ORCHESTRATING SYNTHETIC DATA WITH REASONING

Tim R. Davidson* EPFL
Benoit Seguin Google
Enrico Bacis Google
Cesar Ilharco Google Deepmind
Hamza Harkous† Google

ABSTRACT

Many AI applications of interest require specialized multi-modal models. Yet, relevant data for training these models is inherently scarce. Human annotation is prohibitively expensive, error-prone, and time-consuming. Meanwhile, existing synthetic data generation methods often rely on manual prompts, evolutionary algorithms, or extensive seed data from the target distribution — limiting scalability and control. In this paper, we introduce **Simula**: a novel, seedless framework that balances global and local reasoning to generate synthetic datasets. We utilize taxonomies to capture a global coverage space and use a series of agentic refinements to promote local diversity and complexity. Our approach allows users to define desired dataset characteristics through an explainable and controllable process, without relying on seed data. This unlocks new opportunities for developing and deploying AI in domains where data scarcity or privacy concerns are paramount.

1 INTRODUCTION

Data availability and access have been central to advances in artificial intelligence research. Recently, the abundance of highly diverse internet data enabled the development of increasingly capable generalist models (Gemini et al., 2023; OpenAI et al., 2023; Anthropic, 2024; Touvron et al., 2023). Despite these models' impressive versatility, widespread integration will require them to quickly specialize on novel, uncommon, and critical applications (e.g., medicine, finance, law). Unfortunately, specialized data in these areas is often scarce or inaccessible due to cost or privacy concerns. Creating such datasets manually is expensive, time-consuming, and error-prone (Chen et al., 2023; Gilardi et al., 2023). Synthetic data offers a promising, scalable alternative (Singh et al., 2024a; Abdin et al., 2024; Guo et al., 2025). Nevertheless, how to best balance its various desiderata is an open question.

To optimize generalist models for specific tasks, practitioners typically use techniques such as fine-tuning (Ziegler et al., 2019; Hu et al., 2022; Chung et al., 2024), distillation (Hinton et al., 2015), reinforcement learning (Christiano et al., 2017; Jaech et al., 2024; Guo et al., 2025), and few-shot prompting (Brown et al., 2020). Each of these approaches relies on the availability of relevant example data. Developing scalable methods that can reliably deliver specialized data on-demand is thus vital to accelerate broader AI adoption. Furthermore, synthetic data can increase control and source-attribution, enabling more targeted optimization (Ruis et al., 2024).

Yet, characterizing "good" synthetic data is intrinsically challenging. Generally, "good" is discussed in terms of *quality*, *diversity*, and *complexity* (Havrilla et al., 2024). However, the precise definitions of these terms is contentious. Instead of describing the usefulness of data (Swayamdipta et al., 2020; Marion et al., 2023a), "quality" commonly refers to how well data points fit specific requirements. For example, if the intention is to generate "*an image of a red cat*", does the resulting image have a cat in it, and, is that cat indeed red? Meanwhile "complexity" can refer to how confusing or elaborate a specific data point is (Ethayarajh et al., 2022; Shao et al., 2023), but is often equated with the relative concept of "difficulty". In the case of our red-cat image, a complex example might be a partially obscured cat, or one lying in the shadows. Finally, "diversity" offers both a *global* and *local* perspective: does the generated data globally cover the main factors of interest, and does it locally exhibit sufficient variety within specific factors?

Existing synthetic data generation methods generally optimize only a subset of the above desiderata (Havrilla et al., 2024). They often rely on elaborate custom prompts (Gupta et al., 2024; Xu et al.,

*Work done during an internship at Google.
†Correspondence to: harkous@google.com. Author contributions at the end of the paper

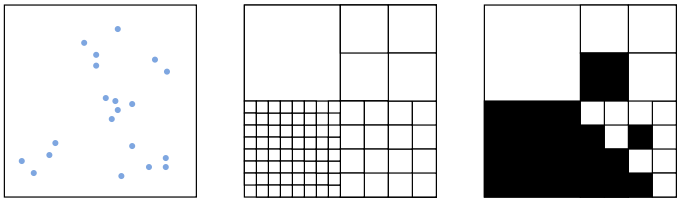

Figure 1: **Synthetic coverage examples.** (a) characterizes "random" sampling behavior; (b) perfect global planning at increasing granularity; (c) global planning with progressive coverage loss.

2024; Yu et al., 2023), or stochastic, evolutionary algorithms (Mehrotra et al., 2024; Fernando et al., 2024). The former generalizes poorly, while the latter lacks explainability and control. Many approaches further require a large number of "seed examples" drawn from the target dataset, which presents an unrealistic assumption in many real-world cases and might hurt global coverage.

In this work we propose Simula: a holistic approach to synthetic data generation that balances global and local reasoning. Given a target dataset description, Simula maps out a global coverage space using synthetic taxonomies. Then, it applies a series of agentic refinements to promote local diversity and complexity. Finally, it performs double-critic rejection sampling to optimize quality. Our approach is seedless and provides clear notions of explainability and control, essential for optimal data curation. We rigorously test the core reasoning assumptions underlying our approach and demonstrate its efficacy on a series of carefully designed experiments.

2 Orchestrating Synthetic Data with Reasoning

Imagine we are interested in creating a dataset with the description $y :=$ *"A dataset of stories about cats"*. Due to the under-specification of y, it is infeasible to exhaustively describe the space of all datasets $\mathcal{Y}$, that fit the description. This is problematic as it prevents us from developing an explainable notion of global coverage, i.e., given a dataset $\mathcal{D}_y \sim \mathcal{Y}$, what area of $\mathcal{Y}$ does it represent?

2.1 Approximating Global Coverage using Synthetic Taxonomies

To regain control, we formulate a first order approximation of $\mathcal{Y}$ by disentangling our target dataset into its prime factors of variation. [1] For example, a dataset that fits the description above might consist of data points considering, e.g., "*cat type*", "*story format*", and "*intended audience*". In Simula, a multi-modal model (M3) is prompted to propose factors based on a set of human-provided instructions, e.g., a description like y, and/or a sample $\mathcal{S}$ of existing data. These factors can be accepted or rejected by a human (or M3). Given factors f_i, an M3 is used to expand them breadth-first into taxonomies, $\mathcal{T}_i$, of a (user-) specified depth d_i:

$$g(\text{M3}, y, \mathcal{S}, (d_0, f_0), \cdots, (d_K, f_K)) = \{\mathcal{T}_i\}_{i=0}^{K} = \mathcal{T}^y \tag{1}$$

Using taxonomies provides granular explainability and control of $\mathcal{Y}$ compared to random sampling (Figure 1.a). Intuitively, as we increase the number of factors and taxonomy depths, we sharpen our coverage control (Figure 1.b). However, this granular control comes at a potential cost: with every taxonomy expansion we risk "missing" nodes of interest, resulting in the progressive coverage loss depicted in Figure 1.c.

To mitigate potential coverage loss resulting from missing nodes, we generate factor taxonomies by alternating between three steps: (1) Given a node, its ancestors and its siblings, an M3 is prompted N times to propose children nodes. This sampling strategy is inspired by the "Best-of-N" literature to increase the proposal distribution and cover edge cases. (2) In a separate call, an M3 is prompted to locally critique the generated nodes, e.g., on completeness, soundness, and specificity, taking advantage of M3s' observed generator-critic gap (Huang et al., 2024). Finally, (3) after generating all nodes of a specific level, an M3 is prompted to generate a "plan" for the next level. This last step

[1] Note that perfect disentanglement is of course not always possible (Locatello et al., 2019).

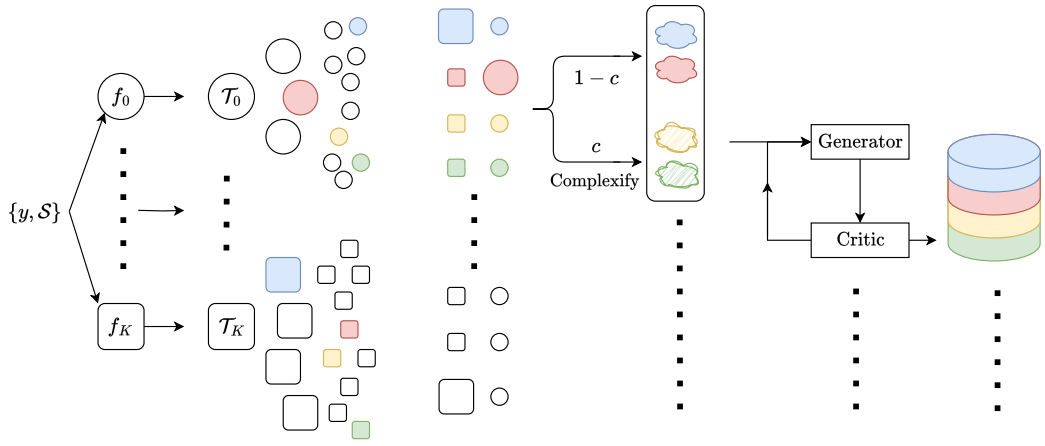

Figure 2: **Schematic of Simula Framework.** Given user instructions y and/or a data sample $\mathcal{S}$, we first determine factors of interest (a), which are expanded into taxonomies (b). Next, nodes of the taxonomies are sampled to obtain mixes (c), and turned into "meta prompts" (d). A user-defined fraction, c, of meta prompts is "complexified". Meta prompts are used to generate data proposals by the `Generator` model, and iteratively refined using a `Critic` model (e).

is necessary to enable consistent and fast parallel generation, e.g., by ensuring a similar degree of granularity at different node expansions on the same level. At each step, the M3 also has access to the user-provided input y, and/or a sample $\mathcal{S}$, from the target distribution. We will empirically show that this alternating generator-critic approach improves over simple 0-shot expansion.

2.2 Generating Controllable and Explainable Synthetic Data at Scale

To generate a synthetic dataset that fits our requirements, we distinguish between two phases: taxonomic sampling (Figure 2.c) and agentic refinement (Figure 2.d-e). Initially, an M3 formulates a plan composed of sampling strategies. A strategy defines which taxonomies can be combined together, and with which weights. This is important, as not all sub-taxonomies make sense to combine (e.g., writing a horror novel about a troubled cat for toddlers seems ill-advised). A practical application of strategies could involve aiming for an equal split between kid and adult audiences, where the M3 might propose two strategies, filtering inappropriate formats like "horror" from the kids' strategy. The generation pipeline then samples a strategy and nodes from the corresponding taxonomies $\mathcal{T}_j$. These sampled "requirements", along with the original dataset instructions y, guide an M3 to construct one or more "meta prompts". For example, M3$(y, \{$*house cat, poem, travel enthusiast*$\})$, becomes "*Compose an exciting haiku about a house cat who goes on an adventure*". Finally, these meta prompts direct an M3 to generate the data outputs.

Optimizing Local Diversity and Complexity. Imagine we want to construct a dataset of size $N = 100$, and our factor and strategy selection has yielded $T = 200$ unique node-pairs. Since $N < T$, our sampling budget allows for at most 100 unique node-pairs with a single meta prompt each, resulting in a global coverage rate of $100/200 = 0.5$. Conversely, for $N > T$, e.g, $N = 800$, we can sample up to four meta prompts for each requirement set. As the number of meta prompts per node-pair grows, we increase *local* diversity. However, as N/T grows, independently generating meta prompts from fixed requirements can lead to mode collapse, i.e., meta prompts that are increasingly similar. This is mitigated by generating multiple meta prompts simultaneously, prompting for maximum sample diversity, then sub-sampling the required fraction. We call this approach "*semantic expansion*". Next, we expand the *complexity* of a fraction of the samples, by prompting the M3 to increase the complexity of the generated meta-prompts and outputs while maintaining our semantic requirements. We refer to this later as "*complexity expansion*. Optimizing local diversity and complexity this way works well for smaller sample sizes, but degrades as N/T grows very large. Instead, for large N/T, Simula can be configured to iteratively prompt for more diverse or complex meta prompts with previous attempts in context. This allows an M3 to reflect on previous generations.

Enhancing Sample Quality with Critics. Next, the system performs a series of agentic refinement steps to optimize sample output quality. It starts with point-wise checks to ensure the generated samples pass the specified semantic and synthetic requirements. This involves prompting the M3 to "critique" the generated samples by providing the meta prompt used for generation and requesting an explanation and a binary verdict. For example, given the generated sample for the adventurous house-cat haiku above, the M3 checks if the cat in the story is indeed a house cat, if the output is a haiku, and if adventures were had. For tasks requiring outputs with a defined notion of correctness (e.g., classification or multiple-choice questions), the system employs an additional "double critic" step, which independently assesses correctness and incorrectness to mitigate sycophancy bias (Sharma et al., 2024). Following these "*critic refinement*" steps, if the M3 responds with a negative verdict, the system either rejects the sample or applies automated modifications based on the explanation, then repeats the critique step.

2.3 Understanding Data Coverage and Complexity

Taxonomic Data Coverage. The importance of data selection during both M3 training and inference time is emphasized by a growing body of research (Swayamdipta et al., 2020; Marion et al., 2023b; Ye et al., 2024; Xia et al., 2024; Hu et al., 2024; Hübotter et al., 2024), *inter alia*. Access to training data that accurately reflects test time conditions further is essential to assess model performance and preparedness. Despite their vital role, most datasets are sparsely labeled, e.g., "*math*", "*harmful*", "*customer complaint*", etc., or not labeled at all. This complicates efforts to curate optimal corpora, allocate resources (Qian et al., 2024), and catch potential misalignment between train and test sets (van Breugel et al., 2024). Using taxonomy mixtures offers a way forward not only for generating synthetic data, but also for better understanding existing data. Given a dataset, we can generate taxonomies and query an M3 to assign nodes of the taxonomies to each data point. This provides a fine-grained view into data composition and actionable insights to expand coverage. Additionally, this opens up the possibility of "taxonomic nearest-neighbor" (TNN) retrieval, alleviating the dependency on limited embedding-based alternatives (Ethayarajh, 2019; Kashyap et al., 2023).

Evaluating Sample Complexity. Depending on the task at hand, more or less complex data samples might be desired. This presents a challenge, as most real data is not annotated for complexity and synthetic data generation is unsupervised. To nevertheless partition data based on complexity, we propose a "batch-wise" comparative evaluation approach: Firstly, batches of data points are sampled such that each data point appears K times. Secondly, an M3 assigns scores to each sample in each batch reflecting their *complexity*, optionally using a dataset description y. Using batches to score individual points provides more context, reducing noise resulting from per-sample overconfidence or poor calibration (Zheng et al., 2023; Xiong et al., 2024). Finally, to further improve our relative scoring signal, we compute ELO scores (Elo & Sloan, 1978) from the score assignments. This method (1) enables complexity comparisons of data points across different datasets and (2) can be used to sample more complex data on demand. We further use this in our evaluation to assess the efficacy of Simula's complexity expansion component. Additional discussion can be found in Appendix D.

3 Experimental Setup

We conduct experiments using Gemini 1.5 Pro as our teacher model (Gemini et al., 2024) and both the pre-trained and instruction-tuned versions of Gemma 2B (Gemma et al., 2024) as our student models. We report confidence intervals as the standard error over three runs when applicable.

3.1 Verifying Core Method Assumptions

Our approach primarily relies on three core assumptions about M3 reasoning capabilities: M3s can (i) generate high-quality taxonomies; (ii) function as effective "critics" of their own outputs; and (iii) distinguish between more and less complex examples. We test each of these assumptions as follows:

M3s Generate High-quality Taxonomies. Evaluating the quality of taxonomies is inherently challenging due to the lack of standardized criteria and methods (Szopinski et al., 2020; Kaplan et al., 2022). We differentiate between grounded taxonomies, e.g., phylogenetic trees, and conceptual ones, e.g., types of harmfulness. Given an expert taxonomy, $\mathcal{T}_E$, we compare to an M3-generated taxonomy for the same topic, $\mathcal{T}_{\text{M3}}$. Structurally, we care about completeness (does $\mathcal{T}_{\text{M3}}$ cover $\mathcal{T}_E$?), soundness

(does $\mathcal{T}_{\text{M3}}$ contain irrelevant or unnecessary nodes?), and novelty (does $\mathcal{T}_{\text{M3}}$ contain relevant nodes *not* in $\mathcal{T}_E$?). We evaluate both the Simula generator-critic approach and 0-shot expansion on six real-world taxonomies. Additional experimental details are available in Appendix B.

M3s Are Effective Critics We take a dataset with ground-truth labels $\mathcal{D}_{\text{true}}$ and create a corrupt copy by prompting an M3 to subtly change the ground-truth labels. We independently prompt a critic model if (1) a label is *correct* and (2) if it is *incorrect*. Because generating synthetic data is unsupervised, we won't know if the critic is judging a correct or incorrect sample at inference time. Hence, for effective critic-based rejection sampling, we need $P(\text{correct}|x_i, \text{do}(y_i = \text{correct}))$ to be high and $P(\text{correct}|x_i, \text{do}(y_i = \text{incorrect}))$ low. We further test if a critic's efficacy in the controlled setting transfers to the empirical setting of model-generated labels. We compare the controlled critic to the empirical one on free-form math problems (MATH, Hendrycks et al. (2021)), and multiple-choice questions in selected languages (MMLU, Singh et al. (2024b)).

M3s Can Distinguish Complexity We investigate if model-assigned complexity (1) agrees with human annotators and (2) how well-calibrated it is to their generative and critic capabilities, We evaluate on the MATH dataset, as each question comes with a 1-5 annotated complexity rating, and selected MMLU subjects with multiple levels of education (elementary, high school, and college).

3.2 Intrinsic Metrics of Synthetic Data

We generate synthetic datasets for a subset of SuperGLUE (Wang et al., 2019), and multilingual subsets of MMLU (Singh et al., 2024b). We ablate four versions of Simula: (i) Taxonomies only (**TO**) with meta prompting, but no expansion, (ii) with semantic expansion only (**TS**), (iii) with complexity expansion only (**TC**), and (iv) full (**TSC**) with both expansion types and critic refinement. As a baseline (**B**), we sample data from the train/validation sets and iteratively prompt an M3 to expand.

Global Diversity. Following Yu et al. (2023), we evaluate average pairwise cosine similarity on the real and generated datasets using embeddings from Lee et al. (2024b).

Local Diversity. We first group data by taking the $k = 10$ closest points to each data point in embedding space to ensure semantically meaningful clusters. We then analyze the average pairwise cosine similarity across these clusters.

Complexity. We use our batch-wise evaluation approach described in Section 2.3, mixing synthetic with real data for valid comparisons. After instructing the M3 to assign scores between 0-100, we report the delta to the real data. For example, if real data has an average complexity of 50, and synthetic 55, we report +5.

3.3 Downstream Metrics of Synthetic Data

We LoRA fine-tune pre-trained student models (Hu et al., 2022) on the explanation (chain of thought generated by the teacher) and the final answer, varying the sample size for multilingual subsets of MMLU . Data is generated using the full version (**TSC**) and one without critic refinement (**TSC\c**), as well as the baseline (**B**). We also include comparisons to a 5-shot setup with pre-trained Gemma 2B and a 5-shot with instruction-tuned Gemma 2B. We report the macro average F1-score over the 4 choices as well as the "Performance Gap Recovery" (PGR) metric from Kim et al. (2024).

4 Results

4.1 Testing Core Assumptions

Quality of Taxonomies. Table 1 suggests Simula taxonomies approximately cover those created by human experts (γ). For conceptual taxonomies, almost all generated nodes are sound (σ), with many novel expansions (ν) resulting in increased total coverage (τ). For both taxonomy types, Simula clearly outperforms 0-shot generation. These results support our approach of approximating a global coverage space using taxonomies. Appendix B contains expanded results, analysis, and examples.

Critic and Complexity. Shown left in Table 2, the probability of our double critic recognizing a correct answer is slightly higher than generating it, $p(y) \geq \mu_{\text{gen}}$, while the probability of accepting

Table 1: **Taxonomy evaluation.** Average completeness (γ), soundness (σ), novelty (ν), and total coverage (τ).

Source	Grounded		Conceptual			
	γ	σ	γ	σ	ν	τ
Simula	**0.74**	**0.75**	**0.78**	**0.97**	**0.94**	**1.72**
0-shot	0.52	0.70	0.50	**0.97**	0.32	0.83

an incorrect answer, $p(y^{\text{corrupt}})$, is much lower than rejecting it. Thus, after a rejection-sampling step on the generated outputs, the mean accuracy is expected to increase:

$$\mu_{\text{gen}} \cdot p(y) + (1 - \mu_{\text{gen}}) \cdot p(y^{\text{corrupt}}) = \mathbb{E}[\mu_{\text{critic}}] > \mu_{\text{gen}} \tag{2}$$

Lastly, the percentage of rejected samples $|\mathcal{D}_{\text{reject}}|$ is expected to grow as the complexity increases.

Right of Table 2, we observe critic-rejection sampling on model outputs empirically increases the remaining accuracy of model outputs, $P(\mu_{\text{gen}}|\checkmark) > P(\mu_{\text{gen}}|\times)$. Further note that, stratified by Complexity, the average ELO score of rejected outputs is consistently higher than those of accepted ones. Taken together, these results provide strong evidence that our critic refinement improves overall sample quality. Additional critic and complexity results are available in Appendices C and D.

Table 2: **Critic-rejection sampling results for MATH test set stratified by complexity.** The controlled critic probabilities of accepting a correct answer $p(y)$, or an incorrect one $p(y^{\text{corrupt}})$. Applying rejection sampling gives the expected change in accuracy $\mathbb{E}[\mu_{\text{critic}}]$ over the generative performance μ_{gen} and the percentage of rejected samples. The right side of the table shows the empirical change in accuracy, μ_{gen} of following critic rejections. Also shown are the average ELO complexity score and the size of the rejected and accepted subsets.

	Controlled Rejection Sampling (SE < 0.01)					Empirical Rejection Sampling			
Complexity	$p(y)$	$p(y^{\text{corrupt}})$	μ_{gen}	$\mathbb{E}[\mu_{\text{critic}}]$	$\lvert\mathcal{D}_{\text{reject}}\rvert$	Critic	μ_{gen}	ELO	$\lvert\mathcal{D}\rvert$
Level 1	0.97	0.24	0.97	0.99	0.04	×	0.33 ± 0.21	373 ± 19	6
						✓	0.98 ± 0.01	328 ± 2	431
Level 2	0.95	0.24	0.95	0.99	0.06	×	0.56 ± 0.12	410 ± 12	18
						✓	0.96 ± 0.01	364 ± 2	875
Level 3	0.93	0.28	0.93	0.98	0.08	×	0.30 ± 0.09	435 ± 7	30
						✓	0.95 ± 0.01	389 ± 2	1101
Level 4	0.89	0.30	0.85	0.94	0.14	×	0.27 ± 0.05	453 ± 4	90
						✓	0.89 ± 0.01	413 ± 1	1124
Level 5	0.81	0.36	0.72	0.86	0.23	×	0.17 ± 0.03	470 ± 2	189
						✓	0.81 ± 0.01	436 ± 1	1134

4.2 Intrinsic Results

A subset of the intrinsic results is shown in Table 3. We note that the full system (**TSC**) consistently improves over the baseline (**B**) and the minimal system (**TO**). Compared to the real data (**R**), our system is capable of generating more complex data while boasting comparable to better diversity in most cases. Note that, as the diversity metrics are based on embeddings, they do not necessarily capture important control details. Additional results can be found in Appendix E.

4.3 Downstream Results

Figure 3 showcases the macro average F1-score across languages and subjects of pre-trained (PT) and instruction-tuned (IT) models. We observe the value of increasing the training data volume on performance, which is a unique advantage of synthetic data. We observe the utility of the critic step, evidenced by the gap between **TSC** and **TSC\c**. We further note the consistent performance gain compared to the baseline PT 5-shot COT variant and the synthetic baseline (**B**) across the various combinations. As the fine-tuning data size increases, the models fine-tuned on **TSC** get closer to or exceed the IT variant. In Appendix F, we show the results of the Performance Gap Recovery metric.

Table 3: **Intrinsic evaluation results.** Global and local diversity (Δ_{global}, Δ_{local}) and global complexity (ϕ). Diversity values have SE < 0.001. For complexity, we report the lift over real data (**R**) with SE < 1.

	boolq			rte			MMLU (English)			MMLU (Nepali)		
Method	$\Delta_{\text{global}} \uparrow$	$\Delta_{\text{local}} \uparrow$	$\phi \uparrow$	$\Delta_{\text{global}} \uparrow$	$\Delta_{\text{local}} \uparrow$	$\phi \uparrow$	$\Delta_{\text{global}} \uparrow$	$\Delta_{\text{local}} \uparrow$	$\phi \uparrow$	$\Delta_{\text{global}} \uparrow$	$\Delta_{\text{local}} \uparrow$	$\phi \uparrow$
R	0.30	0.62	20	0.51	0.66	28	0.36	0.64	40	0.64	0.85	43
B	0.40	0.92	+1	0.52	0.74	+8	0.39	0.82	+2	0.69	0.93	0
TO	0.36	0.71	+5	0.55	0.73	-2	0.36	0.70	+9	0.65	0.87	+8
TSC	0.37	0.66	+17	0.52	0.70	+5	0.36	0.68	+15	0.53	0.80	+14

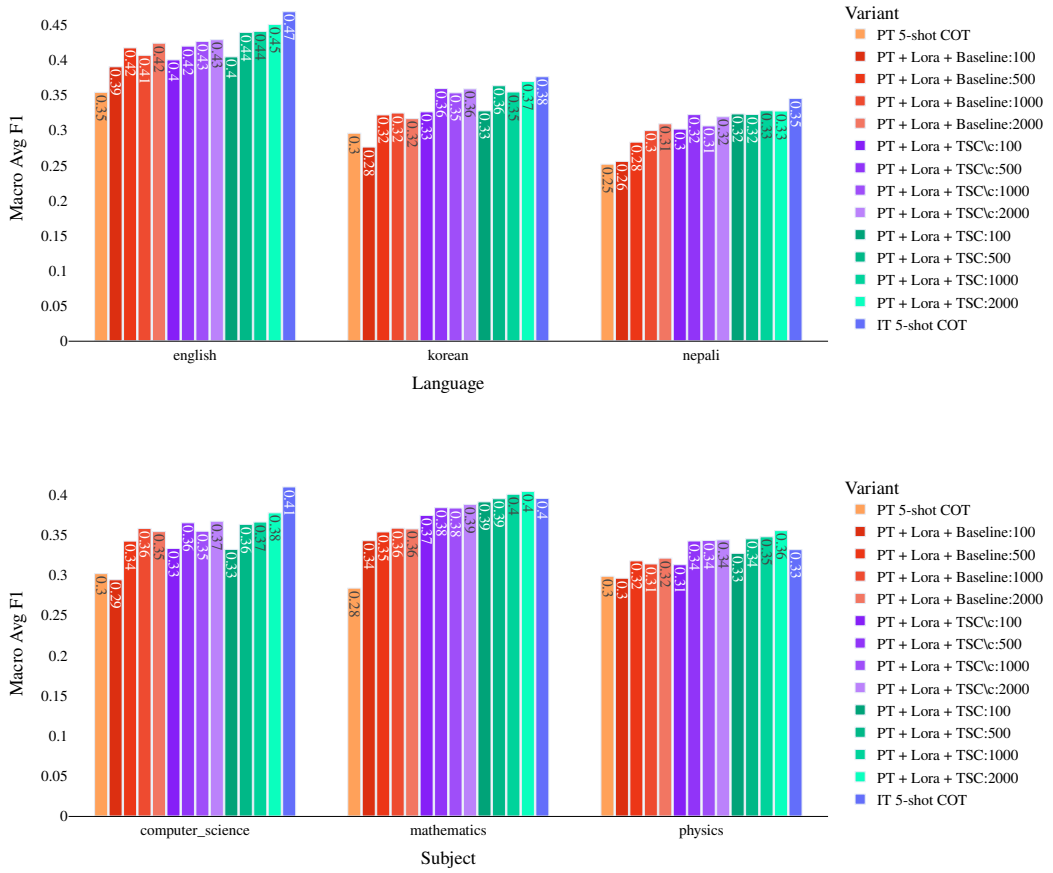

Figure 3: **Performance on MMLU Global.** Split by language (top) and subject (bottom), of synthetic data variants vs. PT & IT 5-shot with COT. The numbers in the legend indicate the data size used per variant.

5 Conclusion

AI is at a junction: just as its potential is becoming evident, the data needed to realize it is unlikely to be generated by humans (Villalobos et al., 2024). With Simula, we address this important need, by providing a scalable synthetic data process built around reasoning. We carefully validated its efficacy across diverse tasks and languages, demonstrating control and explainability. We thus believe our work represents a promising step toward more widespread and equitable access to advanced AI.

AUTHOR CONTRIBUTIONS

Tim R. Davidson: Led the majority of the writing of the manuscript, including system formalization, literature review, experimental setup, and evaluation; designed and evaluated the double critic-rejection sampling approach; developed and evaluated the calibrated complexity scoring metric; introduced data coverage metrics; developed several optimizations for the taxonomy generation algorithm (e.g., best-of-N, critiquing); conducted intrinsic evaluations; provided core insights across the other parts during his internship stay at Google.
Benoit Seguin: Led engineering for Simula; implemented core components for modular, scalable data generation; developed the core library for prompting and large-scale inference used throughout the experiments; established the fine-tuning framework for downstream evaluation.
Enrico Bacis: Built the library for unified data representation across diverse datasets and researched suitable datasets; developed the evaluation framework for few-shot experiments and conducted these experiments.
Cesar Ilharco: Developed metrics for evaluating taxonomies and conducted the corresponding experiments; researched and collected suitable datasets for taxonomy evaluation.
Hamza Harkous: Founded and led research for Simula; designed the core system for end-to-end seedless data generation; researched and implemented the original system components, including mixing strategies, taxonomy building, meta-prompting, critiquing, and refinement; conducted downstream fine-tuning experiments for evaluation; developed the data generation setup for ablation studies.
All authors: Contributed to project ideation, technical discussions, and manuscript writing and revision.

ACKNOWLEDGMENTS

The authors would like to acknowledge Sai Teja Peddinti for feedback on early drafts of this work. We would like also to thank Nina Taft and Amanda Walker for continuous support and advice on the project and the paper. We also thank Coran Corbett for engineering contributions to Simula.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Anthropic. The claude 3 model family: Opus, sonnet, haiku, 2024. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Michele Banko, Brendon MacKeen, and Laurie Ray. A unified taxonomy of harmful content. In Seyi Akiwowo, Bertie Vidgen, Vinodkumar Prabhakaran, and Zeerak Waseem (eds.), *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 125–137, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.alw-1.16. URL https://aclanthology.org/2020.alw-1.16/.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020.

Olaf Bánki, Yury Roskov, Markus Döring, Geoff Ower, Diana Raquel Hernández Robles, Camila Andrea Plata Corredor, Thomas Stjernegaard Jeppesen, Ari Örn, Thomas Pape, Donald Hobern,

Stephen Garnett, Holly Little, R. Edward DeWalt, Keping Ma, Joe Miller, Thomas Orrell, Rolf Aalbu, John Abbott, Carlos Aedo, and Others. Catalogue of life, 1 2025. ISSN 2405-8858. URL https://www.checklistbank.org/dataset/307664.

Hao Chen, Abdul Waheed, Xiang Li, Yidong Wang, Jindong Wang, Bhiksha Raj, and Marah I Abdin. On the diversity of synthetic data and its impact on training large language models. *arXiv preprint arXiv:2410.15226*, 2024.

Zhihong Chen, Feng Jiang, Junying Chen, Tiannan Wang, Fei Yu, Guiming Chen, Hongbo Zhang, Juhao Liang, Chen Zhang, Zhiyi Zhang, et al. Phoenix: Democratizing chatgpt across languages. *arXiv preprint arXiv:2304.10453*, 2023.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *JMLR*, 2024.

Gary N. Curtis. Taxonomy of logical fallacies. https://www.fallacyfiles.org/taxonnew.html, 2023. Accessed 2025-01-23.

Tim R. Davidson, Viacheslav Surkov, Veniamin Veselovsky, Giuseppe Russo, Robert West, and Caglar Gulcehre. Self-recognition in language models. In *EMNLP*, 2024a.

Tim Ruben Davidson, Veniamin Veselovsky, Michal Kosinski, and Robert West. Evaluating language model agency through negotiations. In *ICLR*, 2024b.

Arpad E Elo and Sam Sloan. The rating of chessplayers: Past and present, 1978.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *EMNLP-IJCNLP*, 2019.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *International Conference on Machine Learning*, pp. 5988–6008. PMLR, 2022.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. In *ICML*, 2024.

Richard V. Gaines, H. Catherine W. Skinner, Eugene E. Foord, Brian Mason, and Abraham Rosenzweig. *Dana's New Mineralogy: The System of Mineralogy of James Dwight Dana and Edward Salisbury Dana*. John Wiley & Sons, New York, 8th edition, 1997. ISBN 978-0471193104. With sections by Vandall T. King.

Gemini, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemini, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Gemma, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 2023.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Himanshu Gupta, Kevin Scaria, Ujjwala Anantheswaran, Shreyas Verma, Mihir Parmar, Saurabh Arjun Sawant, Chitta Baral, and Swaroop Mishra. Targen: Targeted data generation with large language models. *COLM*, 2024.

Alex Havrilla, Andrew Dai, Laura O'Mahony, Koen Oostermeijer, Vera Zisler, Alon Albalak, Fabrizio Milo, Sharath Chandra Raparthy, Kanishk Gandhi, Baber Abbasi, et al. Surveying the effects of quality, diversity, and complexity in synthetic data from large language models. *arXiv preprint arXiv:2412.02980*, 2024.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. In *NeurIPS*, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 2017.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, 2015.

Tom Hosking, Phil Blunsom, and Max Bartolo. Human feedback is not gold standard. In *ICLR*, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022.

Yuzheng Hu, Pingbang Hu, Han Zhao, and Jiaqi W Ma. Most influential subset selection: Challenges, promises, and beyond. *NeurIPS*, 2024.

Audrey Huang, Adam Block, Dylan J Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan T Ash, and Akshay Krishnamurthy. Self-improvement in language models: The sharpening mechanism. In *NeurIPS Workshop on Mathematics of Modern Machine Learning*, 2024.

Jonas Hübotter, Sascha Bongni, Ido Hakimi, and Andreas Krause. Efficiently learning at test-time: Active fine-tuning of llms. In *NeurIPS Workshop on Mathematics of Modern Machine Learning*, 2024.

International Union of Pure and Applied Chemistry (IUPAC). Periodic table of the elements. https://iupac.org/what-we-do/periodic-table-of-elements/, 2022. Accessed: 2025-01-23; Standard atomic weights current as of 2022-05-04.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

Angelika Kaplan, Thomas Kühn, Sebastian Hahner, Niko Benkler, Jan Keim, Dominik Fuchß, Sophie Corallo, and Robert Heinrich. Introducing an evaluation method for taxonomies. In *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering*, pp. 311–316, 2022.

Abhinav Ramesh Kashyap, Thanh-Tung Nguyen, Viktor Schlegel, Stefan Winkler, See-Kiong Ng, and Soujanya Poria. A comprehensive survey of sentence representations: From the bert epoch to the chatgpt era and beyond. *arXiv preprint arXiv:2305.12641*, 2023.

Seungone Kim, Juyoung Suk, Xiang Yue, Vijay Viswanathan, Seongyun Lee, Yizhong Wang, Kiril Gashteovski, Carolin Lawrence, Sean Welleck, and Graham Neubig. Evaluating language models as synthetic data generators. *arXiv preprint arXiv:2412.03679*, 2024.

Dennis Kundisch, Jan Muntermann, Anna Maria Oberländer, Daniel Rau, Maximilian Röglinger, Thorsten Schoormann, and Daniel Szopinski. An update for taxonomy designers: methodological guidance from information systems research. *Business & Information Systems Engineering*, pp. 1–19, 2021.

Alon Lavie and Abhaya Agarwal. METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. In Chris Callison-Burch, Philipp Koehn, Cameron Shaw Fordyce, and Christof Monz (eds.), *Proceedings of the Second Workshop on Statistical Machine Translation*, pp. 228–231, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/W07-0734/.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: scaling reinforcement learning from human feedback with AI feedback. In *ICML*, 2024a.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, et al. Gecko: Versatile text embeddings distilled from large language models. *arXiv preprint arXiv:2403.20327*, 2024b.

Nicholas Lee, Thanakul Wattanawong, Sehoon Kim, Karttikeya Mangalam, Sheng Shen, Gopala Anumanchipalli, Michael W. Mahoney, Kurt Keutzer, and Amir Gholami. LLM2LLM: boosting llms with novel iterative data enhancement. In *ACL*, 2024c.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *CoRR*, 2024a.

Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, et al. Synthetic data (almost) from scratch: Generalized instruction tuning for language models. *arXiv preprint arXiv:2402.13064*, 2024b.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.

Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. #instag: Instruction tagging for analyzing supervised fine-tuning of large language models. In *ICLR*, 2024.

Sheng Lu, Shan Chen, Yingya Li, Danielle Bitterman, Guergana Savova, and Iryna Gurevych. Measuring pointwise $\mathcal{V}$-usable information in-context-ly. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15739–15756, 2023.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *NeurIPS*, 36, 2024.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When less is more: Investigating data pruning for pretraining llms at scale. In *NeurIPS Workshop on Attributing Model Behavior at Scale*, 2023a.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. When less is more: Investigating data pruning for pretraining llms at scale. *CoRR*, 2023b.

Anay Mehrotra, Manolis Zampetakis, Paul Kassianik, Blaine Nelson, Hyrum S Anderson, Yaron Singer, and Amin Karbasi. Tree of attacks: Jailbreaking black-box LLMs automatically. In *NeurIPS*, 2024.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *NeurIPS*, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Rajvardhan Patil, Sorio Boit, Venkat Gudivada, and Jagadeesh Nandigam. A survey of text representation and embedding techniques in nlp. *IEEE Access*, 11:36120–36146, 2023.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *NeurIPS*, 2021.

Crystal Qian, Michael Xieyang Liu, Emily Reif, Grady Simon, Nada Hussein, Nathan Clement, James Wexler, Carrie J Cai, Michael Terry, and Minsuk Kahng. The evolution of llm adoption in industry data curation practices. *arXiv preprint arXiv:2412.16089*, 2024.

Emily Reif, Crystal Qian, James Wexler, and Minsuk Kahng. Automatic histograms: Leveraging language models for text dataset exploration. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2024. ISBN 9798400703317. doi: 10.1145/3613905.3650798.

Laura Ruis, Maximilian Mozes, Juhan Bae, Siddhartha Rao Kamalakara, Dwarak Talupuru, Acyr Locatelli, Robert Kirk, Tim Rocktäschel, Edward Grefenstette, and Max Bartolo. Procedural knowledge in pretraining drives reasoning in large language models. *arXiv preprint arXiv:2411.12580*, 2024.

Mikayel Samvelyan, Sharath Chandra Raparthy, Andrei Lupu, Eric Hambro, Aram H Markosyan, Manish Bhatt, Yuning Mao, Minqi Jiang, Jack Parker-Holder, Jakob Nicolaus Foerster, et al. Rainbow teaming: Open-ended generation of diverse adversarial prompts. In *NerIPS*, 2024.

William Saunders, Catherine Yeh, Jeff Wu, Steven Bills, Long Ouyang, Jonathan Ward, and Jan Leike. Self-critiquing models for assisting human evaluators. *arXiv preprint arXiv:2206.05802*, 2022.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. Synthetic prompting: Generating chain-of-thought demonstrations for large language models. In *ICML*, volume 202, pp. 30706–30775. PMLR, 2023.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Esin DURMUS, Zac Hatfield-Dodds, Scott R Johnston, Shauna M Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models. In *ICLR*, 2024.

Avi Singh, John D. Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J. Liu, James Harrison, Jaehoon Lee, Kelvin Xu, Aaron T. Parisi, Abhishek Kumar, Alexander A. Alemi, Alex Rizkowsky, Azade Nova, Ben Adlam, Bernd Bohnet, Gamaleldin Fathy Elsayed, Hanie Sedghi, Igor Mordatch, Isabelle Simpson, Izzeddin Gur, Jasper Snoek, Jeffrey Pennington, Jiri Hron, Kathleen Kenealy, Kevin Swersky, Kshiteej Mahajan, Laura Culp, Lechao Xiao, Maxwell L. Bileschi, Noah Constant, Roman Novak, Rosanne Liu, Tris Warkentin, Yundi Qian, Yamini Bansal, Ethan Dyer, Behnam Neyshabur, Jascha Sohl-Dickstein, and Noah Fiedel. Beyond human data: Scaling self-training for problem-solving with language models. *TMLR*, 2024a.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, et al. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *arXiv preprint arXiv:2412.03304*, 2024b.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. The good, the bad, and the greedy: Evaluation of llms should not ignore non-determinism. *arXiv preprint arXiv:2407.10457*, 2024.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *EMNLP*, 2020.

Daniel Szopinski, Thorsten Schoormann, and Dennis Kundisch. Criteria as a prelude for guiding taxonomy evaluation. In *HICSS*, pp. 1–10, 2020.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *EMNLP*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Muhammad Usman, Ricardo Britto, Jürgen Börstler, and Emilia Mendes. Taxonomies in software engineering: A systematic mapping study and a revised taxonomy development method. *Information and Software Technology*, 85:43–59, 2017.

Boris van Breugel, Nabeel Seedat, Fergus Imrie, and Mihaela van der Schaar. Can you rely on your model evaluation? improving model evaluation with synthetic test data. *NeurIPS*, 2024.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Position: will we run out of data? limits of llm scaling based on human-generated data. In *ICML*, 2024.

Vijay Viswanathan, Kiril Gashteovski, Carolin Lawrence, Tongshuang Wu, and Graham Neubig. Large language models enable few-shot clustering. *TACL*, 12:321–333, 2024.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *ACL*, 2023.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pp. 214–229, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533088. URL https://doi.org/10.1145/3531146.3533088.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: selecting influential data for targeted instruction tuning. In *ICLR*, 2024.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *ICLR*, 2024.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. Wizardlm: Empowering large language models to follow complex instructions. *ICLR*, 2024.

Jiasheng Ye, Peiju Liu, Tianxiang Sun, Yunhua Zhou, Jun Zhan, and Xipeng Qiu. Data mixing laws: Optimizing data mixtures by predicting language modeling performance. *CoRR*, 2024.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. Large language model as attributed training data generator: A tale of diversity and bias. *NeurIPS*, 2023.

Yunhao Zhang and Renée Gosline. Human favoritism, not ai aversion: People's perceptions (and bias) toward generative ai, human experts, and human–gai collaboration in persuasive content generation. *Judgment and Decision Making*, 18:e41, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *NeurIPS*, 2023.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

APPENDICES

A Related Work

A.1 Synthetic dataset evaluation.

Early Evaluation Methods. When evaluating a synthetic dataset, we can differentiate between *comparative* and *intrinsic* evaluation. The former expects the availability of a reference dataset. In the case of "closed" tasks that allow for narrow comparisons, we can directly compare model-generated outputs to reference solutions, e.g., single-word answers, translations, or summaries. Originally, these were done directly on the observed outputs, e.g., by comparing overlapping words or n-grams (Papineni et al., 2002; Lin, 2004; Lavie & Agarwal, 2007), *inter alia*.

Rise of Semantic Metrics. As tasks might have multiple different, but semantically equivalent solutions, evaluations shifted to embedding-based approaches (Patil et al., 2023). Such approaches can even be used in the absence of one-to-one reference mappings, by comparing output distributions on a dataset level (Heusel et al., 2017; Pillutla et al., 2021). Although embedding-based approaches allow for automated evaluation, they can struggle in specialized domains and miss semantic subtleties (Ethayarajh, 2019; Kashyap et al., 2023).

Auto-Verifiable Tasks. Then, we have a class of problems that allow for many trajectories ending in a final solution whose correctness can be automatically verified quickly, e.g., program synthesis and mathematical reasoning (Song et al., 2024), or negotiation games (Davidson et al., 2024b).

The Challenge of Open Tasks. What remains is a large class of "open" tasks, for which there are no well-defined criteria of correctness, e.g., creative writing, advice, and most visual media. For these, we largely rely on the reasoning capabilities of human annotators to provide quality references and to obtain pairwise preference labels (Christiano et al., 2017; Bai et al., 2022). However, creating such datasets manually is expensive, time-consuming, and error-prone (Chen et al., 2023; Gilardi et al., 2023; Hosking et al., 2024).

Emergence of LLM Judges. Recently, the growing capabilities of frontier models have opened up the possibility of model-based reasoning (Lee et al., 2024a; Zheng et al., 2023; Li et al., 2024a; Saunders et al., 2022; Wang et al., 2023; Madaan et al., 2024), *inter alia*. As humans and AI start to prefer AI-generated text (Zhang & Gosline, 2023; Panickssery et al., 2024), and AI prefers text by the strongest AI models (Davidson et al., 2024a), model-based evaluation is set to become increasingly prevalent.

Intrinsic Evaluation Metrics. In the absence of reference data, one must rely on *intrinsic* evaluation. As described in the introduction, we typically focus on the axes of quality, diversity, and complexity. When we treat quality as how well a set of data points meets stated requirements, we are quickly forced to rely on reasoning-based evaluation. Instead, if understood as the utility of a dataset for a specific downstream task, we can directly measure the lift in downstream performance, e.g., through classification accuracy. Diversity is often approximated using pairwise cosine similarity after embedding outputs into a higher dimensional space (Yu et al., 2023; Gupta et al., 2024). Without grouping the data using an appropriate clustering step, e.g., semantic clusters, average statistics are sensitive to outliers and fail to differentiate between global and local diversity. Crucially, such diversity statistics provide few semantic, actionable insights. Attempts to automate complexity scoring range from measuring the length of outputs (Shao et al., 2023), or the relative entropy over output alternatives (Ethayarajh et al., 2022; Lu et al., 2023), to generating a large solution set and using a reward model to estimate correctness (Snell et al., 2024). Reasoning-based approaches instead directly query a model to provide a "difficulty" score (Li et al., 2024a). Because models are generally poorly calibrated (Zheng et al., 2023; Tian et al., 2023; Xiong et al., 2024), and difficulty is a relative concept, such absolute scores can be noisy.

Our Reasoning-Based Approach. Our work continues the trend of incorporating model-based reasoning to evaluate data. Instead of using approximate statistics based on output embeddings, reasoning-based approaches provide explainable traces that can be audited and controlled. Mapping out a global coverage space using taxonomies allows end-users to quickly evaluate if their generated data meets the appropriate global diversity requirements (Sections 2.1, 2.3). We further carefully tested popular assumptions about the use of model-based critics through a series of controlled experiments. On the evaluated datasets, we find that critic-rejection sampling of synthetic outputs consistently succeeds in increasing average sample quality (Sections 2.2, 3.1, 4.1) across different

datasets and complexity levels. We also found that model-assigned complexity scores are promising proxies for human notions of difficulty, and correlate well with models' critic and generation capabilities (Sections 2.3, 4.1, and Appendix D).

A.2 SYNTHETIC DATASET GENERATION.

Seed-Based Expansion Methods. Popular synthetic data methods generally only account for a subset of the quality, diversity, and complexity axes (Havrilla et al., 2024). For example, Wang et al. (2023) start with a set of seed examples and iteratively expand them using hand-crafted semantic diversity prompts. Xu et al. (2024) similarly expand seed examples, but focus on both diversity and complexity. The main quality check performed is to ensure that these expanded examples do not become degenerative. An attempt at increasing global diversity is done by maximizing pairwise cosine similarity through rejection sampling. The authors note that expanding semantic diversity and complexity are both positively correlated with downstream performance after fine-tuning.

Factor Identification Approaches. Reif et al. (2024); Chen et al. (2024); Viswanathan et al. (2024); Lu et al. (2024) use seed examples to automatically detect relevant factors through iterative sampling of the target dataset, after which factors are extracted using reasoning modules. Other approaches side-step the need for seed examples by manually inspecting or reasoning about a target dataset to find globally relevant factors of variation (Yu et al., 2023; Gupta et al., 2024; Samvelyan et al., 2024). They then sample from these factors for conditional generation. Relevant to our framework is work done by Li et al. (2024b), who attempt to generate a single, large taxonomy to cover a variety of topics. Inspired by curriculum-based learning in human education systems, the authors generate a variety of learning modules to implicitly vary complexity.

Leveraging the Critic Gap. Many have by now pointed out the apparent gap between current models' generative and verification capabilities (Huang et al., 2024). This gap allows models to act as critics of their own outputs (Saunders et al., 2022; Madaan et al., 2024) and has been successfully used by many of the above methods (Lee et al., 2024c; Gupta et al., 2024; Samvelyan et al., 2024; Chen et al., 2024), *inter alia*.

Scaling Test-Time Compute. Recent efforts in scaling test-time computation show that models are capable of generating correct outputs even for complex questions, given enough attempts (Song et al., 2024; Brown et al., 2024). Yet, how to best scale such a test-time computation budget depends on the complexity of the particular problem (Snell et al., 2024). The authors suggest that "easier" tasks most benefit from exploiting an existing output attempt through iterative refinement, whereas more "difficult" tasks benefit more from exploring a larger proposal distribution.

Our Orchestration Approach. With Simula, we build on many of the existing insights into the merits of optimizing quality, diversity, and complexity for downstream performance. In contrast to existing methods, we explicitly orchestrate the generative process on a dataset level to increase global control and explainability. We carefully split the data generation process into separate steps, i.e., global diversity, local diversity, complexity, and correctness, allowing end users to tailor datasets to their specific requirements (Section 2). In doing so, it becomes possible to allocate computational resources where they are most desired, e.g., by generating more meta prompts to increase the proposal distribution of complex samples or adding additional critic steps to refine the outputs.

B SUPPLEMENTARY: TAXONOMY EVALUATION

To address the inherent challenges of assessing taxonomy quality and completeness, we use a critic-model based framework for evaluation. Traditional taxonomy evaluation often relies on manual expert review, which is time-consuming, expensive, and often difficult to attain. Our proposed framework leverages the capabilities of multi-modal models (M3s) as "critic models" to provide a more automated, scalable, and reproducible evaluation alternative.

B.1 DEFINING TAXONOMY EVALUATION METRICS

First, a hierarchical representation for each expert taxonomy $\mathcal{T}_E$, $\mathcal{T}_{\text{M3}}$ is provided and the critic model classifies each node into the following category.

- **Good and Overlapping**: The node is good (well-defined, relevant to its parent node, and fits appropriately within the overall taxonomy) AND overlapping (there is a semantically equivalent node in the other taxonomy which represents the same concept).
- **Good and Exclusive**: The node is good (well-defined, relevant to its parent node, and fits appropriately within the overall taxonomy) AND NOT overlapping (there is a no semantically equivalent node in the other taxonomy which represents the same concept; this concept appears uniquely in this taxonomy).
- **Redundant**: The node is a duplicate within its own taxonomy, there is another node in the same taxonomy representing the same concept.
- **Bad**: The node is irrelevant, poorly defined, misclassified, or otherwise inappropriate for its position in the taxonomy.

Based on the critic model's classifications, we compute several quantitative metrics to evaluate each taxonomy:

- **Completeness**: This metric measures the extent to which $\mathcal{T}_{\text{M3}}$ covers the concepts present in $\mathcal{T}_E$. The M3 critic assesses, for each node (concept) in $\mathcal{T}_E$, whether a semantically equivalent node exists in $\mathcal{T}_{\text{M3}}$. This serves as a measure of coverage and recall, quantified by the ratio (Good and Overlapping) / (Total Good) in $\mathcal{T}_E$. Here, Total Good = Good and Overlapping + Good and Exclusive.
- **Soundness**: This metric assesses the proportion of relevant and correct nodes within $\mathcal{T}_{\text{M3}}$. The M3 critic examines each node in $\mathcal{T}_{\text{M3}}$ to judge its relevance to the topic and whether it constitutes a non-redundant entry. Fewer irrelevant or incorrect nodes result in greater soundness. This serves as a measure of precision, quantified by the ratio (Total Good) / (Total Nodes) in $\mathcal{T}_{\text{M3}}$. Here, Total Good = Good and Overlapping + Good and Exclusive.

It is worth noting here, that there are different taxonomy types in practice. *Grounded* taxonomies are typically revised over time as new empirical evidence is gathered through the scientific method. A recently accepted version in the literature can be considered closer to a ground truth than a conceptual taxonomy, in that it offers less scope for an M3 to generate new terms absent new evidence. We use the following grounded taxonomies for evaluation: "*Animal Phylogenetic Classes*" (Bánki et al., 2025), "*Periodic Chemical Elements*" (International Union of Pure and Applied Chemistry (IUPAC), 2022), and "*Mineral Classification*" (Gaines et al., 1997).

Conceptual taxonomies are more subjective than grounded ones; even the definitions and usage of terminology can vary across the literature (Usman et al., 2017; Szopinski et al., 2020; Kundisch et al., 2021; Kaplan et al., 2022). We use the following conceptual taxonomies for evaluation: "*Risk of Language Models*" (Weidinger et al., 2022), "*Online Harmful Content*" (Banko et al., 2020), and "*Logical Fallacies*" (Curtis, 2023)).

Because of this subjectivity, we additionally compute the following metrics for conceptual taxonomies:

- **Novelty**: This metric assesses whether $\mathcal{T}_{\text{M3}}$ contains relevant nodes that are not present in $\mathcal{T}_E$. The M3 critic identifies nodes in $\mathcal{T}_{\text{M3}}$ that are not semantically equivalent to any node in

$\mathcal{T}_E$ and then judges the relevance of these novel nodes. A higher number of relevant novel nodes indicates greater novelty. We define this as the ratio (Good and Exclusive in $\mathcal{T}_{\text{M3}}$) / (Total Good in $\mathcal{T}_E$).

- **Coverage**: This represents the total number of "good" items in $\mathcal{T}_{\text{M3}}$ relative to the total number of "good" items in $\mathcal{T}_E$. Coverage is equivalent to Completeness + Novelty and provides a comparative metric of the number of sound items. It follows that a coverage value greater than 1.0 indicates that $\mathcal{T}_{\text{M3}}$ covers more relevant and correct items than $\mathcal{T}_E$ within the global space for the given taxonomy.

A more elaborate description of the grounded and conceptual taxonomies used can be found in B.2.

B.2 Taxonomy Evaluation Results

Table 4: **Performance metrics for different taxonomies.**

Type	Topic	Method	Completeness (γ)	Soundness (σ)	Novelty (ν)	Coverage (τ)
Conceptual	Online Harmful Content	Simula	**0.749**	**0.980**	**0.865**	**1.614**
		0-shot	0.588	0.957	0.412	1.000
	Logical Fallacies	Simula	**0.726**	0.919	**1.679**	**2.405**
		0-shot	0.458	**0.966**	0.193	0.651
	Risks of Language Models	Simula	**0.867**	**1.000**	0.267	**1.134**
		0-shot	0.467	**1.000**	**0.367**	0.834
Grounded	Animal Phylogenetic Classes	Simula	**0.458**	**0.926**	—	
		0-shot	0.349	0.918		
	Periodic Chemical Elements	Simula	**0.993**	**0.987**	—	
		0-shot	0.775	0.864		
	Mineral Classification	Simula	**0.762**	**0.340**	—	
		0-shot	0.442	0.329		

Description of the taxonomies from Table 4:

- **[Conceptual] Online Harmful Content:** This taxonomy aims to provide a unified classification of harmful content found online. It synthesizes common abuse types described by industry content policies, policy recommendations, community standards, and health expert guidelines. The goal is to create readily usable categories for content moderation, encourage the development of accurate datasets for model training, and raise awareness of less-studied abuse types to improve online safety. This taxonomy categorizes different types of harmful content found online into four main groups: Hate and Harassment; Self-Inflicted Harm; Ideological Harm; and Exploitation — and further branches each into a set of more specific types. (Banko et al., 2020).
- **[Conceptual] Logical Fallacies:** This taxonomy classifies types of logical fallacies into a hierarchical structure. It divides fallacies into two main branches: Formal Fallacies (errors in the structure of the argument) and Informal Fallacies (errors in the content or context of the argument). These main categories are further subdivided into numerous specific types of fallacies, such as Propositional Fallacies, Quantificational Fallacies, and various informal fallacies like Appeal to Ignorance, and Red Herring, and then branching down to increasingly specific types (Curtis, 2023).
- **[Conceptual] Risks of Language Models:** This taxonomy identifies ethical and social risks associated with large-scale language models (LMs). It categorizes these risks into six areas: Discrimination, Hate speech and Exclusion; Information Hazards; Misinformation Harms, Malicious Uses, Human-Computer Interaction Harms, and Environmental and Socioeconomic harms. The taxonomy distinguishes between "observed" risks (already evidenced in LMs) and "anticipated" risks (considered likely but not yet observed). The goal is to provide a comprehensive framework for understanding and mitigating the potential negative consequences of LMs. (Weidinger et al., 2022).

- **[Grounded] Animal Phylogenetic Classes:** This taxonomy represents the hierarchical classification of animals based on their evolutionary relationships. It is truncated to two levels deeper into the animal kingdom, encompassing its phyla and classes (Bánki et al., 2025).
- **[Grounded] Periodic Chemical Elements:** This taxonomy organizes chemical elements into a hierarchical structure based on their atomic number, electron configuration, and recurring chemical properties, primarily reflecting their placement in the periodic table. It positions elements into groups from Group 1 - Alkali Metals through Group 18 - Noble Gases, as well as the Lanthanides and Actinides. Each group further lists individual elements like Sodium (Na) or Gold (Au). The structure represents the periodic trends and shared characteristics within groups, enabling chemists to understand relationships and predict elemental behavior (International Union of Pure and Applied Chemistry (IUPAC), 2022).
- **[Grounded] Mineral Classification:** Presented in Dana's New Mineralogy, Eighth Edition, this taxonomy is a hierarchical classification system for minerals, employing a four-part numerical code to categorize each species (Gaines et al., 1997). This system, analogous to the Linnaean taxonomy for biology, organizes minerals based on both their chemical composition and crystal structure. The first number denotes the mineral's class (e.g., anhydrous carbonates), reflecting broad compositional categories or dominant structural features (especially in silicates). The second number signifies the mineral's type, sometimes based on atomic properties, or formula. The third number groups minerals with similar structural arrangements. Finally, the fourth number uniquely identifies the individual mineral species, such as Calcite or Magnesite within the Calcite Group. This numerical system offers a structured and expandable framework, allowing new minerals to be easily integrated while highlighting the chemical and structural relationships between different mineral species.

B.3 Limitations

- **Preference Bias.** As discussed in the Related Work Section A, there is evidence that M3s prefer model-generated text over text generated by humans. In our case, we do not prompt the M3 to express a preference. Rather, we ask if certain nodes semantically approximate other nodes, or if certain nodes are appropriate given the context. However, we did use models from the same model family for both the generation and the evaluation. Thus, future work might want to repeat this experiment with separate generator and critic models.
- **Stochastic Sensitivity.** We did not optimize prompts for each separate taxonomy. As such, the reported metrics likely represent lower bounds.
- **Downstream Application.** While we performed a comparative evaluation of synthetic and real taxonomies, it is not directly clear which are better suited for certain downstream applications.

B.4 QUALITATIVE EXAMPLES

(a) Harmful Content Taxonomy — Simula

(b) Harmful Content Taxonomy — Expert

Figure 4: Comparison of Online Harmful Content Taxonomy (Simula vs. Expert).

(a) Chemical Elements — Simula

(b) Chemical Elements — Expert

Figure 5: Comparison of Chemical Element Taxonomy (Simula vs. Expert).

C SUPPLEMENTARY: VERIFICATION VS. GENERATION

C.1 MULTILINGUAL MMLU

In Table 5 we show empirical critic-rejection sampling results for a subset of MMLU questions on Mathematics, Computer Science, and Physics. We use the subjects' education levels (elementary, high-school, and college) as ground truth complexity categories. We evaluate performance on languages with different resource categories, e.g., "Low", "Mid", and "High", according to their recorded, written, and catalogued NLP resources per Singh et al. (2024b). We observe that our critic-rejection sampling strategy is effective for each language under each complexity condition.

Table 5: **Critic-Rejection Sampling on Multilingual MMLU Questions.** We evaluate our critic-rejection sampling method for MMLU questions on Mathematics, Physics, and Computer Science. We use the subject education level (elementary, high-school, and college) as the ground-truth Complexity. We display the realized change in accuracy, μ_{gen} of following critic rejections. Also shown are the average ELO complexity score and the size of the rejected and accepted subsets.

		English (High)			Korean (Mid)			Nepali (Low)		
Complexity	Critic	μ_{gen}	ELO	$\lvert\mathcal{D}\rvert$	μ_{gen}	ELO	$\lvert\mathcal{D}\rvert$	μ_{gen}	ELO	$\lvert\mathcal{D}\rvert$
Level 1	×	0.93 $\pm$0.07	290 $\pm$8	14	0.86 $\pm$0.07	306 $\pm$7	29	0.76 $\pm$0.07	308 $\pm$7	41
	✓	0.98 $\pm$0.01	303 $\pm$2	364	0.97 $\pm$0.01	301 $\pm$2	349	0.97 $\pm$0.01	304 $\pm$2	337
Level 2	×	0.52 $\pm$0.07	427 $\pm$7	46	0.64 $\pm$0.06	428 $\pm$5	53	0.58 $\pm$0.06	431 $\pm$6	60
	✓	0.94 $\pm$0.01	427 $\pm$2	578	0.92 $\pm$0.01	426 $\pm$2	571	0.93 $\pm$0.01	425 $\pm$2	564
Level 3	×	0.67 $\pm$0.10	473 $\pm$5	24	0.54 $\pm$0.10	471 $\pm$5	26	0.41 $\pm$0.09	473 $\pm$4	34
	✓	0.89 $\pm$0.02	467 $\pm$2	278	0.88 $\pm$0.02	468 $\pm$2	276	0.86 $\pm$0.02	465 $\pm$2	268

D SUPPLEMENTARY: COMPLEXITY ANALYSIS

We compare model-assigned complexity scores against the ground-truth human annotations. We ablate model-assigned complexity scores varying the number of times each sample is scored (**N**) and the batch size (**BS**) of questions being scored simultaneously. Importantly, for fixed N, the number of samples being scored simultaneously (BS) increases the context length but reduces the number of inference passes. For example, for $|\mathcal{D}| = 1000$, N=10 and BS=1, we require $10,000$ inference passes. Setting BS to 5 instead reduces this to $10,000/5 = 2,000$. All things equal, in practice we would thus like to see a higher BS to have similar or better performance than a lower BS.

D.1 OPEN GENERATION: MATH

In Figure 6, we show results of comparing model-assigned complexity scores to the human-annotated ground truth (Levels 1-5). To enable side-by-side comparison, we scale both the raw average scores (Score) and the computed ELO rankings (ELO) to lie between 0 and 100 for each {BS, N} grouping. As we increase the number of samples, N, clusters become better separated. Increasing BS > 1 enables the use of latent skill methods like ELO to increase consistency. We found BS = N = 5 to strike an appropriate balance between separation and inference cost.

D.2 MULTIPLE-CHOICE GENERATIONS: MMLU GLOBAL

In Table 6, we compare model-assigned complexity scores for a subset of MMLU questions on Mathematics, Computer Science, and Physics. We use the subjects' education levels (elementary, high-school, and college) as ground truth complexity categories. We evaluate performance on languages with different resource categories, e.g., "Low", "Mid", and "High", according to their recorded, written, and catalogued NLP resources per Singh et al. (2024b). After running KMeans on the model-assigned complexity scores, we compute the Normalized Mutual Information (**NMI**) and the Adjusted Rand Index (**ARI**) to evaluate cluster approximation of the ground truth complexity. Finally, we train a logistic regression on model-assigned complexity scores to evaluate them as

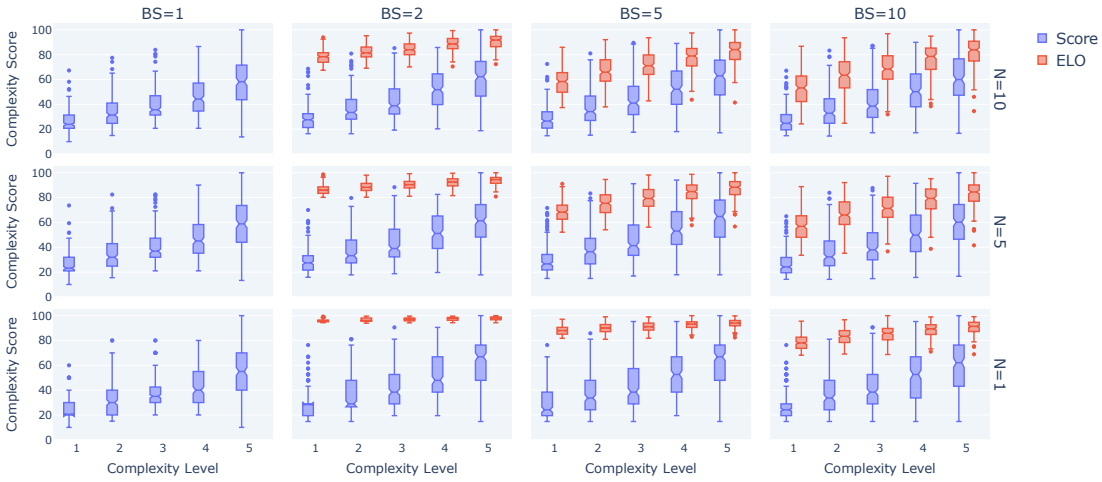

Figure 6: **Complexity score ablation MATH test set.** The MATH dataset comes with ground truth complexity levels ranging from 1-5. We group by the ground truth complexity level and plot the model-assigned complexity scores for each group. We compare using average raw scores (Score) against using ELO rankings (ELO) and ablate the batch size (**BS**) of question scored simultaneously and the number of times each question is scored (**N**). All results are scaled per (BS, N)-grouping to lie between 0 and 100. For BS $= 1$, no pairwise rankings are done, so ELO rankings do not apply.

an estimator for the model's generative performance, reporting the Area Under the Curve (**AUC**). Similar to our findings in Section C.1, we find model-assigned complexity scoring robust across several languages. Choosing BS = N = 5 again emerge as reasonable hyperparameters.

Table 6: **Multilingual MMLU Complexity Scoring.** We use exam questions for the topics Mathematics, Physics, and Computer Science, on education levels elementary (Mathematics only), high-school, and college. Taking the education level as our ground-truth complexity level (1-3), we run KMeans on the complexity scores generated by the model and compute the Normalized Mutual Information (**NMI**) and the Adjusted Rand Index (**ARI**). Finally, we compute the Area Under the Curve (**AUC**) of using the complexity scores as an estimator for the model's generative performance. We ablate the batch size (**BS**) of exam questions scored simultaneously and the number of times each exam question is scored (**N**).

BS	**N**	Enlish (High)			Arabic (High)			Dutch (High)			Korean (Mid)			Nepali (Low)		
		NMI	ARI	AUC	NMI	ARI	AUC	NMI	ARI	AUC	NMI	ARI	AUC	NMI	ARI	AUC
1	1	0.27	0.23	0.61	0.32	0.27	0.60	0.32	0.28	0.59	0.32	0.27	0.59	0.32	0.27	0.56
	5	0.27	0.24	0.62	0.33	0.28	0.61	0.34	0.30	0.59	0.34	0.29	0.59	0.33	0.29	0.56
5	1	0.36	0.31	0.63	0.38	0.32	0.60	0.37	0.31	0.59	0.38	0.33	0.59	0.37	0.32	0.57
	5	0.42	0.36	0.64	0.42	0.35	0.62	0.41	0.35	0.59	0.40	0.34	0.60	0.41	0.35	0.57
10	1	0.38	0.33	0.64	0.39	0.34	0.61	0.37	0.32	0.60	0.40	0.35	0.61	0.39	0.33	0.57
	5	0.42	0.37	0.63	0.41	0.36	0.62	0.40	0.34	0.60	0.43	0.38	0.61	0.43	0.37	0.57

E SUPPLEMENTARY: INTRINSIC EVALUATION RESULTS

Table 7 and Table 8 display intrinsic evaluation results for the various synthetic data method and the real target datasets. We first note that the various Simula components generally improve over the baseline, with the exception of the complexity score on `rte`. We further note that combining optimizations towards diversity (**TS**) and complexity (**TC**), results in both better diversity and complexity metrics (**TSC**). As expected, the "expansion baseline" (**B**) has poor local diversity metrics on most datasets. However, this lack of diversity is less clear from the *global* metric Δ_{global}. Because each method is given the same sample budget, i.e., the size of the real data (**R**), this is a clear indication that average cosine similarity is a limited metric for fine-grained diversity evaluation.

Table 7: **Intrinsic evaluation on Multilingual MMLU.** We compare intrinsic metric for various data generation methods to real data (**R**) on MMLU questions (Mathematics, Physics, and Computer Science). Here Δ_{global} is the *global* average pairwise cosine similarity, Δ_{local} the *local* average pairwise cosine similarity of the $k = 10$ nearest neighbors, and ϕ the global dataset complexity. Δ_{global} and Δ_{local} are diversity measures, so values scores are better. All reported diversity values have SE < 0.001. Complexity values are assigned between 0-100, after which we report the lift (if any) over the real data. Reported complexity values have SE < 1.

	English (High)			Korean (Mid)			Nepali (Low)		
Method	Δ_{global}	Δ_{local}	ϕ	Δ_{global}	Δ_{local}	ϕ	Δ_{global}	Δ_{local}	ϕ
R	0.36	0.64	40	0.59	0.83	41	0.64	0.85	43
B	0.39	0.82	+2	0.58	0.90	+2	0.69	0.93	0
TO	0.36	0.70	+9	0.55	0.82	+13	0.65	0.87	+8
TS	0.36	0.68	+12	0.57	0.81	+13	0.65	0.86	+10
TC	0.36	0.70	+14	0.53	0.80	+15	0.56	0.82	+13
TSC	0.36	0.68	+15	0.54	0.78	+17	0.53	0.80	+14

Table 8: **Intrinsic evaluation on SuperGLUE subsets.** We compare intrinsic metric for various data generation methods to real data (**R**) on a subset of the SuperGLUE tasks (boolq, rte, copa). Here Δ_{global} is the *global* average pairwise cosine similarity, Δ_{local} the *local* average pairwise cosine similarity of the $k = 10$ nearest neighbors, and ϕ the global dataset complexity. Δ_{global} and Δ_{local} are diversity measures, so values scores are better. All reported diversity values have SE < 0.001. Complexity values are assigned between 0-100, after which we report the lift (if any) over the real data. Reported complexity values have SE < 1.

	boolq			rte			copa		
Method	Δ_{global}	Δ_{local}	ϕ	Δ_{global}	Δ_{local}	ϕ	Δ_{global}	Δ_{local}	ϕ
R	0.30	0.62	20	0.51	0.66	28	0.61	0.76	16
B	0.40	0.92	+1	0.52	0.74	+8	0.59	0.82	+4
TO	0.36	0.71	+5	0.55	0.73	-2	0.53	0.71	+12
TS	0.35	0.66	+5	0.55	0.71	0	0.53	0.70	+13
TC	0.37	0.70	+17	0.55	0.73	+2	0.53	0.71	+19
TSC	0.37	0.66	+17	0.52	0.70	+5	0.53	0.70	+19

F SUPPLEMENTARY: DOWNSTREAM RESULTS

In Figure 7 we show the "Performance Gap Recovery" (PGR) metric from Kim et al. (2024), complementing the results in Section 4.3. The metric is computed as:

$$\text{PGR} = (\mu_{\mathcal{D}} - \mu_{\emptyset})/(\mu_{\text{ref}} - \mu_{\emptyset}), \tag{3}$$

where μ_{ref} represents 5-shot Gemma 2B IT with COT, $\mu_{\emptyset}$ 5-shot PT Gemma 2B with COT, and $\mu_{\mathcal{D}}$ our fine-tuned models on synthetic data. To interpret this metric, we note that a PGR value of +50% means that synthetic data generation has recovered 50% of the improvement achieved by the reference instruction tuned model relative to a baseline pre-trained model. A negative value indicates that the training degraded pre-training performance. The PGR results further solidify our design choices around using critic refinement and avoiding the use of simple few-shot data expansion. Simula consistently allows recovering a large percentage of the performance gap across the various combinations of MMLU subjects and languages.

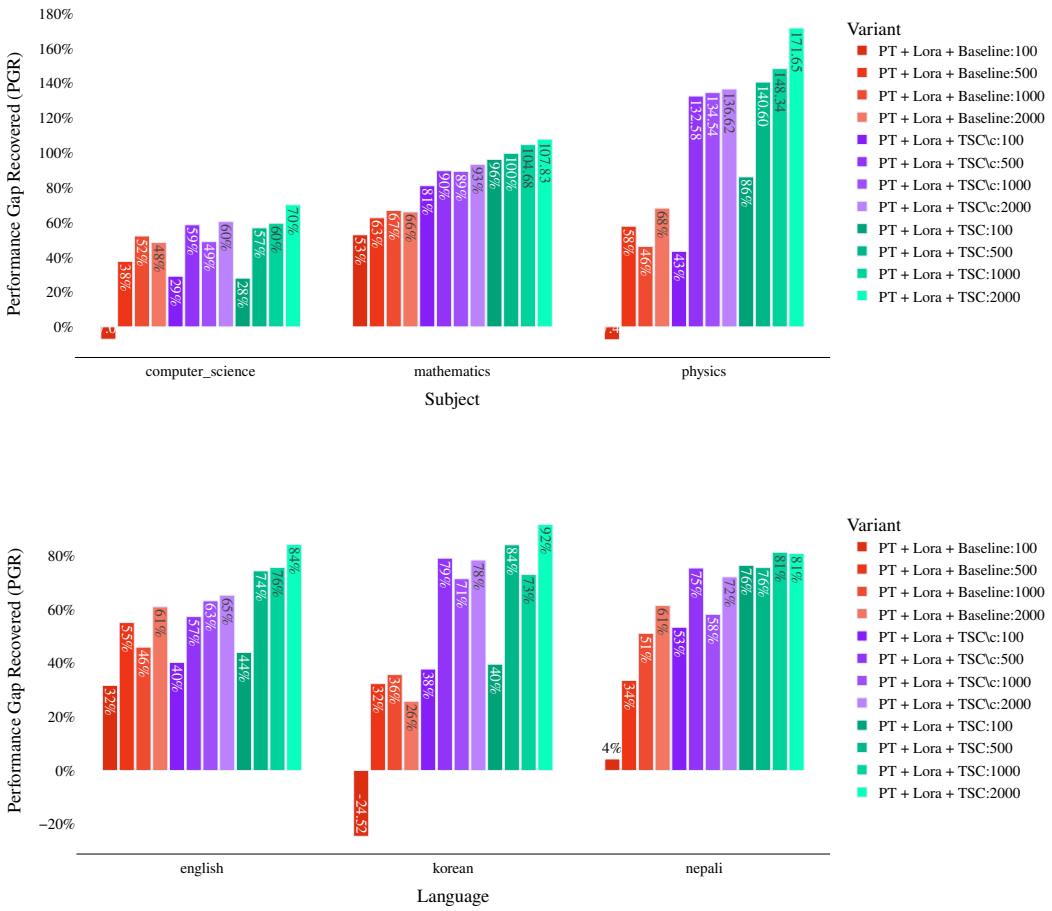

Figure 7: **Performance Gap Recovery (PGR) on subsets of MMLU Global.** Higher PGR implies that the synthetic data generation has been able to recover the gap between the pre-trained model and the instruction-tuned model. The numbers in the legend indicate the data size used per variant.