



Figure 1: **Synthetic Coverage Examples.** The outer squares of (a-b-c) represent a semantic factor of interest, e.g., “*cat type*.” (a) Characterizes “random” sampling behavior, with no notion of a global coverage space. This often results in samples clustered around semantic modes and misses edge cases. The grid-like structures in (b-c) represent discrete semantic spaces defined by a taxonomy’s leaf nodes at increasing levels of granularity. For example, the first level could represent “*cat type*” broken down into “*domestic*, *big wild cats*, *small wild cats*, and *feral*”, whereas a square at a lower level might represent a specific cat breed like the “*British shorthair*.” (b) Represents perfect global planning at increasing granularity; and (c) shows global planning with progressive coverage loss, e.g., missing the branch “*big wild cats*” entirely (bottom left) or missing specific breeds (bottom right).

## 2 Simula: A Reasoning-First Framework for Data Generation and Evaluation

Suppose we want to create 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 this description. This is problematic as it prevents us from developing an actionable notion of coverage, i.e., given a dataset  $\mathcal{D}_y \sim \mathcal{Y}$ , what area of  $\mathcal{Y}$  does it represent?

### 2.1 Using Taxonomies to Capture Dataset Coverage

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

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

A taxonomy is a hierarchical tree structure where the root node represents a broad factor of variation, and child nodes represent increasingly specific sub-categories or instantiations. For instance, “*cat type*” can be broken down into “*domestic*”  $\rightarrow$  “*shorthair*”  $\rightarrow$  “*British shorthair*.” Taxonomies thus serve as a structured map of a concept space, breaking down abstract factors into concrete, sample-able attributes. This 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 granularity comes at a potential cost: with every taxonomy expansion we risk “missing” nodes of interest (e.g., the “*shorthair*” branch), resulting in the progressive coverage loss depicted in Figure 1.c.

To mitigate potential coverage loss resulting from missing nodes, we generate  $\mathcal{T}_i$  by alternating between three steps (full algorithm in App. B.4): (1) Given a node, its ancestors and its siblings, an M3 is prompted  $N$  times to propose an initial set of children nodes. This sampling strategy is inspired by the “Best-of- $N$ ” literature to increase the proposal distribution and cover edge cases (Brown et al., 2024). (2) In a separate call, an M3 is prompted to act as a critic, refining the initial proposals by adding, removing, merging, or editing nodes to improve their completeness, soundness, and specificity, leveraging the generator-critic gap (Huang et al., 2024). Optionally, (3) after generating all nodes of a specific level, an M3 is prompted to generate a “plan” for the next level. This enables consistent and fast parallel generation by ensuring a similar degree of granularity at different node expansions on the same level across independent predictions. At each step, the M3 also has access to the user-provided input  $y$ , and/or a sample  $\mathcal{S}$  from the target distribution.

<sup>1</sup>Note that perfect disentanglement is of course not always possible (Locatello et al., 2019).

## Broader Impact Statement

Simula is a general-purpose framework for generating synthetic datasets. By offering fine-grained control and steerability, Simula introduces a dual-use potential. While our work is motivated by a desire to overcome latent biases and increase explainability in data generation, we recognize that these same capabilities could be leveraged by malicious actors to produce nefarious content or reinforce existing biases. For instance, the ability to precisely manipulate factors of variation could be used to generate datasets that promote harmful stereotypes or misinformation, depending on the intentions of the steering agent (M3 or human).

We highlight that Simula integrates several mechanisms designed to mitigate these risks from happening inadvertently. Firstly, it promotes transparent generation by transforming the opaque nature of real-world data into a “white box” problem. This allows for a clearer understanding and control over potential biases. Secondly, it provides auditing tools that can be used as pragmatic post-checks. The use of taxonomies enables the measurement of data coverage, offering a practical method to detect imbalanced output distributions of specific factors of interest. Similarly, our “calibrated attribute scoring” technique can be employed to identify undesirable shifts in sensitive attributes, such as assessing the level of prejudice. These inherent features aim to facilitate the detection and rectification of biases, thereby promoting responsible and ethical synthetic data generation.