $\rightarrow$  maybe cyclohexane? They compute  $C_6H_{12}$ : 12/84 = 0.1429 matches. So  $Z$  = cyclohexane. Good.

So  $Y$  mixture contains cyclohexane and another liquid. Disproportionation of alkenes on Pt yields alkanes? Starting  $X$  are two alkenes equimolar that decolorize bromine water, no conjugated double bonds, so likely 1-hexene and 2-hexene? But need total H atoms in two liquids of mixture  $X$ . Each molecule formula? Disproportionation (alkene coupling?) On Pt heated, alkenes undergo disproportionation:  $2RCH = CH_2 \rightarrow RCH_2 - CH_2R + RCH = CH_2$ ? Actually yields alkane and alkene? Might produce mixture  $Y$  of two liquids not decolorizing bromine: likely cyclohexane and butane? **But these products do not follow from the given reaction mechanism, and the chain lengths do not match the required hydrogenation product.**

Starting  $X$  could be 1-hexene and 2-hexene (both  $C_6H_{12}$ ). Total H per molecule = 12 each  $\Rightarrow$  24 H atoms. **But 1-hexene hydrogenates to hexane, not cyclohexane, contradicting earlier conclusions.**

Not match. Maybe  $X$  are  $C_4H_8$  (butenes) total H = 16 total matches option C. **But  $C_4H_8$  hydrogenates to butane, not cyclohexane, so this is chemically impossible.**

#### B.4 Prompt used for Preference Data Prompting

Below, we present the prompt template used for Preference Data Prompting. Each in-context example includes the task query, a preferred (good) reasoning trace, and a rejected (bad) reasoning trace.

#### B.3.2 Preferred reasoning trace at the 800-token budget checkpoint

Sentences contributing to effective reasoning are highlighted in blue.

We provide one example pair for each token budget checkpoint. Below is an example prompt we used for NaturalPlan (Trip), in which case each intermediate solution quality is evaluated with CSR:

```

Here are examples showing the difference between GOOD and POOR
reasoning approaches under different progress checkpoints. You
should follow the GOOD reasoning approach to reason step by step.

[Scoring] CSR = (# satisfied constraints) / (total constraints).

[Examples @ Token Budgets]
- Pair (@ Token Budget 100 tokens) -
Question: [EXAMPLE_QUERY]

  • GOOD REASONING (Higher CSR)[CSR=0.8]: [PREFERRED_TRACE] Why
    GOOD:- Satisfaction Rate: 80.0% - Satisfied: 8/10 constraints

  • POOR REASONING (Lower CSR) [CSR=0.2]: [REJECTED_TRACE] Why
    POOR:- Satisfaction Rate: 20.0% - Satisfied: 2/10 constraints

- Pair (@ Token Budget 200 tokens) -
... (omitted preference pairs at budget 200-800 tokens)

Now, please follow the GOOD reasoning traces in the examples to reason
step by step to solve the target problem: [TARGET_QUERY]

```

## B.5 The computational cost of Preference Data Prompting

**PDP does not inherently require a large pool of reasoning traces.** PDP uses multiple CoT traces per example to construct preference pairs, but the number of traces ( $N$ ) is a tunable hyperparameter, not a fixed requirement. In our experiments on NaturalPlan, we chose ( $N = 64$ ) to capture broader variation in traces and obtain more informative preference pairs, but much smaller values of ( $N$ ) (e.g., 8–16) are also possible if compute is limited. Importantly, trace generation and pair selection are one-time, offline steps per model–dataset setting. Once this pool of traces is generated and preference pairs are constructed, the same data is reused for all PDP runs and all budget settings.

**PDP does not require running on the entire dataset or having a “complete” dataset to extract ICL examples.** In our experiments for NaturalPlan, we rely only on the five few-shot examples provided by the benchmark and use them to extract example pairs and construct ICL-style prompts. We do not require a complete training dataset or any additional labeled data beyond these standard few-shot examples. In other domains (e.g. math reasoning), one could similarly use a small set of seed samples (e.g. randomly sample 5 samples from MATH) to generate traces and preference pairs. PDP is compatible with low-data regimes and does not depend on large-scale data.

## B.6 The connection between Anytime Index and Preference Data Prompting

Anytime Index is an evaluation construct: it summarizes the entire quality–budget curve (CSR/accuracy vs. token budget) into a single scalar, in the spirit of performance profiles for anytime algorithms or AUC over trade-off curves in classical evaluation. PDP is a training / prompting construct: it is designed to reshape that quality–budget curve, making the model achieve higher CSR earlier and more consistently across budgets.

Anytime Index effectively measures the (normalized) area under the quality-vs-budget curve over a fixed budget range. A method that: 1) achieves high CSR only at the very largest budget but stays low elsewhere will have a relatively low Anytime Index; 2) achieves moderate-to-high CSR across many budgets, especially early ones, will have a higher AnyIndex, even if the final CSR is similar. PDP is designed to favor the second type of behavior. We construct preference pairs at fixed budgets ( $b_i$ ): for each ( $b_i$ ), we compare reasoning traces that lead to higher vs. lower CSR under that same budget. The model performs in-context learning to prefer reasoning patterns that yield higher CSR at each ( $b_i$ ), not just at the maximum budget. Because we do this across all token budgets, PDP encourages trajectories that are consistently good across the whole budget range, which is precisely what leads to a larger Anytime Index.

In summary, **Anytime Index and PDP are not two unrelated contributions**: Anytime Index provides the formal objective for budget-aware anytime evaluation, and PDP is specifically designed to improve Anytime Index.

## C Details of the Evaluation Metric of NaturalPlan (Trip)

In the trip planning domain in NaturalPlan, there are often many itineraries that are fully valid but do not match the single reference plan in NaturalPlan. Exact Match (EM), by construction, only counts a solution as correct if it matches the ground-truth itinerary exactly (including order, city sequence, and phrasing). In our experiments, we observe numerous cases where: the model’s itinerary satisfies 100% of the constraints (CSR = 1.0), but differs in benign ways from the reference (e.g., a different but feasible city order, or alternative use of valid flight edges), and EM therefore assigns a score of 0. For anytime reasoning, this is prob-

lematic: two itineraries that are equally valid ( $\text{CSR} = 1.0$ ) are treated as different by EM, making it hard to measure improvements in solution quality. Additionally, our core contribution is an evaluation framework for how solution quality improves as reasoning tokens increase.  $\text{CSR}$  provides a measure of partial progress (e.g.,  $6/10 \rightarrow 8/10 \rightarrow 10/10$  constraints satisfied) and thus yields meaningful intermediate scores at every token budget.

## D Constraint Satisfaction Checker Validation

We randomly sampled 100 queries from the validation set we constructed from the NaturalPlan (Trip) dataset. For each query, we used GPT-4.1 to generate a trip itinerary, resulting in 100 model outputs. Then, for each output itinerary, we ran our automatic constraint checker to compute the constraint satisfaction rate ( $\text{CSR}$ ) and to explicitly list which constraints were satisfied vs. unsatisfied for each query. Next, two of the lead authors jointly inspected all 100 cases through discussion, comparing the checker’s reported  $\text{CSR}$  and satisfied/unsatisfied constraints against a careful manual assessment for each itinerary. In 95 out of 100 cases, the checker’s output was exactly correct (both the overall  $\text{CSR}$  and the breakdown of satisfied vs. unsatisfied constraints). The remaining 5 cases involved minor discrepancies (e.g., wrong interpretations of constraints). We consider the constraint checker accurate and reliable for large-scale evaluation.

## E Detailed Dataset Descriptions and Hyper-parameter Configurations

The NaturalPlan dataset (Zheng et al., 2024) contains three task types: trip planning, meeting planning, and calendar scheduling. We focus on the trip planning task, which poses the most complex reasoning challenges. It consists of 1,600 instances, which we split evenly into validation and test sets. The validation set is used for tuning hyperparameters, including the number of sampled CoT traces ( $N$ ), token budget checkpoints ( $b_i$ ), and generation length limits.

For the evaluation pipeline, we use  $N = 3$  to simulate diverse but efficient reasoning traces. For preference data generation, we increase the sampling to  $N = 64$  to capture broader trace variation. We define token budget checkpoints at  $b_i \in \{100, 200, 300, 400, 500, 600, 700, 800\}$ , align-

ing with typical CoT lengths across models. The maximum token limits for generating both CoT traces and solutions are set to 4,096 tokens.

We also experimented with different decoding settings and found that  $\text{temperature} = 0.7$ ,  $\text{top-}p = 1$ , and  $\text{top-}k = 1$  yielded the most stable results. Additionally, we incorporate the five-shot examples provided in NaturalPlan as part of the input when constructing preference data for each model.

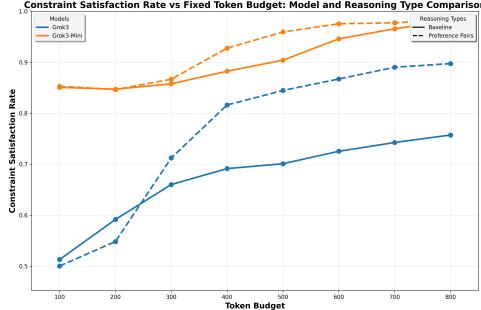
## F Qualitative analysis of results in NaturalPlan

Figure 2 plots constraint satisfaction rate ( $\text{CSR}$ ) curves for Grok-3, GPT-4.1, and GPT-4o, comparing baseline CoT prompting with our Preference Data Prompting (PDP) approach. Across all models, PDP matches baseline performance at lower token budgets (e.g., 100, 200, 300), but begins to outperform CoT at later checkpoints. This trend underscores PDP’s advantage in producing higher-quality solutions as more reasoning tokens become available.

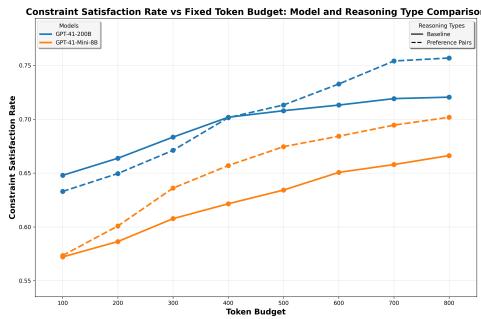
Notably, Grok-3 benefits the most from PDP, surpassing the baseline CoT starting at the second token budget checkpoint (200 tokens). This suggests that reasoning-specialized models like Grok-3 are especially adept at leveraging preference data for self-improvement, resulting in stronger anytime reasoning performance under our PDP approach.

For GPT-4.1 and GPT-4o, the performance trend reveals a key distinction. The smaller variants of GPT-4.1 and GPT-4o exhibit continued improvement even beyond the final token budget checkpoint (800 tokens), whereas the larger models plateau. This indicates greater potential for improvement in smaller models and suggests that our approach can be particularly effective in helping them develop stronger anytime reasoning capabilities. Notably, the performance of GPT-4o shows a non-monotonic trend: as reasoning length increases, performance peaks at a medium budget and then slightly decreases. This decline is due to two main factors: (1) longer traces lead to drift and self-revision, where initial valid plans are altered and sometimes introduce inconsistencies that violate previously satisfied constraints, and (2) longer reasoning provides more opportunities for small errors that accumulate, reducing the fraction of satisfied constraints. These effects are visible in GPT-4o’s  $\text{CSR}$  curve, where performance diminishes

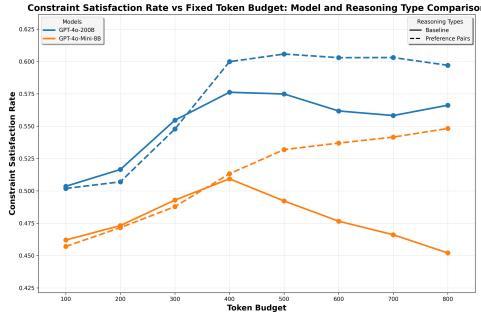
slightly at larger token budgets. Our framework is designed to surface such non-monotonic behaviors, highlighting that more reasoning is not always better and that the quality of intermediate steps is crucial for final performance.



(a) CSR curve for Grok-3 and Grok-3-mini.

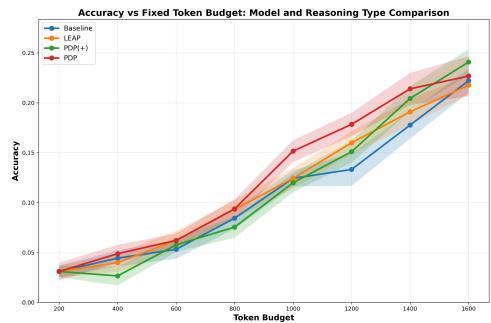


(b) CSR curve for GPT-4.1 and GPT-4.1-mini.

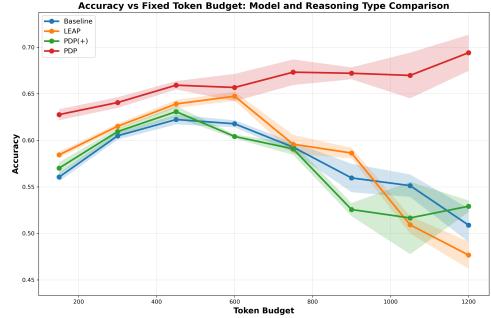


(c) CSR curve for GPT-4o and GPT-4o-mini.

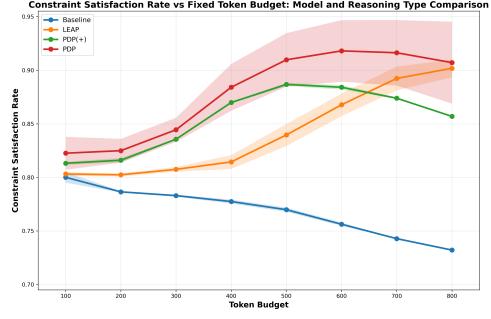
Figure 2: The Constraint Satisfaction Rate evaluated at different token budget checkpoints across different model families: Grok-3, GPT-4.1, and GPT-4o. Preference Data Prompting (dotted line) makes the models better anytime reasoners.



(a) Accuracy curve for Grok-3 on AIME.



(b) Accuracy curve for GPT-4.1 on GPQA.



(c) CSR curve for Grok-3-mini on NaturalPlan.

Figure 3: The accuracy (on AIME and GPQA) and constraint satisfaction rate (on NaturalPlan) evaluated at different token budget checkpoints across different models: Grok-3, Grok-3-mini, and GPT-4.1. Compared to other prompting techniques, Preference Data Prompting (red line) makes the models better anytime reasoners.