SIMPLESTRAT: DIVERSIFYING LANGUAGE MODEL GENERATION WITH STRATIFICATION

Justin Wong UC Berkeley
Yury Orlovskiy UC Berkeley
Michael Luo UC Berkeley
Sanjit A. Seshia UC Berkeley
Joseph E. Gonzalez UC Berkeley

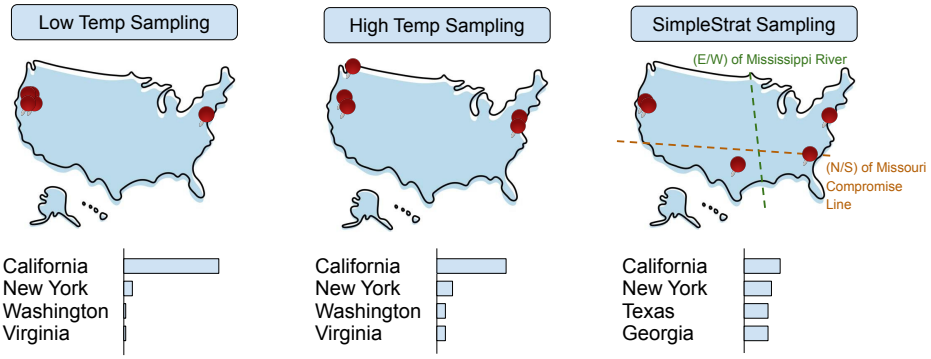

Figure 1: **Stratfied Sampling vs Temperature Scaling** Consider the LLM user request "Name a US State." SimpleStrat employs auto-stratification to utilize the LLM to identify good dimensions of diversity, for instance "East/West of the Mississippi River." Then, SimpleStrat uses stratified sampling to diversify LLM generations.

ABSTRACT

Generating diverse responses from large language models (LLMs) is crucial for applications such as planning/search and synthetic data generation, where diversity provides distinct answers across generations. Prior approaches rely on increasing temperature to increase diversity. However, contrary to popular belief, we show not only does this approach produce lower quality individual generations as temperature increases, but it depends on model's next-token probabilities being similar to the true distribution of answers. We propose SimpleStrat, an alternative approach that uses the language model itself to partition the space into strata. At inference, a random stratum is selected and a sample drawn from within the strata. To measure diversity, we introduce CoverageQA, a dataset of underspecified questions with multiple equally plausible answers, and assess diversity by measuring KL Divergence between the output distribution and uniform distribution over valid ground truth answers. As computing probability per response/solution for proprietary models is infeasible, we measure recall on ground truth solutions. Our evaluation show using SimpleStrat achieves higher recall by 0.05 compared to GPT-4o and 0.36 average reduction in KL Divergence compared to Llama 3.

1 INTRODUCTION.

Large language models (LLMs) are routinely resampled in order to get a wide set of plausible generations. Three key settings where this is important are: 1) improving downstream accuracy with planning or search for agentic tasks (i.e. Tree-of-thought (Yao et al., 2024), AgentQ (Putta et al., 2024)), 2) estimating prediction uncertainty (Aichberger et al., 2024), and 3) generating diverse datasets for post-training (Dubey et al., 2024) and fine-tuning (Dai et al., 2023). All these use cases rely on the model generating multiple plausible generations for the same prompt when multiple answers exists.

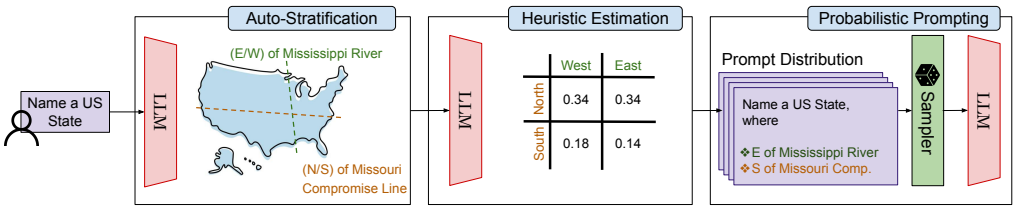

Figure 2: **SimpleStrat workflow.** SimpleStrat employs 3 phases: 1) auto-stratification to identify good dimensions of diversity that divide the solution space into equal partitions, 2) heuristic estimation to estimate the proportion of solutions in each stratum, and 3) probabilistic prompting where a concrete prompt is randomly sampled from the prompt distribution specified by the previous two phases. Critically, diverse resampling comes from both the random choice of prompt as well as the temperature of the LLM decoding.

Naively, increasing temperature, a parameter that controllably flattens an LLM's softmax, can improve an LLM's generation diversity. However, temperature introduces two problems. First, higher temperatures degrades generation quality. Recent evidence suggests removing temperature scaling is desirable for multi-step reasoning to reduce errors compounding (Zhang et al., 2024). This is especially critical in syntax sensitive settings like code generation where low temperatures (≤ 0.15) are often used. Second, controlling for temperature does not necessarily improve diversity in the answer space. In Figure 1, we illustrate increasing temperature doesn't lead to meaningful increase in diversity if the model is excessively confident and suffers from mode collapse. When asked to "Name a US State," the model heavily skews towards answering "California", high temperature only marginally softens the skew while surfacing incorrect answers and hurting instruction following.

Our goal is to improve diversity when resampling LLMs, even in cases of severe mode collapse in next-token probabilities without manual intervention. Our analysis reveals that GPT-4 assigns 87% of its logit weight to "California" when prompted to name a US state. This observed bias can be attributed to the worsening of calibration due to post-training as reported in the GPT-4 tech report (OpenAI et al., 2024). This stark bias mirrors human cognitive bias, exemplified by the blue-seven phenomenon—where individuals disproportionately select blue and seven when asked to choose a random color and number. To counteract similar biases in human populations, social scientists, particularly in political polling, employ stratified sampling techniques (Simpson, 1951; Howell, 1992; Morris, 2022). We propose adapting this method to address mode collapse in LLMs.

We propose SimpleStrat, a training-free sampling approach to increase diversity. SimpleStrat improves LLM generation diversity without degradation to generation quality while ensuring that an LLM's outputs are aligned with the true distribution of answers. SimpleStrat consist of three stages: auto-stratification, heuristic estimation, and probabilistic prompting. Even if a language model cannot generate diverse solutions, we find that it can be prompted to identify useful partitions of the solution space based on the user request. We call this process *auto-stratification*. In Fig. 1, SimpleStrat identifies two semantically significant strata from user request, "Name a US State": "(East/West) of the Mississippi River" and "(North/South) of the Missouri Compromise Line."

Next, the heuristic estimation computes the joint probabilities across all strata. Back to Fig. 1, SimpleStrat then outputs the probability for all four possible regions in US. Finally, SimpleStrat samples from the joint probability distribution to augment the original user prompt with the selected stratas. We note that this approach to diversity is orthogonal to increasing temperature and hence does not affect generation quality.

We evaluate SimpleStrat on underspecified questions, specifically questions that have more than one plausible answer. However, unlike ambiguous questions more widely, an answer to an underspecified question can be easily verified to be a valid without additional context. These questions capture settings where the user is indifferent to the particular answer as long as it's valid or in settings where we wish to resample to get a set of candidates solutions. We introduce CoverageQA, a benchmark of underspecified questions with on average 28.7 equally plausible answers.

We measure diversity by computing the Kullback-Leibler (KL) Divergence from the response distribution to a uniform distribution over all valid answers. By computing the response distribution using next-token probabilities, we show SimpleStrat samples from a less biased distribution. For proprietary