# Estimation of Language Model Generation Length under a Context Window of Fixed Size

Chirag Nagpal
Meta Superintelligence Labs (MSL)

September 2025

**Abstract**

Accurate estimation of the distribution of generation lengths for a language model is critical for various downstream tasks, including post-training and efficient inference infrastructure planning. Standard methods assume all generations are complete, but in practice, modern LLMs operate under a fixed context window size. This constraint frequently results in trajectories being terminated before an end-of-sequence token is generated, leading to censored observations.

I show that ignoring these truncated trajectories, or treating the maximum context length as the decode length, introduces a significant downward bias in the estimated length distribution. This phenomenon, well-known in statistics and epidemiology as right-censoring, necessitates a robust estimation technique. Through illustrative examples and a simulation on a large-scale, multi-turn conversational dataset, I demonstrate that estimators that account for right-censoring result in more accurate and unbiased estimate of the true distribution of decode lengths.

## 1 Introduction

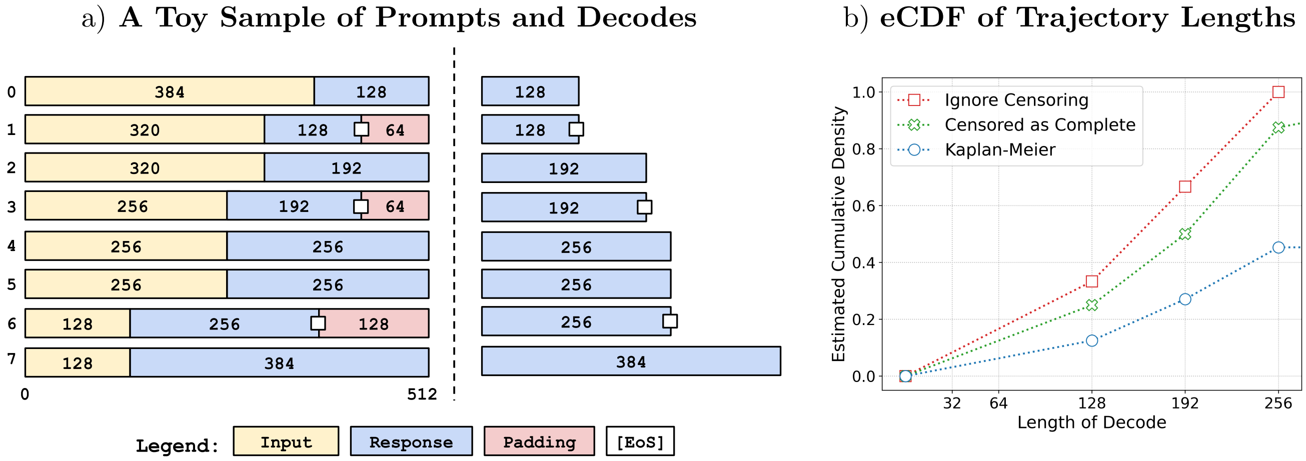
*Alignment* of modern Artificial Intelligence often involves making decisions in terms of the distribution of the length of generations from a large language model. In fact, when training AI models with modern techniques like on-policy Reinforcement Learning, the length of the generated sequences is an important machine tool that teachers like me use when making decisions about the underlying policy model Guo et al. (2025); Wu et al. (2025); Levy et al. (2024); Aggarwal and Welleck (2025); Jin et al. (2024). Furthermore, in efficient inference infrastructure planning, the expected and maximum generation lengths directly dictate memory allocation, latency, and throughput capacity. Hence, accurate estimation of the distribution of the length of generations can help capacity and infrastructure planning for inference when deployed.

In practice, when a large language model is subjected to inference under some distribution of input prompts cannot perform decoding indefinitely and is constrained by the available resources. Typically, therefore, decoding is limited to a fixed horizon called the *context window*.

The standard approach to estimating the distribution of the generation length from a language model involves a simple statistical aggregation of the generated trajectories and their decoded lengths. Here, the decoded length typically refers to the number of tokens that were generated by the language model during the decoding phase up to the `[EOS]` token. This is a special token that marks the end of the decoded trajectory from a language model. This approach assumes that all observed generation lengths represent the true, desired length of the model's output. However, this assumption breaks down in practical settings due to a critical engineering constraint, the fixed context window.

Modern Large Language Models operate under a maximum context size, $L_{\max}$ (for example, 1024 or 8192 tokens). If a generation's length, when combined with the input prompt, exceeds $L_{\max}$ the generation is abruptly terminated before the `[EOS]` token is produced. This results in an incomplete or truncated observation, where the true length of the trajectory is known only to be greater than the observed cutoff.

In statistics and epidemiology, this scenario is precisely defined as *right-censoring* (Aalen et al., 2008; Cox, 2018). The failure to account for right-censored observations is known to introduce a significant and systematic downward bias in the estimated length distribution. Two naive estimation strategies are commonly employed in practice, both of which are inadequate: (**1**) ignoring censored trajectories, which



| **Strategy 1** (Ignore Censoring) | **Strategy 2** (Censored as Complete) | **Kaplan-Meier Estimate** |
|---|---|---|
| $\widehat{\boldsymbol{P}}(L \leq 256) = \frac{\text{Complete}}{\text{Total}} = \frac{3}{3} = 1.0$ | $\widehat{\boldsymbol{P}}(L \leq 256) = \frac{\text{Complete}}{\text{Total}} = \frac{7}{8} = 0.875$ | $\widehat{\boldsymbol{P}}(L \leq 256) = 0.453$ |

Figure 1: **(a) Illustration of Truncated Generations and its effect of Decode Length**: **Yellow** corresponds to the input (prompt), **Blue** corresponds to the generation. and **Red** are padding tokens. Generations marked with □ are trajectories that end in the `[EOS]` token. **(b) The Empirical Cumulative Densities under various strategies**: Ignore Censoring (□), Censored as Complete (✕) and Censoring Adjusted (KM) (○)

severely limits the sample and biases toward shorter examples; and (**2**) treating the maximum context length as the true decode length, which systematically underestimates the true distribution.

For instance, in applications like long-form summarization, open-ended dialogue, or complex code generation, the model might naturally tend toward outputs longer than the context size. When these lengthy, yet truncated, generations are counted as having a length equal to the maximum context window $L_{\max}$ or worse, discarded entirely, the resulting statistics skew heavily toward shorter, unconstrained generations. To address this deficiency, we propose applying the Kaplan-Meier (KM) Product Limit Estimator, a powerful non-parametric technique to estimate language model generation lengths. We demonstrate that by correctly utilizing the available data (including the censored observations), the KM estimator provides a robust, nearly unbiased estimate of the true distribution of length. Our contributions are:

- We formally identify the problem of generation truncation in language model inference as a classic statistical right-censoring problem.
- Through illustrative examples and simulation, I demonstrate that common naive estimation strategies introduce a significant downward bias.
- We propose and apply the Kaplan-Meier estimator as a non-parametric solution, showing that it yields a substantially more accurate estimate of the generation length distribution.

## 2 Motivation: Fixed Context Length and the Problem of *Censoring*

Consider the case of RL post-training for reasoning, the length of responses that one can decode until when performing the online RL step are typically constrained subject to and compute restrictions and available infrastructure. Thus, in practice when drawing generations from an LLM, not all trajectories are decoded till completion. As soon as a trajectory reaches the end of its context window, decoding is terminated. Consider the example in **Figure** 1 (**a**), we have a sample (or *batch*) of prompts that is restricted to the *'Context Window'* length of 512 tokens. This context window is typically shared between the input prompt tokens (or *prefix*), the generated responses (or *decodes*) and the padding tokens. In this illustration, trajectories $\{1, 3, 6\}$ are completed with the `[EOS]` token denoting the completion of the decode.

On the other hand trajectories $\{0, 2, 4, 5, 7\}$ decoded for a few steps till the generations reached the maximum context window size ($L_{\max}$) of 512, but the trajectories were left truncated without completion. In statistical terms, particularly in the context of density estimation, this phenomenon where the value of a measurement or observation is only partially known or incomplete is known as ***censoring***. In **Figure** 1 (**b**), I present the Empirical Cumulative Densities estimated under the censored data. We have a couple of options to estimate length distribution under such a dataset:

**Strategy 1: Ignore Censoring**: One obvious option is to completely ignore trajectories that have not seen the [EOS] token. This has two fundamental problems. (**i**) It reduces the effective sample size (in this case from 8 → 4) preventing robust estimation of the length distribution. (**ii**) Typically censored trajectories are longer thus ignoring them leads to an under-estimation of the trajectory length distribution.

**Strategy 2: Consider Censored Lengths as Decode Lengths**: Another obvious option is to consider the trajectory length till the end of the context window as its complete length. Since the trajectories are in-practice longer than their corresponding censored lengths considering the truncated length as the corresponding length also leads to a significant under-estimation of trajectory length.

This suggests that both the options above are inadequate for accurate estimation of the length distribution. The estimation must still correctly incorporate the partial data points while still avoiding significant bias in the corresponding estimate.

**The Empirical Distribution Function** Consider the simplest case of estimation of the generation length from a large language model. Let us consider a sample of $n$ input prompts and the corresponding length of the decodes $\{(\boldsymbol{x}_i, \ell_i)_{i=1}^n\}$

$$\begin{aligned}\widehat{\boldsymbol{F}}_n(L) &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{\ell_i \leq L\} \\ &= \frac{\text{number of trajectories in the sample} \leq L}{n}\end{aligned}$$

What this estimator essentially says is that probability of the length of responses to be under a certain length can be estimated by number of trajectories under that length, divided by the total number of trajectories in the sample. This simple estimator is known to be optimal owing to the well known Glivenko-Cantelli theorem (Tucker, 1959).

**The KM Estimator for Density Estimation** The Kaplan-Meier (KM) estimator (Kaplan and Meier, 1958) is a non-parametric statistic used to estimate a density function such as that of lengths, which is the probability that a trajectory is smaller than a certain length $\ell$. The KM estimator (also called a product limit estimator) handles censored data (incomplete observations) by assuming that censoring occurs independently of trajectory completion. The result is a step function that changes value only at observed trajectory lengths.

$$\widehat{\widehat{\boldsymbol{F}_n}}(L) = \mathbf{1} - \prod_{i:\ell_i \leq L} \left(1 - \frac{d_i}{n_i}\right) \tag{1}$$

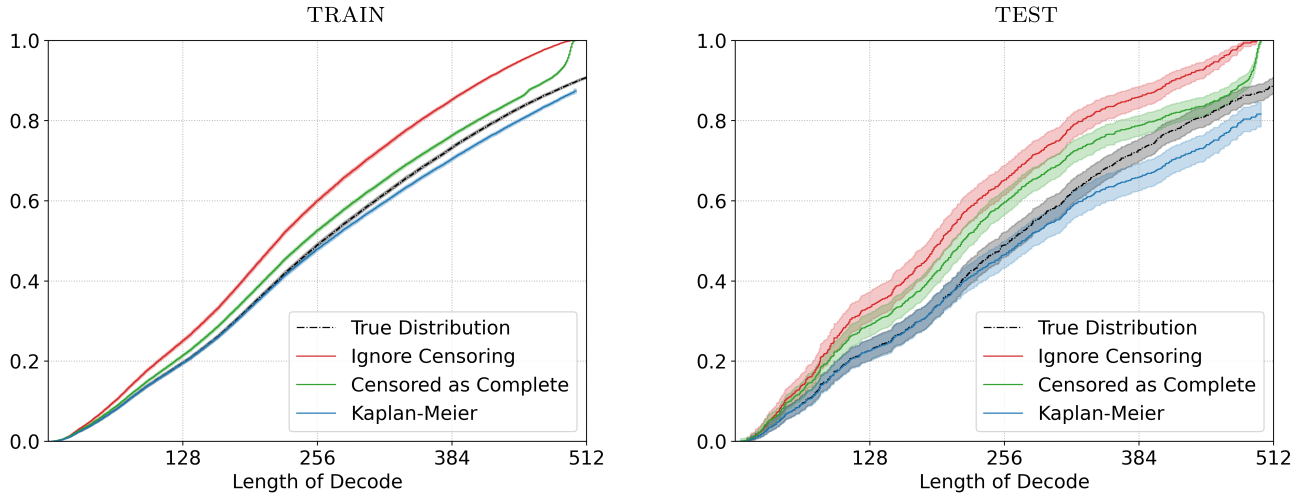
This represents the cumulative distribution function (CDF) for the trajectory length data, where $d_i$ is the number of trajectories that were completely decoding with a length $\ell_i$ and $n_i$ is the number of trajectories yet to finish decoding till completion.

Additionally, summary statistics such as *mean* and *median* length of generations can be obtained relatively easily from the full estimated cumulative density $\widehat{\boldsymbol{F}_n}$. Specifically the median is the length at which estimated cumulative density attains the value of $\mathbf{1/2}$. On the other hand, the mean (or *expected*) generation length can be obtained using the following straightforward formula from elementary statistics:

$$\mathbf{E}[L] = \int_0^\infty \left(1 - \widehat{\boldsymbol{F}}_n(L)\right) \mathrm{d}L, \tag{2}$$

which corresponds to the area under the Kaplan-Meier survival function of the distribution.

## 3 Simulation



| **Strategy** | TRAIN SPLIT | | | TEST SPLIT | | |
|---|---|---|---|---|---|---|
| | True Value | Estimate | Error | True Value | Estimate | Error |
| Censored as Complete | 0.491 | 0.527 | 0.0359 | 0.491 | 0.598 | 0.1066 |
| Ignore Censoring | 0.491 | 0.601 | 0.1103 | 0.491 | 0.653 | 0.1616 |
| Kaplan-Meier | 0.491 | 0.480 | 0.0106 | 0.491 | 0.466 | 0.0247 |

Figure 2: **The empirical cumulative densities under different estimation strategies:** Assuming the Censored Length to be the True Length, as well as Ignoring the Censored Trajectories leads to significant underestimation of Trajectory Length. Error is the $L_1$ distance in estimation of the cumulative density at length of 256 tokens. Adjusting for censoring with a Kaplan-Meier curve result in a more accurate estimate.

In this section, I present a simulation to demonstrate the efficacy of the censoring adjusted approach for estimating generation length distribution using the popular Helpfulness dataset (Bai et al., 2022) involving chatbot style multi-turn conversations. For the purposes of simulation, the last turn of the assistant response is considered as the decode while the rest is considered to be the input prompt. I restrict the context window to be 512 tokens and only consider input prompts with a length less than the total context window. For prompts that have a total length of over 512 tokens we consider the decoded trajectories to be *censored*, while for the others we consider the decodes to be *complete*. Thus by experimental design,

$$\text{Is Trajectory Complete} = \mathbf{1}\{\text{True Decode Length} + \text{Input Prompt} \leq 512\}$$
$$\text{Censored Decode Length} = 512 - \text{Input Prompt Length}$$

Figure 3 presents the estimated cumulative densities from the different strategies as well as the Kaplan-Meier approach on the Train and Test splits of the Helpfulness Dataset. We evaluate the estimated $\mathbf{P}(L < 256)$ and compare it to the true value in the corresponding table in terms of squared error. Notice that the KM approach consistently gives more accurate predictions as evidenced quantitatively as well as in from visual inspection in the Figure.

## 4 Conclusion

In this paper, I present the practical problem of estimating trajectory lengths of generations drawn from a language model under context size constraints. In practice, accurate estimation is challenging and requires accounting for observations that are *censored*. Through an illustrative example and a simulation, I show how the Kaplan-Meier Product Limit estimator is a better alternative for estimation of the distribution of trajectory lengths. This and other recent papers (Nagpal, 2025; Davidov et al., 2025), demonstrate how classic statistical thinking can be useful for the new science of *AI Alignment*.

## References

Aalen, O. O., Borgan, Ø., and Gjessing, H. K. (2008). *Survival and event history analysis: a process point of view*. Springer.

Aggarwal, P. and Welleck, S. (2025). L1: Controlling how long a reasoning model thinks with reinforcement learning. *arXiv preprint arXiv:2503.04697*.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Cox, D. R. (2018). *Analysis of survival data*. Chapman and Hall/CRC.

Davidov, H., Freidkin, G., Feldman, S., and Romano, Y. (2025). Calibrated predictive lower bounds on time-to-unsafe-sampling in llms. *arXiv preprint arXiv:2506.13593*.

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

Jin, M., Yu, Q., Shu, D., Zhao, H., Hua, W., Meng, Y., Zhang, Y., and Du, M. (2024). The impact of reasoning step length on large language models. *arXiv preprint arXiv:2401.04925*.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481.

Levy, M., Jacoby, A., and Goldberg, Y. (2024). Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*.

Nagpal, C. (2025). Preference models assume proportional hazards of utilities. *arXiv preprint arXiv:2508.13189*.

Tucker, H. G. (1959). A generalization of the glivenko-cantelli theorem. *The Annals of Mathematical Statistics*, 30(3):828–830.

Wu, Y., Wang, Y., Ye, Z., Du, T., Jegelka, S., and Wang, Y. (2025). When more is less: Understanding chain-of-thought length in llms. *arXiv preprint arXiv:2502.07266*.