Relating Loss Function Scaling in Neural Language Models to Hilberg's Conjecture

Adrián Raso González

Vienna University of Technology

IMS International Conference on Statistics and Data Science (ICSDS), 2025

Introduction

- Natural language data violates common modeling asumptions (short-range dependence, finite memory, i.i.d.).

Introduction

- Natural language data violates common modeling asumptions (short-range dependence, finite memory, i.i.d.).
- Hilberg's conjecture (1990) suggests that problems in language modeling may be due to a sublinear scaling of block entropy in language sources of the form

$$H(n) \propto n^{\beta}.$$

If true, language contains **unbounded structure**, and a **non-Markovian scaling**.

Introduction

- Natural language data violates common modeling asumptions (short-range dependence, finite memory, i.i.d.).
- Hilberg's conjecture (1990) suggests that problems in language modeling may be due to a sublinear scaling of block entropy in language sources of the form

$$H(n) \propto n^{\beta}.$$

If true, language contains **unbounded structure**, and a **non-Markovian scaling**.

Goal: Empirically estimate $\hat{H}(n)$ growth, fit Hilberg scaling $\hat{\beta}$, and test whether autoregressive neural models of different capacity reproduce this scaling.

Background

Block entropy (or joint entropy) growth in natural language, according to Hilberg, follows

$$H(n) \sim n^{\beta}, \quad 0 < \beta < 1$$

Background

Block entropy (or joint entropy) growth in natural language, according to Hilberg, follows

$$H(n) \sim n^{\beta}, \quad 0 < \beta < 1$$

In modern accounts, this is usually presented through its relaxed form, using an extensive term plus a subextensive correction:

$$H(n) \approx A + hn + Cn^{\beta}, \qquad 0 < \beta < 1.$$

Background

Block entropy (or joint entropy) growth in natural language, according to Hilberg, follows

$$H(n) \sim n^{\beta}, \quad 0 < \beta < 1$$

In modern accounts, this is usually presented through its relaxed form, using an extensive term plus a subextensive correction:

$$H(n) \approx A + hn + Cn^{\beta}, \qquad 0 < \beta < 1.$$

Empirically, estimates often find the following:

$$\beta \sim 0.5$$

Background

Block entropy (or joint entropy) growth in natural language, according to Hilberg, follows

$$H(n) \sim n^{\beta}, \quad 0 < \beta < 1$$

In modern accounts, this is usually presented through its relaxed form, using an extensive term plus a subextensive correction:

$$H(n) \approx A + hn + Cn^{\beta}, \qquad 0 < \beta < 1.$$

Empirically, estimates often find the following:

$$\beta \sim 0.5$$

β is called a *Hilberg exponent*, measuring the degree of long-range statistical dependence. Smaller values of β hint stronger long-range structure and weak rate of information, while $\beta \to 1$ correspond to soft long-range dependencies and Markovian behavior.

Background

A more accessible formulation of Hilberg's conjecture is obtained by considering the conditional entropy

$$h_n \;=\; H(X_n \mid X_1^{n-1}) \;=\; H(n) - H(n-1).$$

Background

A more accessible formulation of Hilberg's conjecture is obtained by considering the conditional entropy

$$h_n \;=\; H(X_n \mid X_1^{n-1}) \;=\; H(n) - H(n-1).$$

If the block entropy grows sublinearly,

$$H(n) \propto n^{\beta}, \qquad 0 < \beta < 1,$$

then the discrete derivative follows the equivalent power law

$$h_n \;\propto\; n^{\beta-1}.$$

This comes handy because thus we don't need the full joint distribution. With models, the per-token loss at context length n gives us an estimate of this h_n, which we can use to study Hilberg scaling.

Why This May Challenge Models?

$$\text{Finite-memory Markov model} \Rightarrow H(n) \propto n \Rightarrow \beta = 1$$

$$0 < \beta < 1 \Rightarrow \text{the source exhibits unbounded effective memory}$$

Why This May Challenge Models?

$$\text{Finite-memory Markov model} \Rightarrow H(n) \propto n \Rightarrow \beta = 1$$

$$0 < \beta < 1 \Rightarrow \text{the source exhibits unbounded effective memory}$$

Yet, other modeling paradigms have stronger performance in advanced natural language tasks, like neural LMs, particularly Transformers.

Why This May Challenge Models?

$$\text{Finite-memory Markov model} \Rightarrow H(n) \propto n \Rightarrow \beta = 1$$

$$0 < \beta < 1 \Rightarrow \text{the source exhibits unbounded effective memory}$$

Yet, other modeling paradigms have stronger performance in advanced natural language tasks, like neural LMs, particularly Transformers.

- Long-range information is mediated through attention and compressed internal representations.
- Their training is local (short-horizon targets), raising the question whether the scaling can implicitly emerge globally.
- Yet, they operate with finite context windows.

Why This May Challenge Models?

$$\text{Finite-memory Markov model} \Rightarrow H(n) \propto n \Rightarrow \beta = 1$$

$$0 < \beta < 1 \Rightarrow \text{the source exhibits unbounded effective memory}$$

Yet, other modeling paradigms have stronger performance in advanced natural language tasks, like neural LMs, particularly Transformers.

- Long-range information is mediated through attention and compressed internal representations.
- Their training is local (short-horizon targets), raising the question whether the scaling can implicitly emerge globally.
- Yet, they operate with finite context windows.

Question: Do neural LMs approximate Hilberg scaling?

Method

- Data: English Wikipedia dump (20 July 2025)

Method

- Data: English Wikipedia dump (20 July 2025)
- Estimate $H(n)$ growth via Prediction by Partial Matching (PPM) over 50 disjoint 1M-token subsets

Method

- Data: English Wikipedia dump (20 July 2025)
- Estimate $H(n)$ growth via Prediction by Partial Matching (PPM) over 50 disjoint 1M-token subsets
- Train autoregressive models with increasing capacity (1M, 5M, 10M parameters) and context (128, 512, 1024 tokens)

Method

- Data: English Wikipedia dump (20 July 2025)
- Estimate $H(n)$ growth via Prediction by Partial Matching (PPM) over 50 disjoint 1M-token subsets
- Train autoregressive models with increasing capacity (1M, 5M, 10M parameters) and context (128, 512, 1024 tokens)
- Compare Hilberg exponent β between the estimated true block entropy $\hat{H}(n)$ growth of the dataset and the discrete derivative h_n of the models loss curves $\ell(n)_{\text{model}}$

PPM estimation

Why PPM?

- Universal coding method widely used for empirical entropy estimation in text
- Provides per-token code lengths (surprisals) using variable-order context models with escape/backoff
- Lets us estimate how conditional uncertainty decreases with available context

PPM estimation

Why PPM?

- Universal coding method widely used for empirical entropy estimation in text
- Provides per-token code lengths (surprisals) using variable-order context models with escape/backoff
- Lets us estimate how conditional uncertainty decreases with available context

Codelength view:

$$L(n) = \sum_{t=1}^{n} -\log_2 \hat{P}(x_t \mid x_1^{t-1}) \qquad \Rightarrow \qquad \hat{H}(n) \approx \mathbb{E}[L(n)].$$

PPM estimation

Why PPM?

- Universal coding method widely used for empirical entropy estimation in text
- Provides per-token code lengths (surprisals) using variable-order context models with escape/backoff
- Lets us estimate how conditional uncertainty decreases with available context

Codelength view:

$$L(n) = \sum_{t=1}^{n} -\log_2 \hat{P}(x_t \mid x_1^{t-1}) \qquad \Rightarrow \qquad \hat{H}(n) \approx \mathbb{E}[L(n)].$$

For context-gain scaling: we analyze the conditional code length as a function of context k and fit

$$\ell(k) \approx \ell_\infty + c\, k^{\beta-1}.$$

H(n) results

n	$\hat{H}(n)$	95% CI
1	60.20	(59.04, 61.36)
2	89.39	(87.23, 91.55)
5	169.73	(165.16, 174.31)
10	277.13	(269.30, 284.95)
20	450.05	(439.15, 460.94)
50	813.98	(791.29, 836.66)
100	1261.48	(1230.45, 1292.51)
200	1973.55	(1923.25, 2023.85)
500	3627.54	(3542.88, 3712.20)
1000	6036.00	(5918.66, 6153.35)
5000	22784.35	(22530.19, 23038.52)
10000	42775.79	(42337.21, 43214.36)

H(n) results

n	$\hat{H}(n)$	95% CI
1	60.20	(59.04, 61.36)
2	89.39	(87.23, 91.55)
5	169.73	(165.16, 174.31)
10	277.13	(269.30, 284.95)
20	450.05	(439.15, 460.94)
50	813.98	(791.29, 836.66)
100	1261.48	(1230.45, 1292.51)
200	1973.55	(1923.25, 2023.85)
500	3627.54	(3542.88, 3712.20)
1000	6036.00	(5918.66, 6153.35)
5000	22784.35	(22530.19, 23038.52)
10000	42775.79	(42337.21, 43214.36)

Hilberg fit: $\beta \approx 0.5601,\ R^2 \approx 0.9804,\ \text{stderr} \approx 0.0354$

Considered models

Three models are considered in the experiment, each with increasing capacity and context, all autoregressive and using a GPT architecture.

Model	**Parameters**	**Context length or block**	**Epochs**
Small (S)	1M	128 tokens	1
Medium (M)	5M	512 tokens	1
Large (L)	10M	1024 tokens	1

The activation function used is GeLU, a standard function for GPT models. For more details on the hyperparameters of the model (hidden size, number of heads, FF dimension, vocabulary size, and such) you can check the Hugging Face collection **Hilberg Scaling in Neural Language Models (ICSDS 2025)** (resources references at the end of the presentation).

Loss function evaluation

- We study the loss function

$$\ell(n) = h_n = H(X_t \mid X_{t-n}^{t-1}) = H(n) - H(n-1)$$

over a range of context blocks

$$n \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\},$$

computing the loss function of the model for each block.

Loss function evaluation

- We study the loss function

$$\ell(n) = h_n = H(X_t \mid X_{t-n}^{t-1}) = H(n) - H(n-1)$$

over a range of context blocks

$$n \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1024\},$$

computing the loss function of the model for each block.

- We fit the context-gain curve with an asymptotic power law:

$$\ell(k) \approx \ell_\infty + c\, k^{\beta-1}.$$

Here $\beta < 1$ corresponds to slow, persistent improvements with additional context.

Results

- Let's remember that the true entropy scaling $\hat{\beta}$ of the dataset is 0.56, hinting sublinear growth, which is a characteristic exponent of any natural language source.

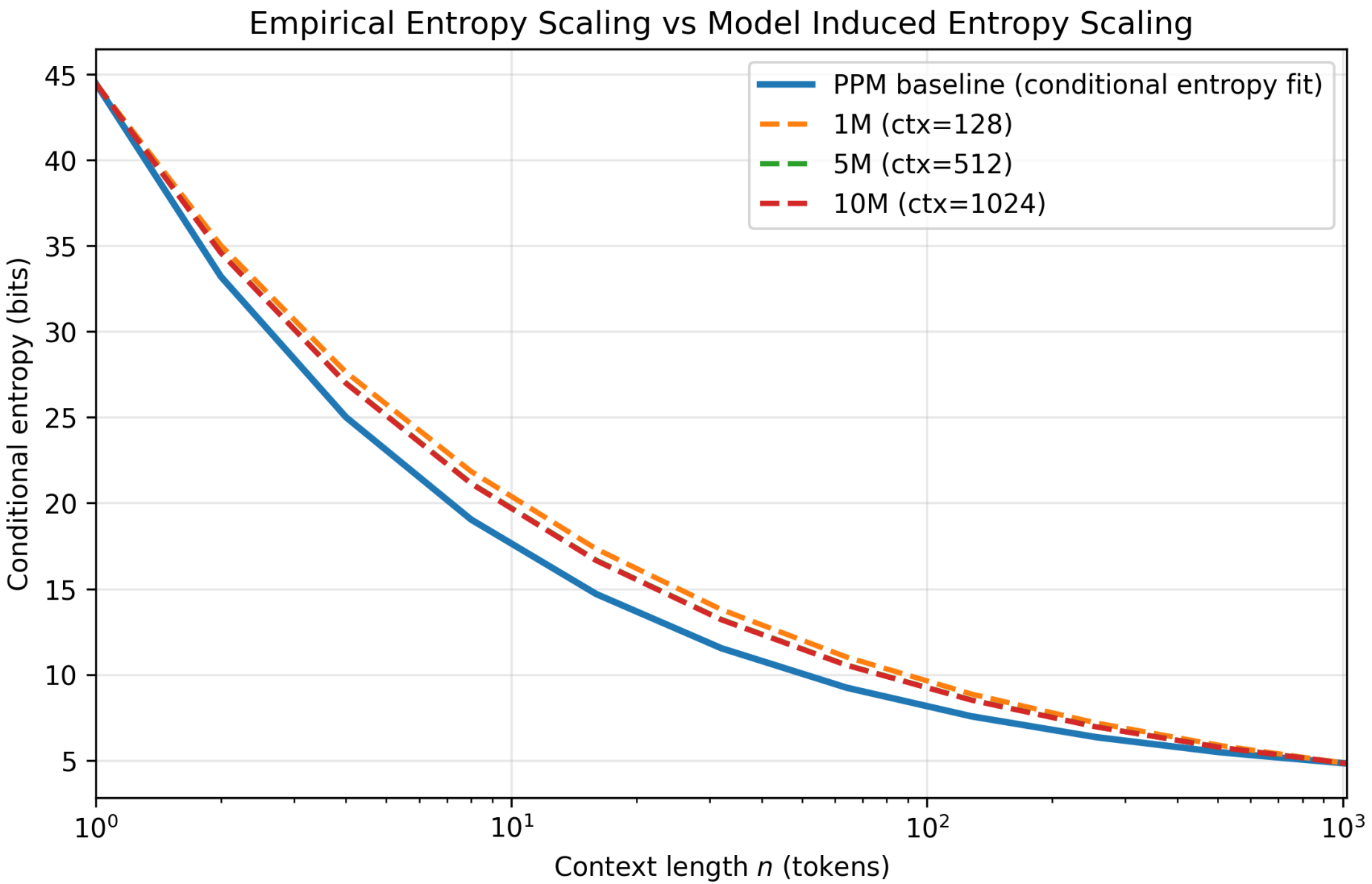
Results

- Let's remember that the true entropy scaling $\hat{\beta}$ of the dataset is 0.56, hinting sublinear growth, which is a characteristic exponent of any natural language source.
- The considered models show the following values for their Hilberg exponent:

Model	**Context length**	β
1M	128	0.6452
5M	512	0.6209
10M	1024	0.6198

- The models exhibit a clear subextensive regime ($\beta < 1$), additional context yielding a slow but persistent reduction in loss.
- Across scales, model size mainly shifts the curve downward (lower ℓ_∞) rather than strongly changing the exponent.

Results: Pre-asymptotic Analysis



Conditional entropy scaling in a pre-asymptotic regime

Conclusions

- Using PPM as an entropy-scaling baseline, we recover a subextensive behavior consistent with Hilberg's growth in the data.
- Transformer LMs exhibit a persistent power law improvement with context: the loss approaches an asymptote as

$$\ell(k) \approx \ell_\infty + c\, k^{\beta-1},$$

  with a stable exponent across model sizes.
- Model scaling primarily improves the level rather than qualitatively changing the shape of the context-gain curve. Long-context gains remain slow but non-negligible.

Conclusions

- Using PPM as an entropy-scaling baseline, we recover a subextensive behavior consistent with Hilberg's growth in the data.
- Transformer LMs exhibit a persistent power law improvement with context: the loss approaches an asymptote as

$$\ell(k) \approx \ell_\infty + c\, k^{\beta-1},$$

  with a stable exponent across model sizes.
- Model scaling primarily improves the level rather than qualitatively changing the shape of the context-gain curve. Long-context gains remain slow but non-negligible.

Important takeaway: Autoregressive Transformers reproduce a Hilberg power law decay of conditional entropy with context, linking classical estimates to neural loss curves in the Transformer family.

References

W. Hilberg,

Der bekannte Grenzwert der redundanzfreien Information in Texten – eine Fehlinterpretation der Shannon-Experimente?

Frequenz, vol. 44, no. 12, pp. 243–248, 1990.

Ł. Debowski,

Hilberg's Conjecture – a Challenge for Machine Learning,

Schedae Informaticae, vol. 23, pp. 33–44, 2014.

W. Ebeling and T. Pöschel,

Entropy and Long-Range Correlations in Literary English,

Europhysics Letters, vol. 26, no. 4, pp. 241–246, 1994.

C. E. Shannon,

Prediction and Entropy of Printed English,

Bell System Technical Journal, vol. 30, no. 1, pp. 50–64, 1951.

W. J. Teahan and J. G. Cleary,

The Entropy of English Using PPM-Based Models,

In Proceedings of the IEEE Data Compression Conference (DCC), pp. 53–62, 1996.

Resources

Code, data, and talk materials available here:

GitHub repo (Method codebase, talk resources)

AdrianRasoOnGit / hilberg-scaling-neural-language-models-icsds-2025

Hugging Face collection (Dataset and models)

AdrianRasoOnHF / hilberg-scaling-in-neural-language-models-icsds-2025

Thank you for your attention.