International Journal of Applied and Advanced Scientific Research (IJAASR)
International Peer Reviewed - Refereed Research Journal, Website: www.dvpublication.com
Impact Factor: 5.655, ISSN (Online): 2456 - 3080, Volume 11, Issue 2, July - December, 2026

RECURSIVE SYNTHETIC TRAINING CONDITIONS AND LANGUAGE MODEL PERFORMANCE DEGRADATION ACROSS PUBLISHED TEXT GENERATION BENCHMARK EXPERIMENTS

Divya Vishwanath* & Paul Johnson Asamoah**
* Janardhan Rai Nagar Rajasthan Vidyapeeth, Udaipur, Rajasthan, India
** Power Life University College, Accra, Ghana, West Africa

Cite This Article: Divya Vishwanath & Paul Johnson Asamoah, "Recursive Synthetic Training Conditions and Language Model Performance Degradation Across Published Text Generation Benchmark Experiments", International Journal of Applied and Advanced Scientific Research, Volume 11, Issue 2, July - December, Page Number 23-30, 2026.
Copy Right: © DV Publication, 2026 (All Rights Reserved). This is an Open Access Article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.
Type of Review: Peer Reviewed as per |C|O|P|E| Guidance.
Disclaimer: The scholarly papers reviewed and published by DV Publication, India, reflect the views and opinions of their respective authors and do not necessarily represent the views or opinions of DV Publication. The publisher disclaims any responsibility for any harm, loss, or damage resulting from the use of the published content by any party.
DOI: https://doi.org/10.5281/zenodo.24039710

Abstract:

This analysis examines whether assessments of language model resilience depend on decoding configuration and measurement scale, addressing international debates on recursive synthetic training. A descriptive reanalysis uses 24 published WikiText 2 perplexity endpoints from 2025, forming 12 matched configurations across GPT 2, SmolLM2, and six decoding strategies at generations zero and nine. Paired absolute changes, percentage changes, ratios, means, and medians reveal deterioration in every configuration, with percentage increases ranging from 10.24% to 513.83%. SmolLM2 records a smaller mean absolute increase than GPT 2, but a larger mean percentage increase, while its median deterioration remains lower on both scales. Different baseline denominators and concentrated deterioration under greedy decoding and beam search explain this aggregation reversal. Sampling strategy rankings also differ between models, and temperature sampling reverses their comparison across measurement scales. These findings qualify interpretations of resilience as an intrinsic model property by demonstrating configuration and measurement dependence. International benchmarking should report matched baselines, absolute and proportional changes, and distributional summaries to support transparent configuration selection.

Key Words: Synthetic Training Data; Model Collapse; Language Model Evaluation; Decoding Strategies; Descriptive Reanalysis

1. Introduction:

A language model can appear more resilient under one measure of deterioration and less resilient under another. This ambiguity matters when developers compare training configurations that reuse generated text, because starting performance, endpoint performance, and proportional change represent different aspects of predictive reliability. Recursive synthetic training connects model outputs to subsequent learning, allowing errors and distributional omissions to enter later training material. Shumailov et al., 2024, demonstrate that repeated learning from generated samples can progressively distort the original distribution, including the loss of uncommon events. Their findings establish a substantive risk, but assessing its severity requires attention to how deterioration is measured and compared. The consequences of synthetic training depend on the process through which generated material enters learning. Kazdan et al., 2025, distinguish replacement of original data from accumulation of real and synthetic data. Their experiments show that these workflows can produce different performance trajectories, making the organisation of training data central to interpreting collapse. This distinction limits the usefulness of treating synthetic content as a uniform exposure. It also directs attention towards comparisons within clearly specified training conditions, where differences in performance can be interpreted against a shared experimental setting.

Within a recursive workflow, decoding strategy determines which outputs become subsequent training examples. Greedy decoding and beam search favour probable sequences, while sampling strategies retain different portions of the predictive distribution. Concentrating selection may narrow the material available for subsequent learning, whereas broader sampling may preserve variation while admitting less reliable continuations. These mechanisms create competing expectations rather than a universal ordering of decoding methods. Wiher et al., 2022, show that relationships between decoding, diversity, and generation quality depend on the task. Consequently, a strategy that performs favourably in one setting cannot be assumed to preserve performance across other model and training configurations. Drayson et al., 2025, provide the closest experimental evidence by examining decoding strategies during recursive language model training and evaluating an approach to mitigating collapse. Their results establish that deterioration varies across decoding methods and model specifications. Those experiments therefore supply both the empirical foundation and the boundary of the present analysis. Rather than presenting the reported decoding differences as a new discovery, this analysis reorganises their published endpoints into matched comparisons that distinguish absolute deterioration, proportional deterioration, and the distribution of changes across configurations. The specific question is whether comparative assessments of model resilience remain consistent when the same observations are expressed on these alternative scales.

Outcome definition is central to that question. Test perplexity measures predictive performance against evaluation text, with higher values indicating poorer prediction. It does not provide a comprehensive measure of generated text quality. Meister et al., 2023, demonstrate that performance under standard model metrics can diverge from perceived coherence and fluency. Even within the narrower domain of predictive performance, endpoint levels and changes from baseline answer different questions. A model may retain lower endpoint perplexity while experiencing a larger proportional deterioration because it began from a stronger baseline. Likewise, a mean change may be dominated by a small number of configurations, while the median describes a different central tendency. These distinctions make measurement and aggregation substantive components of resilience

assessment. This study aims to assess how recursive endpoint deterioration varies across decoding strategies and model specifications, and whether those comparisons depend on measurement scale and aggregation. The analysis uses 24 published perplexity endpoints forming 12 matched configurations for GPT 2 and SmolLM2 under six decoding strategies. Each configuration contains measurements at generations zero and nine on the English WikiText 2 benchmark. The publication falls within the scope of 2010 to 2025; recursive generations identify experimental sequence rather than calendar years. Descriptive paired calculations quantify absolute increases, percentage increases, and endpoint ratios. The objectives are to establish the direction and magnitude of change within configurations, compare decoding strategies within each model, and compare models under matched decoding conditions.

The findings show deterioration in every observed configuration but no uniform ranking of comparative resilience. SmolLM2 records a smaller mean absolute increase than GPT 2 while recording a larger mean percentage increase. Its median deterioration is lower on both scales. Temperature sampling also produces a model comparison that reverses between absolute and proportional change. These results identify a bounded contribution: assessments based on a single scale or aggregate statistic can conceal differences that remain visible in matched configuration comparisons. The analysis qualifies interpretations of recursive degradation by separating its occurrence from its severity and from the construction of comparative rankings. The practical relevance concerns how developers and evaluators report and select recursive training configurations. Reporting baseline and endpoint perplexity alongside absolute and proportional changes makes the basis of comparison explicit. Presenting means together with medians reveals whether large changes in particular configurations dominate an aggregate assessment. These reporting principles are relevant to international benchmarking and model evaluation, while the empirical rankings remain specific to the two models, six decoding strategies, and benchmark examined here. The design supports transparent descriptive comparisons of published results rather than causal attribution or general claims about all synthetic training systems.

2. Theory Conceptual Framework and Research Questions Development:

Recursive synthetic training connects language generation and model learning through a feedback process. A model produces text that becomes training material for a subsequent generation, allowing its existing approximations to influence later learning. Shumailov et al., 2024, explain that repeated learning from generated samples can progressively distort the original distribution, particularly when uncommon events disappear from the sampled data. This mechanism provides the theoretical foundation for examining performance degradation across recursive generations. It suggests a risk of cumulative information loss rather than establishing that every synthetic training configuration must deteriorate equally. This analysis defines recursive synthetic training conditions through three observed dimensions: recursive generation, decoding strategy, and model specification. Language model performance is measured by test perplexity, with higher values indicating poorer prediction of the evaluation text. The analytical sample contains 24 reported endpoints forming 12 matched configurations across two models and six decoding strategies. Generation zero provides the human data training baseline, while generation nine provides the recursive endpoint. These observations support comparisons of performance changes within configurations and differences between configurations. Recursive generation represents repeated exposure to the outputs of preceding models. Each transition can reproduce existing approximation errors while introducing additional sampling variation. The theoretical expectation is therefore that repeated replacement of human continuations with generated continuations may weaken predictive performance against a fixed human text benchmark. The available measurements establish the beginning and endpoint of this process. Accordingly, the analysis measures absolute perplexity changes and proportional changes relative to each configuration's baseline. Research question 1 asks: How does test perplexity change between generation zero and generation nine within each model and decoding configuration?

Decoding strategy determines which parts of a model's probability distribution become training examples. Greedy decoding and beam search favour probable sequences, potentially concentrating subsequent learning on a narrower set of outputs. Sampling methods retain more variation but can also introduce less reliable continuations. These competing mechanisms prevent a general assumption that greater sampling variability always improves recursive training. Meister et al., 2022, explain that highly probable text does not necessarily provide the information content associated with natural language quality. Their argument supports treating decoding as a substantive feature of the training process rather than a neutral implementation choice. The consequences of decoding also depend on the task. Wiher et al., 2022, demonstrate that relationships between decoding, diversity, and text quality vary across language generation settings. This evidence supports comparing decoding strategies within the same benchmark and model rather than imposing a universal ranking. Here, greedy decoding, beam search, pure sampling, temperature sampling, top k sampling, and nucleus sampling are compared through their baseline adjusted perplexity changes. Research question 2 asks: How do changes in test perplexity differ across decoding strategies within each model specification?

Model specification shapes the distribution from which synthetic continuations are generated and the learning process through which those continuations are absorbed. Drayson et al., 2025, report differences in recursive degradation across GPT 2 and SmolLM2 under alternative decoding strategies. This analysis therefore treats model identity as an observed configuration encompassing architecture, pretraining, and parameter count. It compares proportional changes under matched decoding strategies, allowing for different starting perplexities. Research question 3 asks: How do changes in test perplexity differ between GPT 2 and SmolLM2 under the same decoding strategy?

The conceptual framework links the three observed training dimensions to predictive performance through the generated examples available for subsequent learning. Generation identifies cumulative recursive exposure, decoding determines how continuations are selected, and model specification identifies the system producing and learning from those continuations. These relationships organize descriptive comparisons rather than separate causal estimates. The outcome remains predictive performance on the specified benchmark. Meister et al., 2023, show that standard model metrics and perceived generation quality can diverge, reinforcing the need to interpret perplexity within its measurement boundaries. The three research questions consequently concern measured performance changes and configuration differences, without extending those comparisons to unmeasured capabilities.

3. Data:

This analysis uses published experimental endpoints to document data provenance, construct matched comparisons, and measure language model performance degradation.The data support descriptive comparisons across recursive generations, decoding strategies, and model specifications.

3.1 Data Sources, Population, and Sample Construction:

The numerical source is Table 1 of Drayson et al., 2025, published by the Association for Computational Linguistics. The analysis uses the published proceedings version available through ACL Anthology under the Creative Commons Attribution 4.0 licence. The consolidated workbook records its administrative extraction date separately from the publication year. The source does not provide calendar dates for individual training runs. The empirical setting is English text continuation using WikiText 2. The unit of observation is a model and decoding configuration evaluated at a recursive generation endpoint. Two model specifications, GPT 2 with 124 million parameters and SmolLM2 with 360 million parameters, are observed under six decoding strategies at generations zero and nine. The resulting structure contains 24 endpoints grouped into 12 matched configurations. Generation denotes training sequence rather than calendar time. Consequently, these records constitute balanced paired experimental endpoints rather than an annual panel. The target population comprises language model configurations exposed to recursive synthetic training. The accessible sampling frame is narrower and consists of the configurations reported in the source table. Selection retains every reported model perplexity value at the two endpoints. The human reference row supplies no perplexity values and contributes no analytical observations. The temporal scope of 2010 to 2025 defines publication eligibility, while the extracted evidence comes from a publication dated 2025. Coverage is defined by language and benchmark rather than country or economic sector.

Table 1: Sample construction

Stage	Units	Observations	Reason or Rule
Source table screened	12 model configurations and 1 human reference row	24 numerical perplexity endpoints	Inspect reported perplexity at generations zero and nine
Human reference excluded	1 reference row	0 usable perplexity endpoints	Perplexity is unavailable for the human reference
Eligible model sample	2 models and 12 configurations	24 endpoints	Retain all reported model perplexity values
Identifier and completeness checks	12 configurations	24 endpoints retained	No duplicate analytical keys or missing required values detected
Matched comparison sample	12 configurations	12 endpoint pairs	Link each generation nine observation to its corresponding baseline

The selection rules describe the implemented extraction rather than a prospectively registered sampling protocol. No exclusions were introduced after examining performance changes. Both models retain all six decoding strategies and both endpoints. No observations are lost through controls, lag construction, differencing, or estimator requirements.

3.2 Data Integration, Cleaning, and Quality Assurance:

The workbook preserves reported measurements separately from calculated comparisons. Each endpoint has a unique record identifier, while a configuration identifier links observations sharing the same model and decoding strategy. Pairing follows a one-to-one relationship between generation zero and generation nine within each configuration. Checks identified 24 unique record identifiers and 24 unique combinations of model, decoding strategy, and generation. All 12 configurations have complete endpoint pairs. All required fields are populated. Perplexity values are positive and finite, and all baseline denominators are strictly positive. Reported numerical precision is retained at two decimal places. Repeated baseline values across decoding configurations are preserved because they represent distinct reported configurations rather than duplicate analytical records. No external datasets were merged, and no unmatched records arose. The extraction retains the source's model labels, decoding categories, generation identifiers, and perplexity values. No imputation, interpolation, zero replacement, trimming, or winsorisation was applied. Large perplexity values remain in the sample because they are reported experimental outcomes. Intermediate generations are outside the numerical extraction and are not reconstructed from figures. The audit preserves the distinction between source measurements and derived quantities, consistent with the traceability principles discussed by Sandve et al., 2013. Checks establish internal completeness and consistency; they do not constitute an independent replication of model training.

3.3 Variable Measurement, Construction, and Validation:

The outcome is testing perplexity on WikiText 2. Higher perplexity indicates poorer prediction of the evaluation text. Recursive generation, decoding strategy, and model specification identify the observed training conditions. Decoding remains categorical because its methods do not form a common numerical scale. Wiher et al., 2022, show that decoding properties depend on the generation task, supporting comparisons within a shared benchmark.

Table 2: Variable dictionary

Variable or Indicator	Analytical Role	Definition and Construction	Source and Identifier	Unit and Frequency	Transformation and Direction	Missing Value Rule	Coverage and Limitation
Test perplexity	Outcome	Reported predictive performance on the evaluation text	ACL published version, Table 1, perplexity columns	Dimensionless score at each endpoint	Original level; higher means poorer performance	All 24 values present	Measures benchmark prediction
Recursive generation	Training dimension	Position in the recursive training sequence	Table 1, generation columns	Generation zero or nine	Retained as endpoint categories	Both endpoints required	Does not identify intermediate trajectories
Decoding strategy	Training dimension	Greedy, beam search, pure sampling,	Table 1, decoding rows	Nominal category per configuration	No ordinal transformation	All categories present	Six strategies within each model

Variable or Indicator	Analytical Role	Definition and Construction	Source and Identifier	Unit and Frequency	Transformation and Direction	Missing Value Rule	Coverage and Limitation
		temperature, top k, or nucleus					
Model specification	Training dimension	GPT 2 or SmolLM2	Table 1, model groups	Nominal category	Matched comparisons by decoding strategy	Both models retained	Combines architectural and pretraining differences
Parameter count	Model descriptor	Reported model size of 124 or 360 million parameters	Source Section 5 and Appendix A	Millions of parameters	Original scale	Complete	Not separable from model identity
Absolute perplexity change	Derived outcome contrast	Generation nine value minus baseline value	Workbook paired comparisons	Perplexity units per pair	Positive values indicate deterioration	Complete positive baseline required	Twelve matched configurations
Proportional perplexity change	Derived outcome contrast	Endpoint divided by baseline, minus one	Workbook paired comparisons	Fraction, displayed as a percentage	Positive values indicate deterioration	Complete positive baseline required	Sensitive to baseline magnitude
Perplexity ratio	Alternative descriptive scale	Endpoint divided by baseline	Workbook paired comparisons	Dimensionless ratio	Values above one indicate deterioration	Complete positive baseline required	Measures proportional endpoint difference

For configuration c, Equation 1 defines absolute change:

$$\Delta P_c = P_{c,9} - P_{c,0}$$

Here, ΔP_c denotes absolute perplexity change, $P_{c,9}$denotes generation nine perplexity, and $P_{c,0}$denotes baseline perplexity. Equation 2 defines the endpoint ratio:

$$R_c = \frac{P_{c,9}}{P_{c,0}}$$

Here, R_c denotes the perplexity ratio for configuration c, using the endpoint and baseline values defined above. Equation 3 converts the ratio into percentage change:

$$G_c = 100 \times [R_c - 1]$$

Here, G_c denotes percentage change in perplexity and $R_c$denotes the endpoint ratio. This measure expresses percentage change rather than percentage point change.

Table 3: Descriptive statistics

Measure	Valid Count	Mean	Standard Deviation	Median	Minimum	Maximum
Perplexity across all endpoints	24	36.42	24.56	29.27	13.96	85.69
Baseline perplexity	12	21.62	8.00	21.60	13.96	29.31
Generation nine perplexity	12	51.22	26.81	48.66	15.39	85.69
Absolute perplexity change	12	29.61	25.15	19.37	1.43	71.73
Percentage perplexity change	12	154.10	173.44	72.72	10.24	513.83
Perplexity ratio	12	2.54	1.73	1.73	1.10	6.14

Standard deviations describe variation across reported configurations and are not standard errors across independent training runs. The three calculated contrasts express the same endpoint comparison on different scales. They are not separate dependent constructs. Interpretation remains specific to predictive performance because benchmark perplexity and perceived generation quality need not coincide, as Meister et al., 2023, demonstrate.

3.4 Analytical Framework, Identification, and Reproducibility:

The analysis is descriptive. Its quantities of interest are the absolute and proportional changes between paired endpoints, differences across decoding strategies within each model, and differences between model specifications under matched decoding strategies. Equations 1 to 3 implement these comparisons using the same 12 configurations. Baseline adjustment accounts for different starting perplexities without treating those differences as independently identified explanatory effects. Within each model, decoding comparisons use six complete configurations. Between models, comparisons match the decoding strategy and retain each model's own baseline. Absolute changes describe deterioration in perplexity units, while proportional changes express deterioration relative to the starting value. These complementary scales address baseline dependence without changing sample membership.

The available data contain aggregate endpoints rather than independent training repetitions. The analysis therefore reports magnitudes and distributions without significance tests or causal coefficients. Model comparisons describe combined specification differences and do not isolate the contribution of parameter count. Two observed generations establish endpoint differences but do not identify the shape or speed of degradation between them. Reproducibility rests on preserving the published values, source identifier, configuration keys, variable definitions, and spreadsheet formulas. This separation allows the numerical comparisons to be reconstructed without claiming replication of the original training experiments. Peng, 2011, distinguishes computational reproducibility from independent replication, which defines the appropriate scope of the supporting workbook. Consistent pairing, positive denominators, unchanged source precision, and a common comparison sample align measurement with the descriptive framework.

4. Methods:

4.1 Research Design and Comparison Strategy:

This analysis uses a descriptive reanalysis of published experimental endpoints to examine performance deterioration within configurations, differences across decoding strategies, and differences between model specifications. The observational unit is a model and decoding configuration evaluated at a recursive generation endpoint. The analytical unit for measuring deterioration is the matched pair linking generation zero with generation nine. Table 1 documents sample construction, while Table 2 defines the measurements used in these comparisons. The comparison strategy holds model specification and decoding strategy constant when calculating endpoint changes. Comparisons across decoding strategies are conducted separately within each model. Comparisons between models match the decoding strategy and express deterioration relative to each model's own baseline. This structure separates within configuration change from differences between configurations without estimating adjusted causal effects. The analytical sample comprises 12 complete configurations represented by 24 endpoint observations. Each model contributes six configurations, and every configuration contains both endpoints. These are balanced paired observations rather than a calendar time panel. The published experiments supply the observed variation, but the reanalysis does not independently assign training conditions or replicate training runs. Accordingly, the quantities reported describe the supplied experimental configurations rather than population treatment effects. Drayson et al., 2025, provide the original experimental setting and endpoint measurements. Model identity encompasses differences in architecture, pretraining, and parameter count. Comparisons between model specifications therefore retain these characteristics jointly rather than attributing observed differences to any single component.

4.2 Analytical Specification and Measurement Alignment:

The implemented specification consists of paired arithmetic contrasts. Absolute change, endpoint ratios, and percentage change follow Equations 1 to 3. Absolute change measures additional perplexity units, whereas the ratio and percentage measures express deterioration relative to the starting value. All calculations use the original positive baseline values, avoiding undefined denominators. The analysis calculates these contrasts separately for every model and decoding configuration before producing aggregate summaries. This order preserves the matched comparison and distinguishes the average configuration specific percentage change from the percentage change between pooled endpoint means. The former weights configurations equally; the latter implicitly weights their proportional changes by baseline perplexity. For model m, Equation 4 defines the average configuration specific change:

$$\overline{D}_m = \frac{1}{n_m} \sum_{s=1}^{n_m} D_{ms}$$

Here, $\overline{D}_m$ denotes the mean change for model m, n_m denotes its six decoding configurations, s indexes decoding strategy, and D_{ms} denotes either the absolute change defined in Equation 1 or the percentage change defined in Equation 3. The calculation is performed separately for each change measure. Means describe average deterioration across configurations, while medians describe the central configuration without giving disproportionate numerical influence to large changes. Sample standard deviations, minima, and maxima describe dispersion. These summaries are computed across all 12 pairs and separately across the six pairs belonging to each model. Decoding strategy is retained as a nominal category. No numerical ordering is imposed across greedy decoding, beam search, pure sampling, temperature sampling, top k sampling, and nucleus sampling. This choice preserves the substantive differences between selection procedures. Wiher et al., 2022, demonstrate that decoding properties depend on the generation task, supporting interpretation within the common benchmark examined here.

4.3 Processing, Calculation, and Validation Procedures:

Processing begins with the cleaned endpoint records described in Table 1. Records are identified through unique endpoint identifiers and grouped through configuration identifiers. Within each configuration, generation zero is paired with generation nine. Model identity and decoding category are checked for agreement across the paired records before applying Equations 1 to 3. The calculation sequence retains reported perplexity levels, derives paired contrasts, groups those contrasts by model, and produces descriptive summaries. Cross model comparisons align the same decoding strategy. Strategy rankings are assessed separately within each model using absolute and proportional changes. Comparisons of means and medians use identical configurations, ensuring that differences between summaries arise from the calculation rather than changes in sample composition. The supplied workbook contains transparent formulas for the paired contrasts. Descriptive summaries were independently calculated in Python using NumPy, with sample standard deviations computed using a denominator equal to the number of observations minus one. Calculations retain numerical precision internally and are rounded to two decimal places for presentation. No randomisation, resampling, optimisation, or stochastic estimation is involved. Validation checks assess unique record identifiers, unique combinations of model, decoding strategy, and generation, complete endpoint pairing, and positive finite perplexity values. Calculated changes are checked against their corresponding source inputs. All required observations remain in the analytical sample. No imputation, interpolation, replacement with zero, trimming, or winsorisation is applied. Reported extreme values are retained as experimental outcomes. Their influence is examined through the comparison of means with medians and through inspection of individual configuration changes. This procedure evaluates sensitivity to aggregation without redefining eligibility after observing results. Preserving source measurements separately from derived quantities maintains an auditable computational sequence consistent with the reproducibility principles discussed by Sandve et al., 2013.

4.4 Sensitivity, Interpretive Scope, and Reproducibility:

Sensitivity assessment addresses two consequential choices: the scale of deterioration and its aggregation across configurations. Absolute and proportional changes are compared on the same matched sample to identify scale dependent rankings. Means and medians are compared to determine whether large changes in particular configurations alter aggregate assessments. These comparisons form a descriptive examination of published outcomes rather than prospectively registered confirmatory tests. The analysis reports magnitudes and dispersion without confidence intervals, significance tests, or regression coefficients. The endpoint records do not supply independent training repetitions from which experimental uncertainty could be

estimated. Standard deviations therefore describe variation across configurations and are not interpreted as uncertainty around a population effect. Interpretation remains specific to predictive performance on the observed benchmark. Perplexity changes do not directly measure coherence, reasoning, or human assessments of generated text. Meister et al., 2023, show that standard predictive metrics and perceived generation quality can diverge, establishing a relevant boundary for the outcome interpretation. Likewise, two endpoints establish net change but do not identify the intervening trajectory. The replication materials comprise the consolidated endpoint records, source identifier, configuration keys, variable dictionary, and spreadsheet formulas. These materials permit reconstruction of the paired comparisons and descriptive summaries. They support computational reproduction of the reanalysis rather than independent replication of the underlying training experiments, consistent with the distinction developed by Peng, 2011. The methodological contribution lies in applying complementary measurement scales and aggregation summaries transparently to a common matched sample.

5. Findings:

This analysis examines changes in test perplexity across recursive training endpoints, decoding strategies, and model specifications. The findings address three research questions through matched comparisons of reported experimental results. Interpretation focuses on the magnitude, distribution, and consistency of performance changes within the observed configurations.

5.1 Sample Characteristics and Overall Performance Change:

The analytical sample contains 24 endpoint observations forming 12 matched configurations. Each configuration combines one of two models with one of six decoding strategies and includes measurements at generations zero and nine. All configurations have complete perplexity measurements. The evidence concerns English text continuation on WikiText 2 and comes from experiments published in 2025. The comparisons describe variation across reported configurations rather than independent repetitions of model training. Perplexity increased in all 12 configurations. Across configurations, mean perplexity rose from 21.62 at baseline to 51.22 at generation nine. The mean paired increase was 29.61 perplexity units, while the median increase was 19.37 units. Absolute increases ranged from 1.43 to 71.73 units. The difference between the mean and median indicates that configurations with large increases contributed disproportionately to the aggregate result.

Table 4: Distribution of paired performance changes

Measure	Configurations	Mean	Standard Deviation	Median	Minimum	Maximum
Absolute perplexity increase	12	29.61	25.15	19.37	1.43	71.73
Percentage perplexity increase	12	154.10	173.44	72.72	10.24	513.83
Endpoint perplexity ratio	12	2.54	1.73	1.73	1.10	6.14

The mean percentage increase of 154.10% represents the average of the 12 configuration specific percentage changes. It differs from the 136.96% increase obtained by dividing the pooled endpoint mean by the pooled baseline mean. These summaries apply different weighting to starting perplexities. The configuration specific measure preserves each matched comparison as the analytical unit.

5.2 Recursive Generation and Performance Degradation:

Research question 1 asks how perplexity changes between generations zero and nine within each model and decoding configuration. Table 5 shows that every observed configuration experienced poorer predictive performance at the recursive endpoint. The direction is consistent, but the magnitude varies substantially.

Table 5: Performance changes by model and decoding strategy

Model	Decoding Strategy	Baseline Perplexity	Generation Nine Perplexity	Absolute Increase	Percentage Increase
GPT 2	Greedy	29.29	82.74	53.45	182.49
GPT 2	Beam search	29.25	78.06	48.81	166.87
GPT 2	Pure sampling	29.29	58.64	29.35	100.20
GPT 2	Temperature	29.23	44.55	15.32	52.41
GPT 2	Top k	29.31	48.36	19.05	64.99
GPT 2	Nucleus	29.28	48.96	19.68	67.21
SmolLM2	Greedy	13.96	85.69	71.73	513.83
SmolLM2	Beam search	13.96	84.75	70.79	507.09
SmolLM2	Pure sampling	13.96	15.39	1.43	10.24
SmolLM2	Temperature	13.96	24.88	10.92	78.22
SmolLM2	Top k	13.96	19.86	5.90	42.26
SmolLM2	Nucleus	13.96	22.81	8.85	63.40

For GPT 2, absolute increases ranged from 15.32 units under temperature sampling to 53.45 units under greedy decoding. For SmolLM2, the corresponding range extended from 1.43 units under pure sampling to 71.73 units under greedy decoding. Recursive endpoint deterioration was therefore common to all observed configurations, while its severity depended on the configuration. These results resolve research question 1 descriptively: generation nine perplexity exceeded baseline perplexity in every matched pair. The endpoint structure establishes the direction and magnitude of the observed changes. It does not establish whether deterioration progressed steadily or concentrated within particular intermediate generations.

5.3 Differences Across Decoding Strategies:

Research question 2 asks how performance changes differ across decoding strategies within each model. Greedy decoding produced the largest absolute and proportional increase for both models, followed by beam search. Their consistent positions identify a shared pattern across model specifications despite different baseline performance. Among the four sampling

strategies, the ordering differed between models. GPT 2 recorded its smallest increase under temperature sampling at 52.41%, followed by top k at 64.99%, nucleus at 67.21%, and pure sampling at 100.20%. SmolLM2 recorded its smallest increase under pure sampling at 10.24%, followed by top k at 42.26%, nucleus at 63.40%, and temperature sampling at 78.22%. Pure sampling therefore occupied opposite positions within the sampling group. It produced the greatest proportional deterioration for GPT 2 but the smallest for SmolLM2. Temperature sampling showed the reverse pattern. These differences establish that the observed ranking of sampling strategies depends on model specification. Top k and nucleus sampling showed greater consistency in their relative ordering. Top k produced smaller increases than nucleus sampling for both models. However, neither strategy prevented endpoint deterioration. Research question 2 is resolved through a common disadvantage for greedy decoding and beam search, alongside model dependent differences among sampling strategies.

5.4 Model Comparisons and Sensitivity to Measurement Scale:

Research question 3 asks how performance changes differ between model specifications under matched decoding strategies. Model comparisons depended on both the decoding strategy and the scale used to express deterioration.

Table 6: Model specific summaries of performance change

Model	Configurations	Mean Baseline Perplexity	Mean Endpoint Perplexity	Mean Absolute Increase	Median Absolute Increase	Mean Percentage Increase	Median Percentage Increase
GPT 2	6	29.28	60.22	30.94	24.52	105.70	83.71
SmolLM2	6	13.96	42.23	28.27	9.89	202.51	70.81

SmolLM2 had lower mean baseline and endpoint perplexity than GPT 2. Its mean absolute increase was also smaller, at 28.27 compared with 30.94 units. Nevertheless, its mean percentage increase was larger, at 202.51% compared with 105.70%. The lower starting denominator and large increases under greedy decoding and beam search account for this reversal between absolute and proportional summaries. The medians provide a different comparison. SmolLM2 recorded a lower median absolute increase and a lower median percentage increase. Its median percentage increase was 70.81%, compared with 83.71% for GPT 2. The contrast between its mean and median demonstrates the influence of the two configurations with increases exceeding 500%. Matched comparisons further qualify the aggregate pattern. SmolLM2 deteriorated more than GPT 2 under greedy decoding and beam search on both measurement scales. Under pure sampling, top k, and nucleus sampling, SmolLM2 deteriorated less on both scales. Temperature sampling produced a scale dependent result: SmolLM2 had a smaller absolute increase of 10.92 units compared with 15.32 units, but a larger percentage increase of 78.22% compared with 52.41%.

Research question 3 is therefore resolved by conditional rather than uniform model differences. SmolLM2 showed smaller absolute increases under all four sampling strategies, while GPT 2 showed smaller increases under the two deterministic strategies. Proportional comparisons preserved this pattern except under temperature sampling. The findings establish that baseline performance, decoding strategy, and summary scale jointly determine the descriptive assessment of degradation. Across the three research questions, the evidence supports a consistent direction of endpoint deterioration but rejects a uniform ranking of model resilience. The measured differences concern predictive performance within the reported benchmark configurations. Their contribution is the distinction between common deterioration, configuration dependent severity, and aggregation sensitive comparisons.

6. Discussion:

The apparent resilience of a language model can reverse when degradation is measured against its starting performance. Tables 4 and 6 distinguish absolute deterioration from proportional deterioration, showing that these measures answer different questions. SmolLM2 exhibits a smaller mean absolute increase but a larger mean percentage increase than GPT 2, while its median deterioration remains lower on both scales. Equations 1 to 3 explain this divergence through baseline denominators and the distribution of configuration specific changes. Shumailov et al., 2024, establish recursive distributional degradation as a central concern in synthetic training. The present reanalysis adds a measurement qualification: conclusions about comparative resilience depend on the scale and summary statistic used. Its contribution is an explicit assessment of this dependence within published experimental endpoints, rather than a claim to have discovered recursive degradation independently. Decoding strategy changes the observed relationship between model specification and deterioration. Table 5 shows that pure sampling produces the smallest proportional increase for SmolLM2 but the largest increase among sampling strategies for GPT 2. Temperature sampling reverses that ordering. Drayson et al., 2025, already document decoding related differences in the underlying experiments; the matched reanalysis clarifies their implication for comparison: neither model identity nor decoding strategy provides a sufficient standalone ranking of resilience. Theoretically, token selection determines which parts of a model's distribution become subsequent training material, making configuration dependent outcomes plausible. These endpoints identify patterns consistent with that explanation, without separating the contribution of individual transmission mechanisms. The resulting refinement is to treat recursive performance as a property of the model and decoding configuration rather than an intrinsic attribute of either component alone.

The distribution of deterioration also changes the meaning of aggregate model comparisons. In Table 6, SmolLM2's mean percentage increase exceeds its median substantially because greedy decoding and beam search occupy the upper end of the observed distribution. Table 5 shows that both strategies produce greater deterioration than the sampling alternatives for each model. This pattern supports distinguishing concentration around probable outputs from sampling-based selection when interpreting recursive training outcomes. However, the reversal among sampling strategies prevents a general prescription that greater sampling variability always preserves performance. Wiher et al., 2022, demonstrate that decoding properties vary across generation tasks. The present comparisons sharpen that boundary within a shared benchmark: the ordering also varies across model specifications. Aggregate averages therefore require configuration level interpretation before they can support operational choices. Measurement creates an additional constraint on what resilience means. Under temperature sampling, Table 5 assigns SmolLM2 a smaller absolute increase but a larger percentage increase than GPT 2. Equation 1 measures the additional perplexity

burden, whereas Equations 2 and 3 measure deterioration relative to baseline. Neither scale supersedes the other, because each preserves a different comparison. Moreover, predictive performance remains distinct from generation quality. Meister et al., 2022, explain why highly probable language need not correspond to desirable text, while Meister et al., 2023, show that standard model metrics can diverge from perceived output quality. The analytical implication is precise: the rankings reported here concern benchmark prediction. Reporting paired levels, absolute changes, proportional changes, and both means and medians makes the basis of those rankings visible and reduces dependence on a selectively favourable summary.

The international relevance lies in evaluation practices for synthetic training rather than geographical differences. Kazdan et al., 2025, distinguish training workflows that replace original data from workflows that accumulate real and synthetic material, showing that collapse depends on how generated data enter learning. The deterioration in Table 5 therefore belongs to a specific recursive training setting and should inform decisions within that boundary. For developers comparing such configurations, Table 6 supports retaining separate assessments of starting performance, endpoint performance, absolute deterioration, and proportional deterioration. For benchmark reporting, Table 4 supports presenting distributional summaries alongside averages. Together, these findings refine global debates by distinguishing whether deterioration occurs from how severely configurations deteriorate and how comparative rankings are constructed. The practical contribution is a transparent basis for configuration selection and performance reporting, grounded in matched observations rather than a universal claim about synthetic data.

7. Conclusion and Implications:

Reliable comparisons of artificial intelligence systems require clarity about what deterioration means. This analysis shows that repeated learning from generated material produces different observed losses across configurations, while starting performance and aggregation choices can reverse comparative assessments. Its contribution is a descriptive reanalysis demonstrating that resilience rankings depend jointly on experimental configuration and measurement scale. This qualifies interpretations that treat resilience as an intrinsic property of a model or generation method. The theoretical implication is that the occurrence of deterioration must be distinguished from its severity and from the construction of comparative rankings. For developers, configuration selection should consider both retained predictive performance and deterioration relative to the starting point. For evaluators, reporting absolute and proportional changes alongside distributional summaries makes consequential differences visible. Procurement and benchmarking guidance can incorporate these reporting practices to make comparisons more transparent across suppliers and evaluation settings. Operationally, preserving matched baselines and reviewing individual configurations before relying on aggregate averages can improve performance monitoring. These implications extend to international evaluation practices, although the observed rankings remain specific to the examined experiments. More transparent comparisons can support accountable deployment decisions and reduce the risk that favourable summaries obscure substantial deterioration in particular configurations.

The published endpoints provide a basis for extending the analysis through independent experimental replication. Additional training repetitions would permit uncertainty estimation, while intermediate observations would reveal whether deterioration accumulates gradually or concentrates at particular stages. Broader comparisons across model families, languages, benchmarks, and training workflows would assess the transferability of the observed ranking reversals. Direct measurements of generated training material could distinguish distributional narrowing from other explanations for predictive deterioration. Combining predictive evaluation with independently assessed output quality would clarify when measurement scales yield consistent operational conclusions and when different dimensions of performance require separate decisions.

References:

1. Drayson, G., Yilmaz, E., & Lampos, V. 2025. Machine generated text detection prevents language model collapse. In C. Christodoulopoulos, T. Chakraborty, C. Rose, & V. Peng, Eds., Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 29657 to 29673. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1506
2. Kazdan, J., Schaeffer, R., Dey, A., Gerstgrasser, M., Rafailov, R., Donoho, D. L., &Koyejo, S. 2025. Collapse or thrive: Perils and promises of synthetic data in a self generating world. Proceedings of Machine Learning Research, 267, 29469 to 29494. https://proceedings.mlr.press/v267/kazdan25a.html
3. Meister, C., Pimentel, T., Wiher, G., & Cotterell, R. 2023. Locally typical sampling. Transactions of the Association for Computational Linguistics, 11, 102 to 121. https://doi.org/10.1162/tacl_a_00536
4. Meister, C., Wiher, G., Pimentel, T., & Cotterell, R. 2022. On the probability–quality paradox in language generation. In S. Muresan, P. Nakov, & A. Villavicencio, Eds., Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers, pages 36 to 45. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-short.5
5. Peng, R. D. 2011. Reproducible research in computational Science. Science, 334, Issue 6060, 1226 to 1227. https://doi.org/10.1126/science.1213847
6. Sandve, G. K., Nekrutenko, A., Taylor, J., & Hovig, E. 2013. Ten simple rules for reproducible computational research. PLOS Computational Biology, 9, issue 10, Article e1003285. https://doi.org/10.1371/journal.pcbi.1003285
7. Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. 2024. AI models collapse when trained on recursively generated data. Nature, 631, issue 8022, 755 to 759. https://doi.org/10.1038/s41586-024-07566-y
8. Wiher, G., Meister, C., & Cotterell, R. 2022. On decoding strategies for neural text generators. Transactions of the Association for Computational Linguistics, 10, 997 to 1012. https://doi.org/10.1162/tacl_a_00502