Towards Label-Free Biological Reasoning Synthetic Dataset Creation via Uncertainty Filtering

Josefa Lia Stoisser Novo Nordisk
Lawrence Phillips Novo Nordisk
Aditya Misra Novo Nordisk
Tom A. Lamb University of Oxford
Philip Torr University of Oxford
Marc Boubnovski Martell Novo Nordisk
Julien Fauqueur Novo Nordisk
Kaspar Märtens Novo Nordisk

Abstract

Synthetic chain-of-thought (CoT) traces are widely used to train large reasoning models (LRMs), improving generalization by providing step-level supervision. Yet most approaches require ground-truth labels to seed or filter these traces—an expensive bottleneck in domains like biology where wet-lab data are scarce. We propose a label-free alternative: *uncertainty-based filtering*, which uses a model's own confidence—quantified through established uncertainty metrics like self-consistency and predictive perplexity—as a substitute for external labels. We sample multiple reasoning traces and retain only low-uncertainty subsets. Applied to biological perturbation prediction, a domain where wet-lab labels are especially costly, we show that the filtered subset has higher accuracy, and that supervised fine-tuning (SFT) on uncertainty-filtered data outperforms unfiltered synthetic data, narrows the gap to ground-truth training, and surpasses strong LRM baselines. Ablations show that per-class filtering corrects for class-specific uncertainty scales and that hybrid uncertainty metrics yield higher-quality datasets. Our results suggest that model-internal confidence is a powerful signal for efficient reasoning dataset creation, enabling LRMs in domains where supervision is expensive.

1 Introduction

Synthetic chain-of-thought (CoT) traces have become a cornerstone for training large reasoning models (LRMs), providing step-level supervision that improves generalization across mathematics, coding, and symbolic tasks [9, 11, 8]. However, most pipelines for generating such traces rely on ground-truth labels to filter sampled generations [9, 11, 8]. While feasible in domains with abundant labels or automatic checkers, this creates a bottleneck where high-quality labels are costly or unavailable.

Applications in biology particularly highlight this challenge. Ground-truth labels, when available at all, often require costly experimental measurement, limiting the scale of supervision. In particular, *cellular perturbation prediction*—predicting how a given perturbation (e.g. drug or gene knockout) affects target gene expression levels (up, down, or unchanged)—is a fundamental task underlying drug discovery and disease modeling. The challenge is compounded by fundamental epistemic uncertainty: even when outcomes can be measured, the underlying causal mechanisms (e.g. gene regulatory networks) remain poorly understood, precluding external validation of synthetic reasoning traces [20]. Moreover, approaches that distill carefully curated reasoning traces into open-source models have shown to achieve task-specific performance that exceeds that of frontier LRMs [14].

We address these challenges with *uncertainty-filtered synthetic reasoning*. Our method samples multiple reasoning traces per example and filters them using the model's *own confidence*—quantified

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Efficient Reasoning.

by self-consistency and predictive perplexity—without any external supervision or verifiers. This approach aims to simultaneously mitigate label scarcity (by reducing dependence on wet-lab outcomes) and guard against epistemic gaps (by discarding examples where the model itself is least confident), to yield cleaner and more reliable synthetic training data. Our contributions are threefold:

- We introduce a *label-free dataset curation pipeline* that filters synthetic reasoning traces by uncertainty, enabling efficient reasoning data construction in unlabeled domains.
- Applied to *biological perturbation prediction* and evaluated on the established PerturbQA benchmark, we show that filtering traces by internal uncertainty yields subsets with higher accuracy on final predictions. Moreover, training on uncertainty-filtered data outperforms unfiltered synthetic data and narrows the gap to ground-truth training, reducing reliance on costly wet-lab experiments.
- Through ablations, we find that per-class filtering corrects for class-specific uncertainty scales and that using an existing hybrid uncertainty score [21] leads to highest quality synthetic data, suggesting *general principles for data-efficient LLM reasoning*.

2 Background

Synthetic Reasoning Datasets. Chain-of-thought (CoT) prompting [24] elicits intermediate reasoning steps, and reliability can be improved via sampling and consistency filtering [30, 22]. LLM prompting has become a general tool for generating synthetic datasets [27, 7, 23]. Recent work combines these directions to produce synthetic reasoning traces [9, 11, 8, 29, 28, 18, 17]. However, most methods still require ground-truth labels [9, 11], sometimes falling back on heuristics like self-consistency when labels are absent [29, 28], and are therefore not fully label-free.

Biological Perturbation Prediction. Predicting how genetic or chemical perturbations alter cellular states is a central challenge in computational biology. Many aspects of perturbation prediction remain unknown [20] , making it a canonical testbed for label-scarce domains. Classical approaches (e.g., GEARS [15], scGPT [5]) often underperform simple baselines [2, 10]. Recent LLM-based models—GenePT [4], SUMMER [25], and SynthPert [14]—adapt embeddings, retrieval, or synthetic reasoning. SynthPert shows that CoT traces can aid generalization, but the method relies on labeled outcomes.

LLM Uncertainty Quantification. LLM uncertainty measures [16, 6, 19] include predictive entropy and probability margins [26, 12], or consistency-based agreement across multiple generations [1, 3]. Hybrid methods such as CoCoA show that combining both yields the strongest correlation with correctness [21], which we use as a substitute for lables.

3 Methods

Problem formulation. We study perturbation prediction: given a tuple (c, g, p) of cell type c, perturbation p, and gene g, the task is to predict whether g's expression increases (`up`), decreases (`down`), or remains unchanged (`non-regulated`). This three-class formulation follows [25] and reflects realistic biological workflows. As prior knowledge of gene responses is often unavailable, we generated synthetic reasoning data for this task, as illustrated in Figure 1.

Synthetic reasoning generation. To address the scarcity of labeled perturbation outcomes, we generate synthetic chain-of-thought (CoT) traces using a frontier LLM. For each input tuple (c, g, p), we produce $k + 1$ reasoning paths: one greedy-decoded trace r_0 and k high-temperature sampled traces $\{r_i\}_{i=1}^{k}$, keeping r_0 as the response. Each trace consists of a natural-language explanation paired with a final prediction in $\{\texttt{up}, \texttt{down}, \texttt{non-regulated}\}$. See appendix H for a runtime analysis.

Uncertainty filtering. We estimate the reliability of synthetic traces using the *CoCoA* metric [21], which combines semantic consistency across sampled traces $\{r_i\}_{i=0}^{k}$ with predictive perplexity of r_0. We compute CoCoA on the combined reasoning and final answer. For each class, we retain the top-$x\%$ traces with the lowest CoCoA, ensuring balanced coverage across outcomes. Details are in appendix A.

Supervised fine-tuning. Finally, we perform supervised fine-tuning (SFT) of a base LLM on the filtered dataset. The model is trained to reproduce both the reasoning trace and the final prediction,

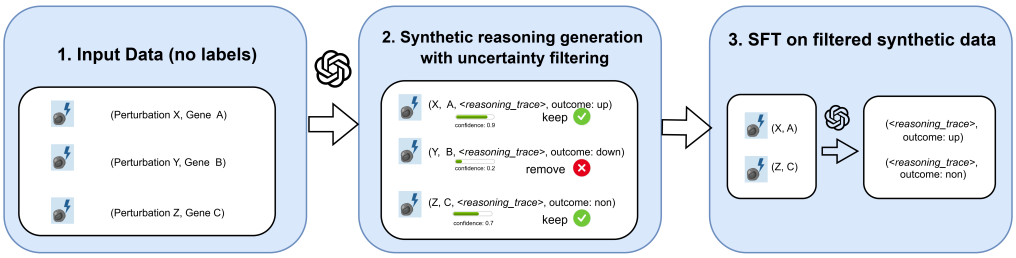

Figure 1: **Uncertainty-filtered synthetic reasoning pipeline.** Step 1: Generate multiple synthetic reasoning traces with predicted outcomes from unlabeled perturbation–gene pairs. Step 2: Score each trace for internal uncertainty (self-consistency and perplexity) and retain only low-uncertainty traces. Step 3: Use the retained traces as a label-free dataset for supervised fine-tuning (SFT), improving reasoning models without ground-truth labels.

distilling patterns from the synthetic pool. We choose SFT over reinforcement learning objectives, since in this setting the base model is not strong enough for a cold-start RL setup. This procedure does not require ground-truth experimental labels, enabling adaptation to perturbation prediction under extreme label scarcity.

4 Experiments

Dataset. We evaluate on the established *PerturbQA* benchmark [25], which reformulates perturbation prediction as natural language tuples (cell type, perturbation, gene) with labels `up`, `down`, `non-regulated`. Class imbalance (`non-regulated` dominates) motivates per-class filtering. We generate 48k synthetic traces and retain the top $x = 10$ percent per class under CoCoA [21], yielding a training set of 4.8k samples. We use the official train/test split provided with the dataset.

Baselines. We benchmark our label-free approach against both zero-shot and supervised baselines: (i) *Zero-shot*: out-of-the-box performance of teacher and student models. (ii) *Ground Truth + Synthetic Data SFT [14]*: Augmentation with label-conditioned synthetic traces, followed by filtering to retain only those with correct predictions. This represents the best-performing label-dependent strategy. (iii) *Random-sampling (10%)*: a size-matched control that selects traces uniformly at random, isolating the effect of uncertainty filtering. (iv) *Unfiltered (100%)*: training on the entire synthetic pool, testing whether more data alone suffices.

We use Gemini 2.5 Pro for synthetic data generation due to strong reasoning performance and access to token-level log-probabilities, with results for Qwen3-32B model in Appendix D, and we train Qwen3-32B. Implementation details are reported in Appendix B. We report means with stratified bootstrapped standard errors (5,000 resamples) for each metric.

4.1 Main Results

Lower-uncertainty subsets contain more predictive traces. Table 1 shows that uncertainty filtering provides clear signal: quality improves monotonically as we retain progressively lower-uncertainty data. Accuracy rises from 0.42 (full 48k) to 0.49 (top 1%, 480 examples), with consistent gains at every threshold. Per-class F1 for minority classes improves substantially—from 0.12/0.14 (Up/Down) to 0.30/0.21—demonstrating that uncertainty identifies not just correct predictions but more balanced, higher-quality reasoning. Appendix Figures 2–3 further visualize this trend across deciles. As a qualitative check, expert annotation of sample traces confirmed this pattern: low-uncertainty examples contained sound biological reasoning, while high-uncertainty ones exhibited errors undermining their conclusions (Appendix E).

Finetuning on uncertainty filtered traces leads to better performance. Table 2 shows that zero-shot Qwen3-32B achieves only 0.40 accuracy. Ground-truth SFT remains strongest at 0.62 accuracy. Among label-free methods, unfiltered training (100%) reaches 0.52, and random 10% sampling falls to 0.48. Crucially, our *uncertainty-filtered* 10% subset achieves 0.57 accuracy and F1 scores of 0.26 (Up) and 0.28 (Down), substantially improving over both random and unfiltered subsets, and

Data	Up			Down			Non-reg.			Acc	# samples
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1		
All data	0.07 ± 0.001	0.31 ± 0.005	0.12 ± 0.004	0.08 ± 0.003	0.64 ± 0.009	0.14 ± 0.003	$\mathbf{0.93} \pm 0.001$	0.39 ± 0.001	0.55 ± 0.001	0.42 ± 0.009	48k
Top 20%	0.11 ± 0.006	0.34 ± 0.009	0.16 ± 0.011	0.10 ± 0.010	0.65 ± 0.012	0.17 ± 0.013	0.91 ± 0.010	0.55 ± 0.009	0.32 ± 0.009	0.44 ± 0.019	9.6k
Top 10%	0.12 ± 0.004	0.36 ± 0.006	0.18 ± 0.007	0.11 ± 0.008	0.66 ± 0.010	0.19 ± 0.009	0.88 ± 0.015	0.39 ± 0.007	0.54 ± 0.018	0.45 ± 0.013	4.8k
Top 5%	0.14 ± 0.002	0.35 ± 0.003	0.20 ± 0.005	$\mathbf{0.12} \pm 0.012$	0.68 ± 0.014	0.20 ± 0.016	0.87 ± 0.021	0.39 ± 0.013	0.54 ± 0.024	0.46 ± 0.015	2.4k
Top 1%	$\mathbf{0.23} \pm 0.015$	$\mathbf{0.43} \pm 0.020$	$\mathbf{0.30} \pm 0.018$	$\mathbf{0.12} \pm 0.017$	$\mathbf{0.69} \pm 0.028$	$\mathbf{0.21} \pm 0.020$	0.88 ± 0.012	$\mathbf{0.41} \pm 0.018$	$\mathbf{0.56} \pm 0.014$	$\mathbf{0.49} \pm 0.023$	480

Table 1: **Lower-uncertainty subsets contain more predictive traces.** on (Prec), recall (Rec), and F1 score of synthetic data generated zero-shot by Gemini 2.5 Pro via the approach described in Figure 1. The rows correspond to the top $x\%$ of data retained after filtering by CoCoA metric. Lower uncertainty subsets achieve higher Acc and F1, indicating improved quality.

Method	Up			Down			Non-reg.			Acc
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	
Zero-shot baseline										
Zero-shot Gemini 2.5 Pro	0.20 ± 0.02	0.19 ± 0.01	0.20 ± 0.01	0.18 ± 0.01	0.22 ± 0.01	0.20 ± 0.01	0.79 ± 0.02	0.58 ± 0.02	0.67 ± 0.02	0.50 ± 0.02
Zero-shot Qwen3-32B	0.11 ± 0.01	0.27 ± 0.02	0.16 ± 0.01	0.18 ± 0.01	0.10 ± 0.01	0.13 ± 0.01	0.78 ± 0.02	0.45 ± 0.02	0.57 ± 0.02	0.40 ± 0.02
Label-based Training (Upper bound)										
Ground truth + Synth data	0.28 ± 0.02	0.51 ± 0.03	0.36 ± 0.02	0.16 ± 0.01	0.77 ± 0.02	0.27 ± 0.02	0.97 ± 0.01	0.61 ± 0.02	0.75 ± 0.02	0.62 ± 0.01
Label-free Training										
100%-Unfiltered	0.12 ± 0.04	0.31 ± 0.05	0.17 ± 0.04	0.17 ± 0.02	0.21 ± 0.03	0.19 ± 0.02	0.88 ± 0.02	0.59 ± 0.02	0.71 ± 0.02	0.52 ± 0.01
10%-Random-sampling	0.11 ± 0.02	0.28 ± 0.03	0.16 ± 0.03	0.17 ± 0.02	0.19 ± 0.05	0.18 ± 0.03	0.79 ± 0.02	0.49 ± 0.03	0.60 ± 0.03	0.48 ± 0.03
10%-Uncertainty-filtered (Ours)	$\mathbf{0.22 \pm 0.03}$	$\mathbf{0.32} \pm 0.04$	$\mathbf{0.26} \pm 0.03$	$\mathbf{0.19} \pm 0.02$	$\mathbf{0.55} \pm 0.05$	$\mathbf{0.28} \pm 0.02$	$\mathbf{0.91} \pm 0.02$	$\mathbf{0.60} \pm 0.01$	$\mathbf{0.72} \pm 0.02$	$\mathbf{0.57} \pm 0.02$

Table 2: **Finetuning on uncertainty filtered traces leads to better performance.** Precision (Prec), recall (Rec), F1 per class (Up, Down, Non-regulated), and overall accuracy (Acc) are shown for zero-shot baselines (teacher: Gemini 2.5 Pro; student: Qwen3-32B), fully supervised methods (with/without synthetic reasoning traces), and label-free training on uncertainty-filtered, random, or full synthetic datasets. Best scores for label-free training are bolded.

surpassing the strong LRM Gemini 2.5 Pro. This demonstrates that how data is selected matters: uncertainty filtering enables strong performance with only 10% of synthetic traces, outperforming both random sampling at the same scale and unfiltered training on 10× more data.

4.2 Ablation Studies

Per-class filtering outperforms global selection. Table 4 shows that random sampling keeps performance near unfiltered data (Up F1 0.10 vs 0.12). Global filtering improves recall for Down (0.71 vs 0.64), but collapses Non-regulated F1 to 0.12, hurting overall accuracy (0.16). In contrast, per-class filtering maintains balance across classes (Up F1 0.18, Down F1 0.19, Non-reg. F1 0.31) and yields the highest accuracy (0.25), suggesting the importance of class-aware selection.

Hybrid uncertainty metrics outperform single signals. Table 5 shows that perplexity alone performs worst (Up F1 0.14, Down F1 0.15). Consistency improves quality (Up F1 0.16, Down F1 0.20, Acc 0.24). The hybrid CoCoA score achieves the strongest balance (Up F1 0.18, Down F1 0.19, Non-reg. F1 0.31, Acc 0.25), suggesting that combining perplexity and self-consistency produces a more reliable label-free filter than either approach individually.

5 Discussion

We demonstrate that model-internal uncertainty enables label-free filtering of synthetic reasoning traces, addressing an underexplored axis of efficiency: supervision efficiency. While prior work focuses on algorithmic or system-level gains, experimental sciences and many other domains are constrained by the cost and availability of labels. On biological perturbation prediction—where each label requires costly wet-lab experiments—filtering by self-consistency and perplexity yields training data with higher accuracy than random or unfiltered baselines. Fine-tuning on this filtered data narrows the gap to fully supervised training without ground-truth labels.

Our results suggest that LLMs can self-curate training data, decoupling reasoning improvements from costly supervision. Although validated on biology — a domain with both label scarcity and epistemic uncertainty about underlying mechanisms — our method is domain-agnostic and applicable to any setting where experimental data or supervision is costly.

Limitations and future directions. Uncertainty filtering adds computational overhead for trace generation and scoring, though this cost is parallelizable and amortized across datasets (see Appendix H). Like any self-supervised signal, uncertainty filtering relies on the model's learned representations and may be sensitive to distribution shift. However, combining filtered synthetic data with even small amounts of ground-truth labels could provide validation and improve robustness. Future work should explore complementary uncertainty signals, validate generalization across diverse domains and shifts, and investigate such semi-supervised setups to improve robustness while reducing reliance on costly ground-truth labels.

References

[1] Y. Abbasi Yadkori, I. Kuzborskij, A. György, and C. Szepesvari. To believe or not to believe your LLM: Iterative prompting for estimating epistemic uncertainty. *Advances in Neural Information Processing Systems*, 37:58077–58117, 2024.

[2] C. Ahlmann-Eltze, W. Huber, and S. Anders. Deep learning-based predictions of gene perturbation effects do not yet outperform simple linear baselines. *BioRxiv*, pages 2024–09, 2024.

[3] J. Chen and J. Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. *arXiv preprint arXiv:2308.16175*, 2023.

[4] Y. Chen and J. Zou. GenePT: a simple but effective foundation model for genes and cells built from chatgpt. *bioRxiv*, pages 2023–10, 2024.

[5] H. Cui, C. Wang, H. Maan, K. Pang, F. Luo, N. Duan, and B. Wang. scGPT: toward building a foundation model for single-cell multi-omics using generative ai. *Nature methods*, 21(8):1470–1480, 2024.

[6] J. Geng, F. Cai, Y. Wang, H. Koeppl, P. Nakov, and I. Gurevych. A survey of confidence estimation and calibration in large language models. *arXiv preprint arXiv:2311.08298*, 2023.

[7] M. Goyal and Q. H. Mahmoud. A systematic review of synthetic data generation techniques using generative ai. *Electronics*, 13(17):3509, 2024.

[8] E. Guha, R. Marten, S. Keh, N. Raoof, G. Smyrnis, H. Bansal, M. Nezhurina, J. Mercat, T. Vu, Z. Sprague, et al. OpenThoughts: Data recipes for reasoning models. *arXiv preprint arXiv:2506.04178*, 2025.

[9] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi, et al. Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[10] E. Kernfeld, Y. Yang, J. S. Weinstock, A. Battle, and P. Cahan. A systematic comparison of computational methods for expression forecasting. *BioRxiv*, pages 2023–07, 2023.

[11] L. H. Li, J. Hessel, Y. Yu, X. Ren, K.-W. Chang, and Y. Choi. Symbolic chain-of-thought distillation: Small models can also" think" step-by-step. *arXiv preprint arXiv:2306.14050*, 2023.

[12] C. Ling, X. Zhao, W. Cheng, Y. Liu, Y. Sun, X. Zhang, M. Oishi, T. Osaki, K. Matsuda, J. Ji, et al. Uncertainty decomposition and quantification for in-context learning of large language models. *CoRR*, 2024.

[13] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[14] L. Phillips, M. Boubnovski Martell, J. Stoisser, C. Prada-Medina, R. Donovan-Maiye, and K. Märtens. SynthPert: Enhancing biological reasoning in LLMs via synthetic reasoning traces for cellular perturbation prediction. *arXiv preprint arXiv:2509.25346v1*, 2025.

[15] Y. Roohani, K. Huang, and J. Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 42(6):927–935, 2024.

[16] O. Shorinwa, Z. Mei, J. Lidard, A. Z. Ren, and A. Majumdar. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*, 2025.

[17] J. L. Stoisser, M. B. Martell, and J. Fauqueur. Sparks of tabular reasoning via text2sql reinforcement learning. *arXiv preprint arXiv:2505.00016*, 2025.

[18] J. L. Stoisser, M. B. Martell, L. Phillips, C. Hansen, and J. Fauqueur. STRuCT-LLM: Unifying tabular and graph reasoning with reinforcement learning for semantic parsing. *arXiv preprint arXiv:2506.21575*, 2025.

[19] J. L. Stoisser, M. B. Martell, L. Phillips, G. Mazzoni, L. M. Harder, P. Torr, J. Ferkinghoff-Borg, K. Martens, and J. Fauqueur. Towards agents that know when they don't know: Uncertainty as a control signal for structured reasoning. *arXiv preprint arXiv:2509.02401*, 2025.

[20] A. Tejada-Lapuerta, P. Bertin, S. Bauer, H. Aliee, Y. Bengio, and F. J. Theis. Causal machine learning for single-cell genomics. *Nature Genetics*, pages 1–12, 2025.

[21] R. Vashurin, M. Goloburda, A. Ilina, A. Rubashevskii, P. Nakov, A. Shelmanov, and M. Panov. Uncertainty quantification for LLMs through minimum bayes risk: Bridging confidence and consistency. *arXiv preprint arXiv:2502.04964*, 2025.

[22] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[23] Y. Wang, Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*, 2022.

[24] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[25] M. Wu, R. Littman, J. Levine, L. Qiu, T. Biancalani, D. Richmond, and J.-C. Huetter. Contextualizing biological perturbation experiments through language. *arXiv preprint arXiv:2502.21290*, 2025.

[26] Y. Xiao and W. Y. Wang. On hallucination and predictive uncertainty in conditional language generation. *arXiv preprint arXiv:2103.15025*, 2021.

[27] H. Xin, D. Guo, Z. Shao, Z. Ren, Q. Zhu, B. Liu, C. Ruan, W. Li, and X. Liang. Deepseek-prover: Advancing theorem proving in LLMs through large-scale synthetic data. *arXiv preprint arXiv:2405.14333*, 2024.

[28] P. Yu, J. Lanchantin, T. Wang, W. Yuan, O. Golovneva, I. Kulikov, S. Sukhbaatar, J. Weston, and J. Xu. CoT-Self-Instruct: Building high-quality synthetic prompts for reasoning and non-reasoning tasks. *arXiv preprint arXiv:2507.23751*, 2025.

[29] E. Zelikman, Y. Wu, and N. D. Goodman. Star: Self-taught reasoner. In *arXiv preprint arXiv:2203.14465*, volume 22, 2022.

[30] Z. Zhang, A. Zhang, M. Li, and A. Smola. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*, 2022.

Synthetic Data	Up			Down			Non-reg.			Acc	# samples
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1		
1% lowest uncertainty	**0.12** $\pm$ 0.025	**0.33** $\pm$ 0.041	**0.18** $\pm$ 0.033	0.08 $\pm$ 0.021	**0.50** $\pm$ 0.052	0.14 $\pm$ 0.027	**0.95** $\pm$ 0.018	**0.54** $\pm$ 0.035	**0.69** $\pm$ 0.030	**0.53** $\pm$ 0.038	40
5% lowest uncertainty	**0.12** $\pm$ 0.015	0.27 $\pm$ 0.025	0.17 $\pm$ 0.020	0.09 $\pm$ 0.014	0.43 $\pm$ 0.032	0.14 $\pm$ 0.018	0.90 $\pm$ 0.010	0.54 $\pm$ 0.020	0.68 $\pm$ 0.017	0.51 $\pm$ 0.021	121
10% lowest uncertainty	0.09 $\pm$ 0.012	0.23 $\pm$ 0.021	0.13 $\pm$ 0.016	0.07 $\pm$ 0.011	0.38 $\pm$ 0.025	0.11 $\pm$ 0.013	0.90 $\pm$ 0.009	0.53 $\pm$ 0.018	0.67 $\pm$ 0.015	0.50 $\pm$ 0.019	160
20% lowest uncertainty	0.10 $\pm$ 0.010	0.29 $\pm$ 0.018	0.15 $\pm$ 0.013	0.07 $\pm$ 0.009	0.29 $\pm$ 0.021	0.11 $\pm$ 0.011	0.89 $\pm$ 0.008	0.53 $\pm$ 0.015	0.66 $\pm$ 0.013	0.50 $\pm$ 0.016	200
All data (100%)	0.10 $\pm$ 0.003	0.24 $\pm$ 0.018	0.14 $\pm$ 0.014	**0.10** $\pm$ 0.011	0.34 $\pm$ 0.032	0.15 $\pm$ 0.011	0.86 $\pm$ 0.004	0.53 $\pm$ 0.007	0.65 $\pm$ 0.007	0.49 $\pm$ 0.008	4000

Table 3: **Synthetic reasoning dataset quality using Qwen3-32B model.** Each row shows the precision (Prec), recall (Rec), and F1 score per class, as well as overall accuracy (Acc). Uncertainty filtering via CoCoA metric.

A Uncertainty Definition

The *CoCoA* score [21] combines semantic consistency and perplexity. Let r_0 be the greedy trace and $\{r_i\}_{i=1}^k$ the sampled traces. We define uncertainty as:

$$\text{CoCoA}(x) = \frac{2}{k} \sum_{i=1}^{k} \big(1 - \text{sim}(r_0, r_i)\big) \cdot U_{\text{PPL}}(r_0), \tag{1}$$

where $\text{sim}(r_0, r_i)$ is the semantic similarity between r_0 and r_i (computed via a cross-encoder [13] as in [21]), and $U_{\text{PPL}}(r_0)$ is the perplexity of r_0. Higher CoCoA indicates higher uncertainty. For each class, we retain the top-$x\%$ traces with the lowest CoCoA, ensuring balanced coverage across outcomes.

Our aim is not to design a new uncertainty estimator — indeed, both perplexity and self-consistency are well-studied. Instead, our contribution is to demonstrate that when combined and repurposed, these familiar measures provide a practical, scalable criterion for filtering synthetic chain-of-thought traces, yielding label-free datasets that are significantly more effective for downstream fine-tuning.

B Implementation details

Synthetic traces are sampled from Gemini 2.5 Pro. For each input we draw $k = 8$ high-temperature (temperature=1, top-p=1.0, top-k=50) completions plus one greedy trace. The student is Qwen3-32B, fine-tuned with QLoRA ($1e^{-5}$ learning rate, batch size 4, 20 epochs) on one A100 GPU for 4 hours. We report per-class precision, recall, F1, and overall accuracy on the held-out test set, bootstrapped over 5,000 resampling iterations to compute standard errors and 95% confidence intervals for each metric. Prompt are reported in G.

We set $k = 8$ samples per query because this provides a tractable balance between (i) sufficient diversity for self-consistency estimation and (ii) manageable computational cost. Larger k increases stability but with diminishing returns." We retain the top 10% per class based on CoCoA because this threshold reliably yields a strong quality–quantity trade-off (see Table 1): retaining fewer examples leads to under-coverage of minority classes, while higher thresholds reintroduce noisy traces. In preliminary training sweeps (5–20%), 10% consistently provided the best downstream accuracy.

We plan to open source our code and dataset upon publication.

C Visualising Uncertainty as a Dataset Quality Signal

We present visualizations of the fact that that groups with lower uncertainty consistently achieve higher F1 scores. Figures 2 3 show that F1 scores monotonically decrease across CoCoA uncertainty deciles, with the lowest-uncertainty bin yielding the cleanest traces.

D Results for Qwen3-32B model

To illustrate the trends in uncertainty and data quality also for an open source model, we present a synthetic dataset generated by the Qwen3-32B model. The details are summarized in Table 3.

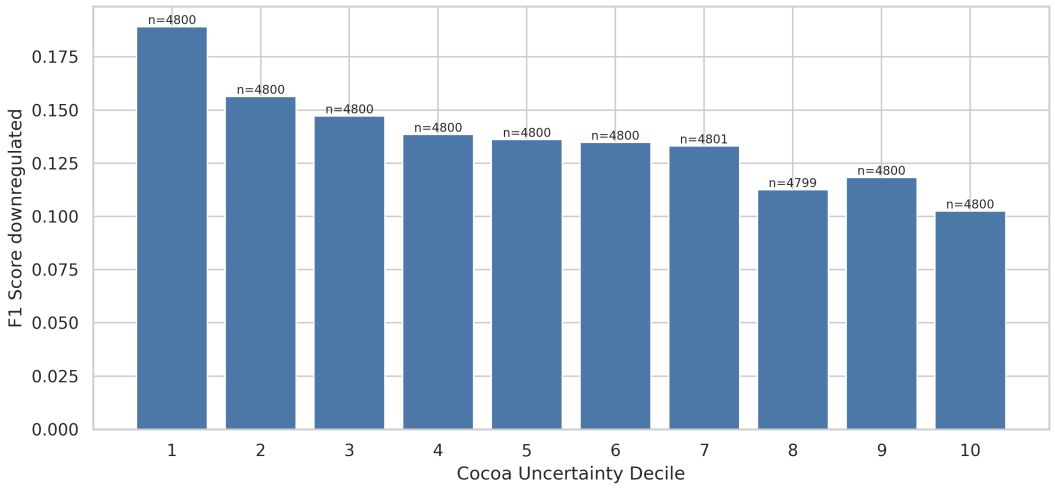

Figure 2: F1 score of upregulated genes stratified by CoCoA uncertainty deciles. Lower-uncertainty subsets yield consistently higher F1, with a clear monotonic trend across deciles. This confirms that uncertainty is strongly predictive of reasoning quality.

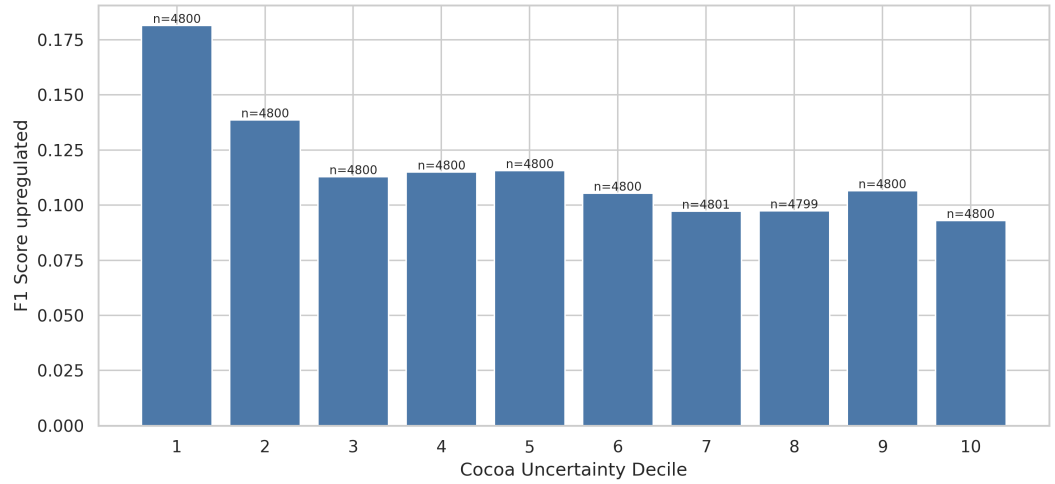

Figure 3: F1 score of downregulated genes stratified by CoCoA uncertainty deciles. Lower-uncertainty subsets yield consistently higher F1, with a clear monotonic trend across deciles. This confirms that uncertainty is strongly predictive of reasoning quality.

E Illustrative Expert Annotation of Reasoning Traces

We asked a PhD-trained biologist to annotate reasoning traces. Below is one drawn from the low-CoCoA (low uncertainty) subset and one from the high-CoCoA (high uncertainty) subset. The low-uncertainty trace was found to be correct throughout, whereas the high-uncertainty trace contained an early factual error that propagated and rendered the overall conclusion incorrect.

Low-Uncertainty Example (All Steps Correct)

Prompt

Analyze the regulatory effect of knocking down the `ALG2` gene on the `PDIA6` gene in a single-cell K562 cell line using CRISPR interference.

Dataset	Up			Down			Non-reg.			Acc
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	
All data	0.07 ± 0.001	0.31 ± 0.005	0.12 ± 0.004	0.08 ± 0.003	0.64 ± 0.009	0.14 ± 0.003	**0.93** ± 0.001	**0.19** ± 0.002	**0.31** ± 0.002	0.22 ± 0.003
10% random sampling	0.06 ± 0.004	0.26 ± 0.007	0.10 ± 0.006	0.08 ± 0.005	0.63 ± 0.012	0.14 ± 0.004	**0.93** ± 0.003	**0.19** ± 0.005	**0.31** ± 0.004	0.22 ± 0.006
Keep lowest 10% per class	**0.12** ± 0.005	0.36 ± 0.008	**0.18** ± 0.007	**0.11** ± 0.006	0.66 ± 0.014	**0.19** ± 0.007	0.88 ± 0.004	**0.19** ± 0.006	**0.31** ± 0.005	**0.25** ± 0.006
Keep lowest 10% global	**0.12** ± 0.006	**0.38** ± 0.010	**0.18** ± 0.008	**0.11** ± 0.007	**0.71** ± 0.016	**0.19** ± 0.009	0.88 ± 0.005	0.06 ± 0.007	0.12 ± 0.008	0.16 ± 0.009

Table 4: **Synthetic data quality under different filtering strategies.** Random sampling selects traces uniformly, global uncertainty filtering selects the lowest 10% CoCoA overall, and per-class filtering selects the lowest 10% CoCoA within each class. Metrics include per-class precision (Prec), recall (Rec), F1, and overall coverage (Acc).

Expert annotation (selected points):

- ALG2 functions in N-linked glycosylation (TRUE).
- PDIA6 is an ER chaperone induced by UPR (TRUE).
- ALG2 knockdown impairs glycosylation, induces ER stress, and activates UPR (TRUE).
- UPR upregulates PDIA6 via XBP1s/ATF6 (TRUE).
- Context: K562 cells are sensitive to ER stress (TRUE).

Conclusion: Expert judged the chain of reasoning correct, predicting `PDIA6` upregulation.

High-Uncertainty Example (Early Factual Error)

Prompt

Analyze the regulatory effect of knocking down the `CD3EAP` gene on the `RPTOR` gene in a single-cell K562 cell line using CRISPR interference.

Expert annotation (selected points):

- CD3EAP is incorrectly identified as calpastatin (FALSE).
- Downstream reasoning (calpain hyperactivation $\rightarrow$ mTORC1 suppression) is partly biologically plausible, but depends on the incorrect gene assignment.
- Final conclusion (RPTOR downregulation) judged not supported.

Conclusion: A single early factual error misdirects the reasoning chain, making the overall conclusion unreliable despite plausible intermediate statements.

F Ablation Details

Table 4 evaluates different filtering strategies for the synthetic dataset. Random sampling selects traces uniformly, global uncertainty filtering selects the lowest 10% CoCoA across all examples, and per-class filtering selects the lowest 10% within each class. The results show that per-class filtering consistently yields higher-quality traces, as uncertainties are not directly comparable across classes; without per-class selection, minority classes tend to be underrepresented.

Table 5 compares different uncertainty metrics for selecting synthetic traces: perplexity alone, consistency across multiple generations, and the CoCoA score, which combines both signals. The results indicate that using CoCoA produces the most reliable traces across all classes, demonstrating that combining perplexity and self-consistency is superior to either metric alone for identifying high-quality reasoning data.

Dataset	Up			Down			Non-reg.			Acc
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	
All data	0.07 ± 0.001	0.31 ± 0.005	0.12 ± 0.004	0.08 ± 0.003	0.64 ± 0.009	0.14 ± 0.003	**0.93** ± 0.001	0.19 ± 0.002	**0.31** ± 0.002	0.22 ± 0.003
Keep lowest 10% CoCoA	**0.12** ± 0.005	**0.36** ± 0.008	**0.18** ± 0.007	0.11 ± 0.006	0.66 ± 0.014	0.19 ± 0.007	0.88 ± 0.004	0.19 ± 0.006	**0.31** ± 0.005	**0.25** ± 0.006
Keep lowest 10% consistency	0.11 ± 0.005	0.30 ± 0.008	0.16 ± 0.007	**0.12** ± 0.006	**0.69** ± 0.014	**0.20** ± 0.007	0.89 ± 0.004	**0.20** ± 0.006	**0.31** ± 0.005	0.24 ± 0.006
Keep lowest 10% perplexity	0.09 ± 0.004	0.34 ± 0.007	0.14 ± 0.006	0.08 ± 0.005	0.64 ± 0.012	0.15 ± 0.004	0.91 ± 0.003	0.18 ± 0.005	**0.31** ± 0.004	0.23 ± 0.006

Table 5: **Synthetic data quality when selecting traces based on different uncertainty metrics.** Perplexity measures fluency, consistency captures agreement across multiple traces, and CoCoA combines both signals. Metrics include per-class precision (Prec), recall (Rec), F1, and overall coverage (Acc).

Prompt

System message: You are an molecular and cellular biology expert analyzing gene regulation upon CRISPRi knockdown. First, provide your reasoning process within <think> </think> tags. Consider relevant pathways (e.g., cell-type specific biology, ribosome biogenesis, transcription, mitochondrial function, stress response), gene interactions, and cell-specific context. Then, choose one option from the following and place your choice within <answer> </answer> tags: 'upregulated', 'downregulated', or 'not differentially expressed'. Example: <think> [Your reasoning here] </think><answer> [upregulated / downregulated / not differentially expressed] </answer>
User message: Analyze the regulatory effect of knocking down the perturbation gene on the gene gene in a single-cell cell_type cell line using CRISPR interference.

Figure 4: **Prompt template used for data generation, SFT, and evaluation.**

G Prompt

Figure 4 shows the prompt used both for synthetic data generation, SFT training and evaluation.

H Computational Efficiency

We measured and verified the compute requirements of our generation and filtering pipeline, reporting both per-example and dataset-level costs. Table 6 summarizes the results for 50k examples.

Component	Operation	Cost per example	Total (50k)
Trace generation	9 traces (∼2k tokens)	∼0.0019 GPUh	95 GPUh
Self-consistency	8 cross-encoder passes	∼0.000088 GPUh	4.4 GPUh
Perplexity	log-probs of greedy trace	≈0 (free at generation)	≈0
Aggregation + ranking	CPU-side sort	negligible	negligible
Total (ours)	generation + filtering	∼0.0020 GPUh	99.4 GPUh

Table 6: **Measured compute costs for generating and filtering 50k synthetic examples.** GPUh denotes GPU-hours. Estimates are based on runs with Gemini 2.5 Pro (trace generation) and a medium cross-encoder (filtering).

Overall, generation dominates compute requirements at ∼95 GPUh, while filtering adds only ∼4–5% overhead (driven entirely by the cross-encoder self-consistency checks). Perplexity calculation incurs no additional cost when log-probs are retained during decoding. Aggregation and ranking are negligible.

By comparison, SynthPert's label-conditioned generation requires both synthetic trace generation and label checking. The latter depends on experimental outcomes and cannot be parallelized across GPUs, making the cost effectively dominated by human/biological supervision. In practice, one wet-lab label requires days of bench work; thus, even a $10\times$ compute overhead is negligible compared to the cost of experimental supervision.