What Makes a Transformer Work?
A Controlled Ablation Study of Causal Language Modeling

Introduction

Transformer language models are often treated as monolithic architectures, yet their success rests on a small number of interacting structural constraints. This report presents a controlled ablation study of a minimal decoder-only Transformer, aimed at separating *problem-defining components* from *representation refinements* and *optimization stabilizers* in causal language modeling.

Rather than maximizing absolute performance, the goal is mechanistic clarity: to identify **which components define the probabilistic modeling problem itself, and which primarily enable effective training**.

Data and Evaluation

All experiments use the **Tiny Shakespeare** corpus with a fixed train/validation split and **character-level** tokenization. The task is next-character prediction, exposing both short-range statistical structure (e.g. spelling and punctuation) and longer-range dependencies (e.g. syntax and simple discourse patterns).

Models are evaluated using negative log-likelihood (NLL), bits-per-character (BPC), perplexity (PPL), and top-1 accuracy. For a sequence $(x_1, \ldots, x_T)$, perplexity is defined as

$$\mathrm{PPL} = \exp\left(-\frac{1}{T}\sum_{t=1}^{T}\log p(x_t \mid x_{<t})\right),$$

i.e. the exponential of the average NLL under the autoregressive model. These likelihood-based metrics quantify local next-token uncertainty, but can be misleading when architectural constraints of the modeling objective are violated.

To probe generation quality beyond likelihood, we additionally measure *generation diversity* using **distinct-3**, the fraction of unique character trigrams in generated samples. This metric is sensitive to repetition and mode collapse, and serves as a coarse indicator of whether a model produces varied, non-degenerate text rather than memorized or trivially predictable sequences.

Experimental Design

We train a minimal GPT-style decoder-only Transformer with identical hyperparameters across all runs. Each experiment performs a *single-component ablation*, removing or altering exactly one architectural element while keeping the remaining structure, parameterization, and training procedure fixed.

This design isolates the *functional role* of each component: performance differences can be attributed directly to the removed mechanism, rather than confounded by capacity or optimization effects. All models share the same optimizer, learning-rate schedule, context length, batching, and number of update steps. Single-seed runs suffice here, as the goal is to identify robust qualitative effects.

Ablations studied. No positional encoding; no attention (replaced by a per-token linear mapping); no MLP; no residuals and LayerNorm; and no causal mask.

The reference architecture is shown on the following page (Figure 1) to preserve vector-quality rendering.

Model

The model uses autoregressive factorization

$$p(x_1, \ldots, x_T) = \prod_{t=1}^{T} p(x_t \mid x_{<t}),$$

where each conditional distribution is parameterized by a Transformer operating on the prefix $(x_1, \ldots, x_{t-1})$. Causal masking enforces this structure by preventing attention to future tokens, ensuring that likelihood corresponds to a valid generative process.

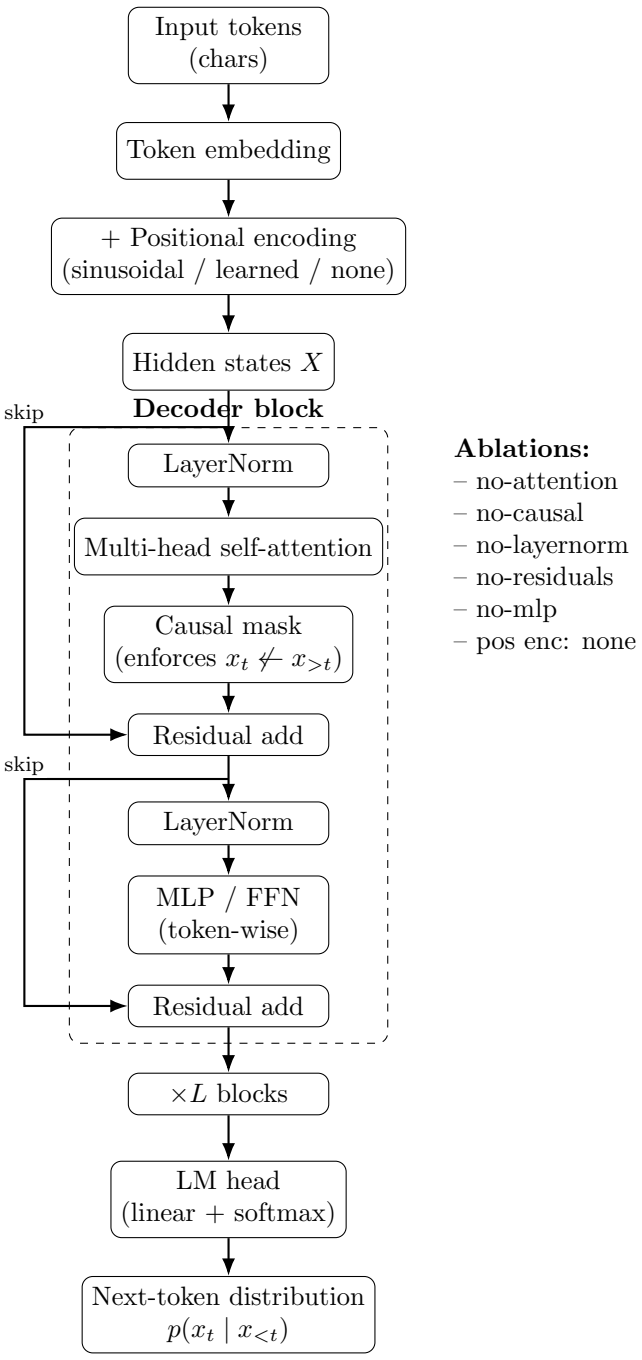

Figure 1: Decoder-only Transformer (GPT-style) used in this study. The model implements the autoregressive factorization $p(x_1, \ldots, x_T) = \prod_t p(x_t \mid x_{<t})$ via masked self-attention. Causal masking prevents access to future tokens and is essential for a valid generative interpretation of likelihood-based metrics. Self-attention provides content-dependent mixing across positions and is the primary mechanism for modeling long-range dependencies. Positional encodings inject sequence order, breaking the permutation invariance of attention. The MLP refines token-wise representations through nonlinear feature mixing, while residual connections and LayerNorm stabilize optimization by preserving gradient flow and controlling activation scale.

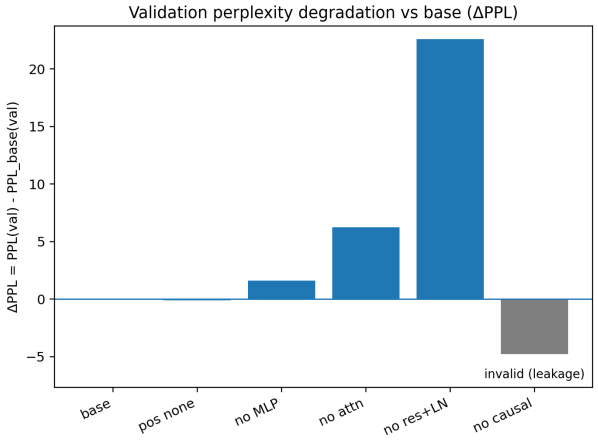

Figure 2: Validation perplexity degradation under single-component ablations. Positive values indicate genuine performance loss; the negative bar for the no-causal model reflects objective invalidation due to future-token leakage.

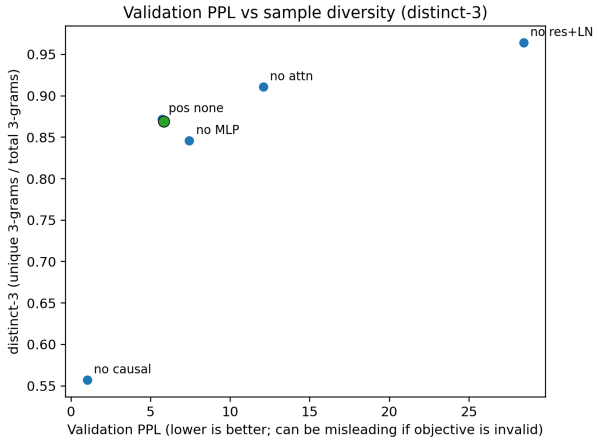

Figure 3: Validation perplexity vs. generation diversity (distinct-3). The base model is shown in green, with ablated models in blue. Low perplexity does not guarantee meaningful generation: the no-causal model achieves near-perfect likelihood yet produces degenerate text.

Variation	BPC	PPL	Top-1
base	2.541	5.823	0.474
positional	2.513	5.710	0.486
MLP	2.893	7.432	0.403
attention	3.593	12.075	0.268
res + no LN	4.829	28.435	0.149
causal mask	0.029	1.020	0.994

Table 1: Validation-set metrics. The causal row is *not comparable* due to future-token leakage.

Results

Figure 2 and Table 1 summarize the primary quantitative findings. The most severe degradation occurs when removing **residual connections and LayerNorm**, leading to a dramatic increase in perplexity and a collapse in accuracy. This failure reflects *optimization breakdown* rather than a loss of representational capacity, highlighting the central role of these components in enabling stable training.

Removing **attention** also produces a substantial performance drop, confirming its importance as the primary mechanism for sequence modeling. By contrast, removing the **MLP** degrades performance more moderately, consistent with its role as a representational refinement rather than a core time-mixing mechanism.

Removing **positional encodings** yields only a small degradation in this setting. This reflects the weak positional demands of character-level modeling on a small corpus, rather than the irrelevance of sequence order as a modeling assumption. Finally, removing the **causal mask** yields artificially low perplexity by allowing future-token access, invalidating the autoregressive objective.

Likelihood-based metrics are therefore only meaningful when architectural constraints enforce the intended probabilistic structure; when causality is violated, perplexity ceases to reflect generative competence.

Figure 3 reinforces this point by contrasting likelihood with generation diversity. Models that violate causal structure achieve deceptively good perplexity while exhibiting highly repetitive or degenerate outputs, demonstrating that likelihood alone is insufficient to assess generative quality when architectural assumptions are broken.

Discussion

These results disentangle architectural roles that are often conflated in practice. **Causal masking** is indispensable for defining a valid autoregressive objective: removing it invalidates likelihood as a modeling criterion. **Attention** remains the primary mechanism for modeling sequence dependencies, and its removal substantially restricts expressivity. However, the most severe empirical failures arise from removing **residual connections and LayerNorm**, which do not define the task itself but are essential for optimization. Their absence leads to training collapse despite retained representational capac-

ity, making optimization stabilizers the dominant practical bottleneck in this setting.

By contrast, removing **positional encodings** causes only a modest change in validation perplexity. This should not be interpreted as evidence that sequence order is unimportant, but rather as a limitation of both the dataset and the metric. At the character level, strong local statistical regularities allow models to infer relative order implicitly, masking positional deficiencies under next-token likelihood. Perplexity therefore provides a weak probe of positional reasoning in this regime, as it emphasizes local predictability rather than sequence-level coherence or generalization.

Limitations. The analysis is restricted to a small, character-level corpus with short effective context lengths, which under-stresses explicit positional reasoning and long-range dependency modeling. In such settings, positional encodings primarily affect robustness, extrapolation, and global structure—properties not well captured by in-distribution perplexity. Datasets or tasks requiring explicit index-based reasoning, long-context generalization, or weak local cues (e.g., synthetic position queries or length extrapolation tests) would more directly expose positional encodings as a problem-defining architectural component. While the qualitative distinction between objective-defining, expressive, and optimization-critical components is robust, the quantitative effects reported here should be interpreted as task- and metric-dependent.