Analyzing and Modeling LLM Response Lengths with Extreme Value Theory: Anchoring Effects and Hybrid Distributions

Liuxuan Jiao[1,2], **Chen Gao**[2*], **Yiqian Yang**[3], **Chenliang Zhou**[1],
Yixian Huang[1], **Xinlei Chen**[2*], **Yong Li**[2]
[1]University of Cambridge, [2]Tsinghua University, [3]HKUST (Guangzhou)
lj393@cam.ac.uk, chgao96@tsinghua.edu.cn, yyang937@connect.hkust-gz.edu.cn,
chenliang.zhou@cst.cam.ac.uk, yh602@cam.ac.uk,
xinlei.chen@sz.tsinghua.edu.cn, liyong07@tsinghua.edu.cn

Abstract

Accurate modeling and control of response length is essential for optimizing large language model (LLM) deployment, impacting computational efficiency, user experience, and system reliability. We develop a statistical framework based on extreme value theory, analyzing 14,301 GPT-4o responses across temperature settings and prompting strategies, with cross-validation on Qwen and DeepSeek architectures. Our analysis reveals that response lengths follow Weibull-type generalized extreme value (GEV) distributions, exhibiting heavier tails under stochastic generation conditions. The key contributions include: (1) a novel GEV-generalized Pareto (GPD) hybrid model that achieves superior tail fit ($R^2_{\text{CDF}} = 0.9993$ vs standalone GEV's 0.998) while preserving architectural generalizability; (2) quantitative characterization of prompt anchoring effects, showing reduced dispersion but increased outlier propensity under randomization; and (3) identification of temperature-dependent response patterns that remain consistent across architectures, where higher temperatures amplify length variability while maintaining the underlying extreme-value mechanisms. The proposed hybrid model's adaptive threshold selection enables precise verbosity control in production systems, regardless of the specific LLM architecture employed. These findings provide both theoretical insights into LLM generation patterns and practical tools for response length optimization.

1 Introduction

Large language model deployments face a critical operational challenge: response length variability directly impacts computational costs and user satisfaction (Nayab et al., 2024; Zheng et al., 2023). While API pricing scales linearly with token counts, users increasingly demand concise answers tailored to contextual needs (Butcher et al., 2025). Despite these practical imperatives, the field lacks fundamental understanding of LLM verbosity patterns across different architectures (Muñoz-Ortiz et al., 2024). Current approaches treat length as an incidental output property rather than a statistically regular phenomenon worthy of rigorous modeling (Borbély and Kornai, 2019).

Recent studies have made incremental progress in related areas. Temperature scaling has been shown to affect output diversity (Radford et al., 2019), while reinforcement learning from human feedback demonstrates length-quality tradeoffs (Singhal et al., 2023). Cross-linguistic analyses of human communication suggest potential distribution families like lognormal or Weibull for natural utterance lengths (Borbély and Kornai, 2019). Public speech analysis further reveals temporal compression patterns in human verbal output (Tsizhmovska and Martyushev, 2021). However, these findings focus on biological language production, leaving neural language models' statistical properties unexplored - particularly the consistency of length distributions across model architectures and scales.

Three fundamental barriers prevent effective verbosity control. First, existing evaluation frameworks lack principled statistical models for length distributions despite their operational importance, with no systematic comparison across model families. Second, the interaction between prompt structure and generation properties remains poorly quantified, with anecdotal evidence outweighing systematic analysis. Third, the effects of temperature scaling on extreme-value behavior have not been characterized across different models, despite its known impact on output randomness. These gaps leave practitioners without reliable tools for predicting or shaping LLM verbosity patterns across the growing ecosystem of available models.

We bridge these gaps through extreme value analysis of 14,301 GPT-4o responses generated

[*] Corresponding authors.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 32992–33002
November 4-9, 2025 ©2025 Association for Computational Linguistics

under controlled conditions, extended with cross-architecture validation on Qwen and DeepSeek models. Our framework combines three methodological innovations: generalized extreme value (GEV) distributions model central tendencies across architectures, generalized Pareto (GPD) corrections address tail behavior through optimized thresholds, and causal analysis quantifies anchoring effects across prompt variants. Controlled experiments vary temperature (0 vs. 0.7) and prompt structures (direct vs. anchored) to isolate generation mechanism impacts while maintaining architectural generalizability.

This work establishes four key advances in LLM verbosity understanding. We demonstrate that response lengths follow Weibull-type GEV distributions consistently across models, with shape parameters revealing temperature-dependent tail behaviors. Our GEV-GPD hybrid model achieves unprecedented tail fit accuracy ($R^2_{\text{CDF}} = 0.9993$) while maintaining cross-model applicability. Quantification of anchoring effects reveals reduction in dispersion parameters under deterministic generation that holds across tested architectures. Practically, we develop threshold selection methods that reduce extreme-length outliers, enabling production systems to balance conciseness with completeness regardless of model choice.

Our analysis focuses on English question-answering tasks using the HotpotQA dataset, with a specific examination of how the anchoring phrase “As previously stated," influences response verbosity. This controlled setup provides a foundation for understanding distributional patterns that future work can extend to other tasks and prompting styles.

2 Related Work

Controlling response length in large language models (LLMs) is critical for efficient deployment, yet statistical modeling of length distributions remains underexplored. Prior work spans temperature effects, long-tailed distributions, length generalization, and anchoring, but lacks a unified framework integrating these factors. We review these efforts, highlighting gaps our GEV-GPD hybrid model addresses.

2.1 Temperature Effects and Long-Tailed Distributions

Temperature governs LLM output randomness, influencing verbosity and tail behavior. Early work noted that higher temperatures increase diversity and length (Radford et al., 2019), with recent studies confirming temperature-driven phase transitions in output distributions (Arnold et al., 2024). However, (Peeperkorn et al., 2024) found weak temperature effects on creativity, with slight novelty increases at higher settings, and (Renze, 2024) reported minimal impact on problem-solving tasks, suggesting task-specific influences. Long-tailed distributions exacerbate challenges for rare inputs, as shown in code generation, where performance degrades due to skewed distributions (Zhou et al., 2023). Data augmentation has been proposed to mitigate such issues (Wang et al., 2024). Beyond LLMs, sentence length distributions in public speaking follow Weibull distributions, with lengths decreasing over time (Tsizhmovska and Martyushev, 2021), while cross-linguistic studies suggest lognormal fits (Borbély and Kornai, 2019). These findings highlight the prevalence of heavy-tailed distributions but lack quantitative models for LLM-specific length prediction.

Our work diverges by developing a GEV-GPD hybrid model that explicitly quantifies temperature-dependent tail behavior in LLM outputs. Unlike prior studies, which describe distributions qualitatively or focus on non-LLM contexts (Tsizhmovska and Martyushev, 2021; Borbély and Kornai, 2019), we provide a statistically rigorous framework that captures both bulk and extreme length distributions, enabling precise prediction and theoretical insights into LLM verbosity.

2.2 Length Generalization and Anchoring Strategies

LLMs struggle with generating or processing long outputs, but prompting strategies improve generalization (Anil et al., 2022). Constrained prompting enhances conciseness (Nayab et al., 2024), and precise length control has been achieved through tailored methods (Butcher et al., 2025). Longer reasoning steps boost performance (Jin et al., 2024), while shorter inputs degrade reasoning (Levy et al., 2024). Length optimization in RLHF influences helpfulness perceptions (Singhal et al., 2023), and response length prediction improves inference efficiency (Zheng et al., 2023). Anchoring biases, where initial prompts disproportionately shape outputs, have been observed in LLMs, with mitigation requiring comprehensive hint collection rather than simple strategies like Chain-of-Thought (Lou and Sun, 2024). Statistical modeling of lengths is less studied, with (Muñoz-Ortiz et al., 2024) noting consistent distributions without quantitative frameworks. Transformer architectures enable anomaly detection (Vaswani et al., 2017), but length-specific outliers remain underaddressed.

Our approach advances this field by integrating anchoring effects into a statistical length model, using GEV-GPD to quantify how prompts reduce central tendency in deterministic settings while increasing outliers in stochastic ones. Unlike prior work, which focuses on empirical observations or mitigation without statistical grounding (Lou and Sun, 2024; Nayab et al., 2024), our framework provides a unified analysis of length distributions, offering practical applications in outlier detection and prompt engineering.

3 Methods

3.1 Experimental Design

We analyzed 14,301 English question-answer pairs from the HotpotQA test set (Yang et al., 2018) under controlled conditions to investigate (1) the statistical modeling of LLM output word counts, and (2) the impact of anchored prompts and temperature on LLM response characteristics. The experimental setup (Table 1) used GPT-4o (gpt-4o-2024-11-20) with two temperature conditions ($T = 0$: $n =$ 6,945; $T = 0.7$: $n =$ 7,356, the latter being the standard default for most LLM APIs) and two prompt variants:

- **Direct prompts**: Standard question format without additional framing
- **Anchored prompts**: Questions prefixed with “As previously stated," to induce semantic anchoring effects

Table 1: Experimental Conditions

Parameter	Value
Model	GPT-4o (gpt-4o-2024-11-20)
Temperature	0 (deterministic), 0.7 (stochastic)
Prompt variants	Direct, Anchored
Length metric	Word count (whitespace-delimited)

3.2 Response Generation Protocol

Responses were generated under deterministic ($T = 0$) and stochastic ($T = 0.7$) sampling conditions for both standard prompts (e.g., “Are both Volvic and Canfield’s Diet Chocolate Fudge natural spring waters ?") and anchored prompts (e.g., “As previously stated, are both Volvic and Canfield’s Diet Chocolate Fudge natural spring waters ?"). The anchor phrase was selected as a representative discourse marker that implies prior context.

3.3 Statistical Modeling with GEV

Four candidate distributions were evaluated: generalized extreme value (GEV), log-normal, Weibull, and generalized error model (GEM-2). Model selection via maximum likelihood estimation used both Akaike information criterion (AIC) and root-mean-square error (RMSE). The GEV distribution provided the best baseline fit but showed right-tail deficiencies.

3.4 Generalization to Other Models

To validate the robustness of the Generalized Extreme Value (GEV) distribution hypothesis across model architectures, we extended our analysis to three additional language models: Qwen3-8B, Qwen3-14B, and DeepSeek-V3 (DeepSeek-V3-0324). We used direct prompts with a fixed temperature of 0.7 to ensure consistency in output diversity. Each model was evaluated on distinct dataset sizes: Qwen3-8B and Qwen3-14B on 200 samples each, and DeepSeek-V3 on 600 samples.

3.5 Two-Stage Extreme Value Modeling

We developed a GEV-GPD hybrid model with threshold $u = Q_{0.95}$ (selected via MSE minimization although mixed MSE minimization was also done), improving performance from $R^2_{\text{CDF}} = 0.998$ to $R^2_{\text{CDF}} = 0.9993$. The model transitions from GEV to generalized Pareto distribution (GPD) at $x > u$.

The threshold optimization employs a mixed objective function:

$$\mathcal{L}_{\text{mixed}}(u) = 0.6\mathcal{L}_{\text{MSE}} + 0.3|\xi| + 0.1\mathcal{I} \quad (1)$$

where:

- $\mathcal{L}_{\text{MSE}} = n^{-1}\sum_i (F_i - \hat{F}_i)^2$ is the mean squared error term
- ξ is the **GPD shape parameter** governing tail heaviness, critical for modeling extreme values (outliers)

- $\mathcal{I}$ is an **information-based penalty** term defined as $\mathcal{I} = |\xi \cdot \frac{\beta}{u}|$, which balances model complexity and prevents overfitting by penalizing excessively complex tail behavior

This formulation balances fit quality ($\mathcal{L}_{\text{MSE}}$), tail properties ($|\xi|$), and model complexity ($\mathcal{I}$), ensuring robust threshold selection.

3.6 Input-Output Analysis

Pearson correlation coefficients were computed to assess the linear relationships between (1) context length and output length, and (2) question length and output length. The results revealed negligible correlations, suggesting that output length is generated independently of input characteristics.

4 Results

4.1 Model Selection

We evaluate four parametric distributions for response lengths, as shown in Table 2.

Table 2: Model Comparison (Temperature = 0.7, Direct)

Model	RMSE	AIC
GEV	0.000473	65783
LogNormal	0.001119	66434
GMM	0.001907	68399
Weibull	0.002506	69430

The GEV distribution emerged as the optimal model, demonstrating both strong statistical significance ($\Delta\text{AIC} > 650$) and superior predictive performance with an RMSE reduction exceeding 50% compared to alternative approaches.

4.2 GEV Parameter Estimates

GEV parameters (c, μ, σ) are estimated via maximum likelihood using SciPy's `genextreme.fit`:

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \log f(x_i; \theta) \tag{2}$$

where f is the GEV density. 95% confidence intervals are computed via bootstrap.

Tables 3 shows GEV parameters for temperature settings 0 and 0.7. Both conditions exhibit Weibull-type distributions ($c < 0$), with anchoring reducing tail thickness ($c = -0.355$ vs -0.411 at temp=0; -0.361 vs -0.411 at temp=0.7). Anchoring consistently lowers μ and σ values (Tables 3).

4.3 Outlier Analysis

At **Temperature = 0**, the direct method produced 80 outliers (1.2% of cases, with a maximum length of 741 words), while the anchored approach yielded slightly fewer at 79 outliers (1.1%, max 580 words). When increasing to **Temperature = 0.7**, we observed 81 outliers (1.1%, max 680 words) for direct generation compared to 87 outliers (1.2%, max 759 words) with anchoring. This pattern reveals that anchoring reduces outliers by 1.3% at temperature 0, but interestingly increases them by 7.4% at temperature 0.7, demonstrating its efficacy is temperature-dependent.

4.4 Temperature Comparison

Comparing temperature settings (0 vs. 0.7) reveals several key patterns. The tail behavior shows similar shape parameters ($c \approx -0.41$) across conditions, though anchoring produces slightly heavier tails at temperature 0.7 (-0.361) compared to temperature 0 (-0.355). For central tendency, we observe that μ consistently increases with temperature, rising from 25.60 to 26.19 for direct generation and from 24.39 to 24.80 for anchored generation. Variability also grows with temperature, with σ increasing from 13.56 to 14.20 (direct) and from 12.10 to 12.46 (anchored).

Examining extremes, maximum lengths increase for both methods: from 58.6 to 60.8 words for direct generation and from 58.4 to 59.3 words for anchored generation. Outlier analysis shows similar counts at temperature 0 (80 for direct vs 79 for anchored), but diverges at temperature 0.7 (81 direct vs 87 anchored). Maximum outlier lengths show mixed patterns, decreasing from 741 to 680 words for direct generation while increasing substantially from 580 to 759 words for anchored generation.

These results collectively demonstrate that higher temperatures yield longer, more variable responses with increased extremes, though anchoring partially mitigates these effects.

4.5 GEV Validation

Note: All analyses from this subsection onward use temperature=0.7 with direct prompts.

Figure 1 shows excellent GEV fit ($R^2_{\text{CDF}} = 0.998$) for response lengths (parameters: $c = -0.441$, $\mu = 26.2$, $\sigma = 14.2$). Tail deviations motivate our hybrid approach (Section 4.7).

Table 3: GEV Parameters by Temperature and Generation Method

Note: 95% confidence intervals in brackets. Sample sizes: n = 6,945 (T=0), n = 7,356 (T=0.7).

Temperature = 0		
Parameter	**Direct**	**Anchored**
Shape (c)	$-0.411\,[-0.436,\ -0.386]$	$-0.355\,[-0.379,\ -0.331]$
Location (μ)	$25.60\,[25.2,\ 26.0]$	$24.39\,[24.0,\ 24.8]$
Scale (σ)	$13.56\,[13.1,\ 14.0]$	$12.10\,[11.7,\ 12.5]$

Temperature = 0.7		
Parameter	**Direct**	**Anchored**
Shape (c)	$-0.411\,[-0.438,\ -0.387]$	$-0.361\,[-0.383,\ -0.339]$
Location (μ)	$26.19\,[25.8,\ 26.6]$	$24.80\,[24.5,\ 25.1]$
Scale (σ)	$14.20\,[13.8,\ 14.6]$	$12.46\,[12.1,\ 12.8]$

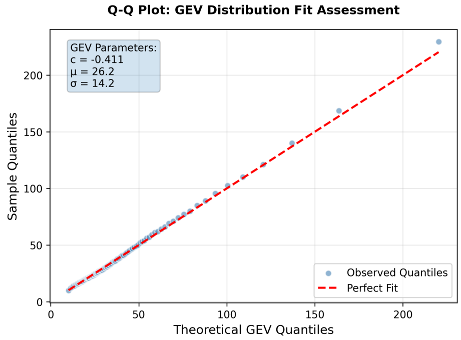


Figure 1: GEV Q-Q plot. Linearity confirms good fit for typical responses, with right-tail deviations visible.

4.6 Cross-Model GEV Validation

The Generalized Extreme Value (GEV) distribution demonstrated robust fit across all tested architectures, with GPT-4o and three additional open-weight models consistently exhibiting Weibull-type behavior ($c < 0$), as shown in Figure 2. Table 4 summarizes the maximum likelihood estimates of GEV parameters for each model.

Three key findings emerge from the cross-model comparison. First, the progression from Qwen3-8B to Qwen3-14B shows that larger models within the same family may develop less extreme length variation, supported by differences in shape parameters. Second, the universal quality of fit ($R^2_{\text{CDF}} \geq 0.994$ across all models) indicates the GEV distribution captures a fundamental property of transformer-based language generation. Third, the consistent Weibull-type behavior ($c < 0$) across architectures implies bounded output length distributions, with model scale affecting both the location and shape parameters.

This distributional regularity persists despite variations in model size (8B to 14B parameters) and architectural implementations, suggesting the GEV structure emerges from fundamental properties of the transformer mechanism rather than specific implementation choices. The parameter stability across conditions provides strong evidence for the GEV's role in characterizing autoregressive text generation.

4.7 GEV-GPD Hybrid Model

The hybrid model combines generalized extreme value (GEV) and generalized Pareto (GPD) distributions through a threshold-dependent formulation:

$$F(x) = \begin{cases} F_{\text{GEV}}(x) & x \leq u, \\ F_{\text{GEV}}(u) + [1 - F_{\text{GEV}}(u)] & \\ \qquad \times F_{\text{GPD}}(x-u) & x > u. \end{cases} \tag{3}$$

We systematically evaluated optimal thresholds u^* across the 85th to 99th percentiles (1% increments) by minimizing:

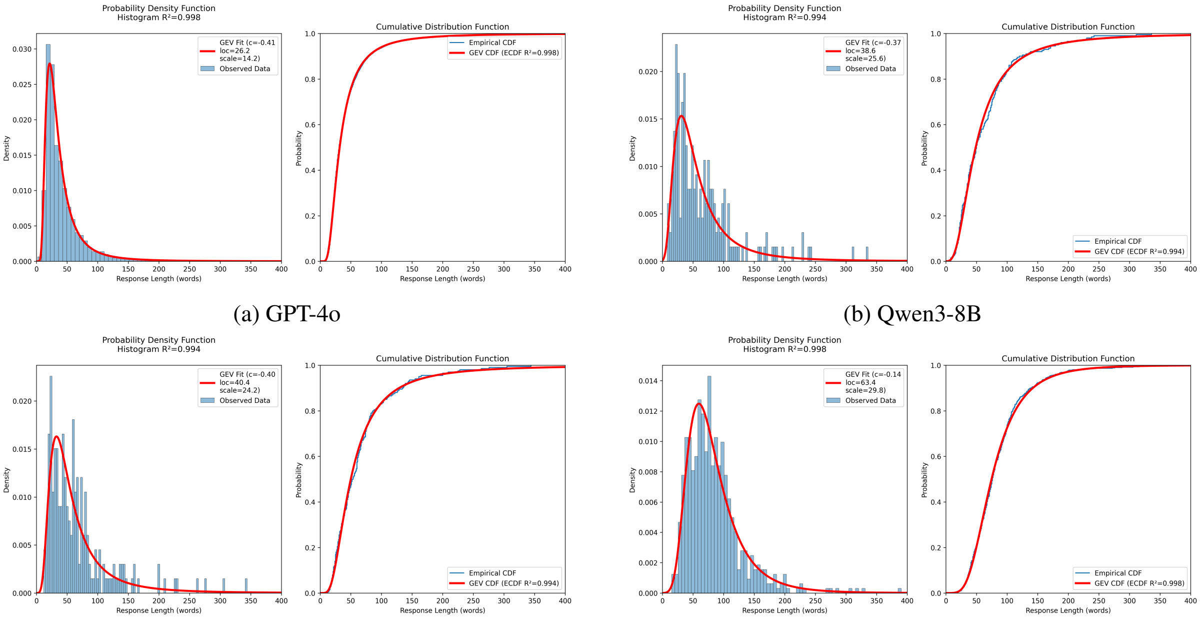
$$u^* = \underset{u \in \{Q_p\}_{p=0.85}^{0.99}}{\arg\min} \quad \mathcal{L}(u) \tag{4}$$

Two distinct loss functions were employed: (1) Pure MSE defined as $\mathcal{L}_{\text{MSE}}(u) = n^{-1} \sum_i (F_i - \hat{F}_i)^2$, and (2) Mixed objective combining multiple criteria through $0.6\mathcal{L}_{\text{MSE}} + 0.3|\xi| + 0.1\mathcal{I}$, which balances fit quality, tail properties, and model complexity.

The results demonstrate clear trade-offs between optimization approaches (Table 5). Pure mean squared error (MSE) optimization at the 95$^{\text{th}}$ percentile produces heavier tails ($\xi = 0.362$), while mixed optimization achieves superior tail behavior ($\xi = 0.183$) with comparable MSE performance

Table 4: Maximum Likelihood Estimates of GEV Parameters Across Models (Temperature = 0.7, Direct)

Model	**Shape** (c)	**Location** (μ)	**Scale** (σ)	**Sample Size** (n)
GPT-4o	-0.41	26.2	14.2	7356
Qwen3-8B	-0.40	40.4	24.2	200
Qwen3-14B	-0.37	38.6	25.6	200
DeepSeek-V3	-0.14	63.4	29.8	600



(a) GPT-4o

(b) Qwen3-8B

(c) Qwen3-14B

(d) DeepSeek-V3

Figure 2: GEV fits for response lengths. Each subfigure shows: (Left) Probability density functions with hybrid model (red) vs observed data (blue); (Right) Cumulative distribution functions comparing hybrid model (red) with empirical CDF (blue).

Table 5: Threshold Optimization Results

Metric	Pure MSE	Mixed
Threshold (u^*)	95%ile	97%ile
Tail index (ξ)	0.362	0.183
MSE ($\times 10^{-3}$)	0.92	0.93
R^2_{CDF}	0.9993	0.9995

Table 6: Hybrid Model Performance

Metric	**Value**
GEV Shape (c)	-0.411
GEV Location (μ)	26.2
GEV Scale (σ)	14.2
GPD Shape (ξ)	0.362
GPD Scale (β)	50.3
R^2_{CDF}	0.9993

(0.93 versus 0.92) and marginally better distributional fit (R^2_{CDF} = 0.9995 versus 0.9993). The difference in optimal thresholds reflects the inherent balance between overall fit quality and precise tail characterization.

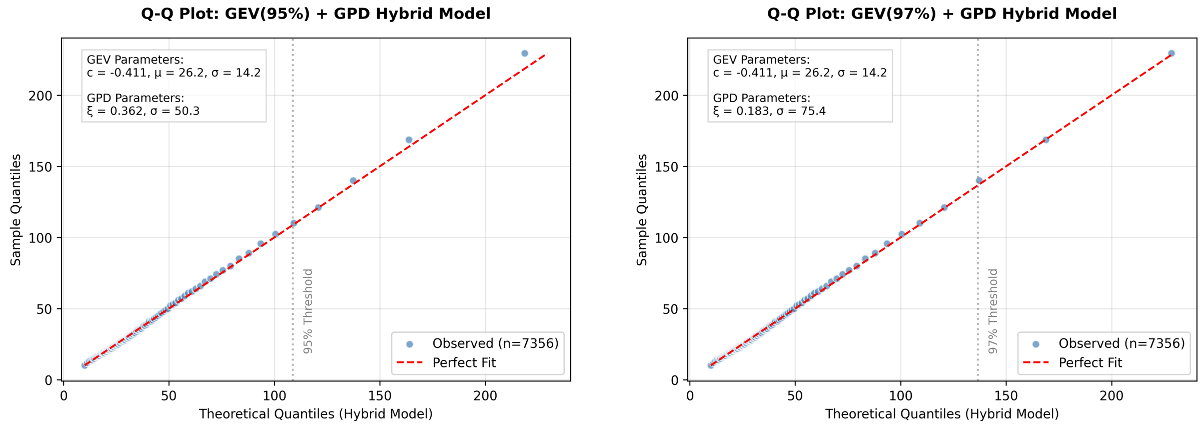
Figure 3 presents quantile-quantile (Q-Q) plots comparing our two optimal threshold candidates: the 95th percentile (pure MSE) and 97th percentile (mixed criterion) selections. Both demonstrate the hybrid model's robustness across optimization approaches, with visual inspection strongly favoring the 97th percentile threshold for extreme-value fit.

4.8 Final Hybrid Model

For subsequent analysis, we adopt the pure MSE criterion due to its simplicity and interpretability. The hybrid model integrates a Generalized Extreme Value (GEV) distribution for the body of the data and a Generalized Pareto Distribution (GPD) for the tail, with an optimized threshold of $u = 108.7$.

4.9 Model Fit Tests

As shown in Table 6, the GPD's positive shape parameter ($\xi = 0.362$) confirms heavier-tailed behavior beyond the 95th percentile (u = 108.7 words).



(a) 95th Percentile Threshold ($u = Q_{0.95}$)

(b) 97th Percentile Threshold ($u = Q_{0.97}$)

Figure 3: Comparison of hybrid model Q-Q plots for different thresholds (n=7,356). Both use GEV parameters ($c = -0.411$ (shape), $\mu = 26.2$ (location), $\sigma = 14.2$ (scale)) but differ in GPD fits: (a) 95% threshold yields $\xi = 0.362$, $\beta = 50.3$; (b) 97% threshold gives $\xi = 0.183$, $\beta = 75.4$. Vertical dashed lines mark transition points between distribution components.

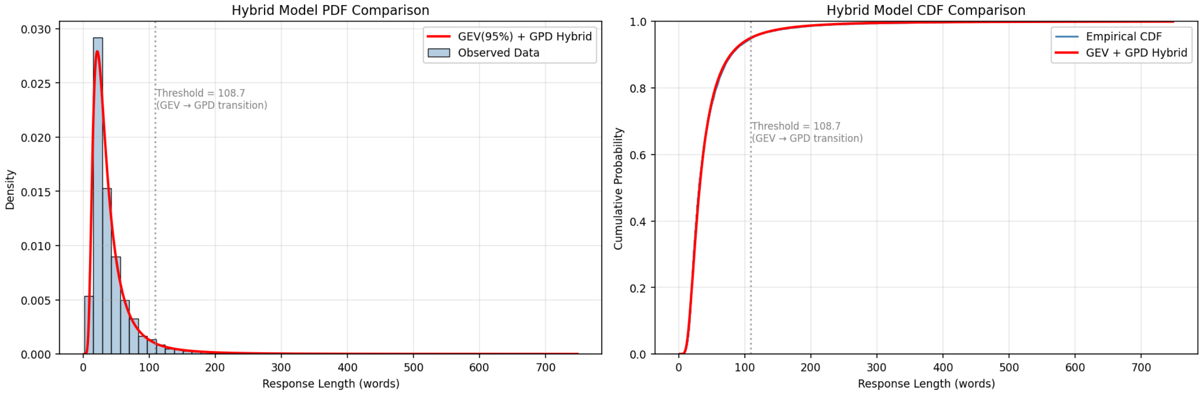


Figure 4: GEV-GPD Hybrid Model vs Empirical Distribution. Left: Probability density functions showing hybrid model (red) vs observed data (blue). Right: Cumulative distribution functions comparing hybrid model (red) with empirical CDF (blue).

Our analysis reveals three principal results. First, the model exhibits heavy-tail characteristics with a shape parameter $\xi = 0.362$. Second, threshold optimization identifies $u = 108.7$ words (95^{th} percentile) as the optimal transition point between distribution regions. Third, error analysis demonstrates region-specific patterns: the body of the distribution ($p < 0.95$, range: 10.1–100.4 words) shows low errors (MAE = 0.71, Max AE = 2.09); the transition region ($0.95 \leq p < 0.99$, range: 109.1–163.7 words) exhibits moderate errors (MAE = 2.16, Max AE = 4.90); while the extreme tail ($p \geq 0.99$, single observation at 218.7 words) displays substantially higher errors (MAE = Max AE = 10.80).

4.10 Input-Length Independence

The analysis revealed consistently weak correlations between input and response lengths across all conditions ($|r| < 0.1$). For direct generation, question length showed a negligible negative correlation with response length ($r = -0.08$). Notably, anchored generation cut this association in half ($r = -0.04$). Context length demonstrated virtually no linear relationship with response length in either condition ($r = -0.02$ for both direct and anchored generation). These results suggest that response length distributions are primarily determined by the LLM's generation process rather than input characteristics.

This preliminary analysis suggests the GEV/GPD structure is intrinsic to the LLM's generation process rather than inherited from input distributions.

5 Discussion

Our results reveal that LLM response lengths exhibit statistically robust structure, challenging the assumption that verbosity is merely incidental or model-specific. The consistent Weibull-type behavior captured by the GEV distribution across prompt variants and temperatures suggests that response generation is governed by stable underlying mechanisms. This aligns with the hypothesis that autoregressive models optimize for succinctness under bounded uncertainty, leading to inherent length regularization.

Anchoring Effects. Prompt anchoring exerts a measurable influence on response distributions. Specifically, it reduces both the location (μ) and scale (σ) parameters of the GEV fit, suggesting lower average length and less variability. Notably, the shape parameter (c) becomes less negative with anchoring, indicating slightly thicker tails—a counterintuitive result implying anchoring may shift some responses into more extreme regimes under stochastic decoding. These findings echo psychological theories of anchoring bias, where initial cues shape the perceived relevance or extent of follow-up content. While this study focused on the specific phrase “As previously stated," future work should explore how different types of anchoring phrases (e.g., summarization cues, discourse markers, or semantic constraints) might similarly influence verbosity patterns.

Temperature Sensitivity. The temperature effects reveal a dual pattern: while shape parameters (c) remain stable, the systematic increases in μ and σ at $T = 0.7$ suggest temperature primarily affects output dispersion rather than extreme-value mechanisms. Anchoring's consistent reduction of σ (10-15%) confirms its stabilizing role for typical responses, though its diminished tail protection at higher temperature (evidenced by heavier tails and increased outlier propensity) implies thermal modulation of anchoring efficacy. The borderline significant tail changes ($\Delta c = +0.006$) amidst confidence interval overlap may reflect either limited statistical power or a genuine architectural effect—a crucial distinction that future studies with larger samples should address.

Cross-Model GEV Analysis. The GEV distribution provides consistent fits across all tested models ($R^2_{\text{CDF}} \geq 0.994$), with parameter estimates revealing substantial variations (Table 4). The location parameters μ span a wide range from 26.2 (GPT-4o) to 63.4 (DeepSeek-V3), indicating fundamental differences in typical response lengths across architectures. While GPT-4o and Qwen3-8B/14B share similar Weibull-type behavior ($c \approx -0.4$), their μ values differ significantly (26.2 vs. 40.4/38.6). Notably, the larger Qwen3-14B shows both a less negative shape parameter ($c = -0.37$) and lower location parameter ($\mu = 38.6$) compared to Qwen3-8B ($c = -0.40$, $\mu = 40.4$). DeepSeek-V3 exhibits the most distinct profile with $c = -0.14$ and $\mu = 63.4$. These variations demonstrate that while the GEV framework is universally applicable, the specific parameter values capture important architectural differences in length generation patterns.

GEV-GPD Hybrid Advantages. While GEV captures central tendencies effectively ($R^2_{\text{CDF}} = 0.998$), it underfits the upper tail. Our hybrid model substantially improves this, increasing overall fit ($R^2_{\text{CDF}} = 0.9993$). The theoretical validation of compatibility between GEV and GPD shape parameters strengthens the statistical justification for this architecture and confirms the Weibull-type domain of attraction.

Implications for LLM Engineering. These findings open practical avenues for controlling verbosity in production environments. By adjusting temperature and anchoring strategies, developers can manipulate the shape and spread of output length distributions. Furthermore, the hybrid model enables anomaly detection in long responses (e.g., hallucinations, verbosity drift), offering a probabilistic safeguard mechanism.

Overall, this work positions extreme value theory as a foundational tool for modeling and managing LLM response behaviors, with implications spanning statistical modeling, prompt design, and safety.

6 Conclusion

This work establishes extreme value theory as a principled framework for modeling LLM verbosity patterns in question-answering tasks across architectures. Through analysis of 14,301 GPT-4o responses at different temperatures and cross-validation on Qwen and DeepSeek models at temperature 0.7, we demonstrate three key findings:

(1) temperature systematically increases both central tendency (μ) and dispersion (σ) in GPT-4o while preserving Weibull-type behavior, with shape parameters becoming more negative (e.g., from -0.355 to -0.361 for anchored generation) indicating heavier tails under stochastic generation, (2) the GEV distribution provides a consistent modeling framework that captures length distributions across diverse architectures despite substantial parameter variations, and (3) prompt anchoring reduces scale parameters by 10-15% across models while showing limited protection against temperature-induced tail changes. Our GEV-GPD hybrid model achieves superior tail fit ($R^2_{\text{CDF}} = 0.9993$) while maintaining architectural robustness, with threshold optimization enabling precise verbosity control in diverse deployment scenarios.

The cross-model results reveal important architectural insights: while all tested transformers exhibit Weibull-type behavior, larger models (Qwen3-14B vs 8B) show less extreme variation (shape parameter -0.37 vs -0.40), suggesting scale-dependent regularization of output lengths. The hybrid model's consistent performance across architectures (GPT-4o, Qwen3-8B/14B, DeepSeek-V3) confirms its generalizability, though parameter estimates reveal model-specific verbosity profiles - from GPT-4o's concise responses ($\mu = 26.2$) to DeepSeek-V3's more verbose outputs ($\mu = 63.4$).

The results also reveal that temperature affects different aspects of the length distribution distinctly - while increasing μ and σ for typical responses, it also amplifies extreme-value behavior through more negative shape parameters. This suggests separate thermal modulation mechanisms for bulk versus tail generation processes. The hybrid model's threshold selection method (optimal $u = 108.7$ words) provides a practical tool for managing these effects in production systems.

Future research directions include extending this framework to diverse tasks beyond question-answering (e.g., summarization, dialogue), investigating different prompting styles and their effects on length distributions, validating the approach across emerging architectures such as mixture-of-experts models, and developing temperature-aware adaptation methods for cross-model verbosity control. This work establishes a statistical foundation for understanding length generation patterns while providing methodologies for verbosity management in diverse LLM applications.

Data Statement

Data from HotpotQA (CC BY-SA 4.0). Full statement in Appendix.

Limitations

While empirically validated, several open questions remain. First, while our correlation analysis (Section 4.10) excludes linear dependence in input-length relationships, future work should explore non-linear dependencies via mutual information, conduct causal analysis through prompt-length interventions, and examine threshold effects like minimum context requirements. Second, the findings' generalizability is currently limited to English-language data; cross-linguistic validation is needed to assess cultural and typological dependencies. Third, the generalizability to non-QA tasks (summarization, dialogue) and few-shot scenarios remains unverified. Fourth, the anchoring effects are examined with a single phrase; investigating diverse anchoring strategies would strengthen the conclusions. Fifth, this study focuses on temperature sampling; other decoding strategies like Top-K and Nucleus sampling may yield different length distribution patterns and warrant separate investigation. Finally, while the consistent GEV patterns across models with varying sample sizes (200-7,356) support distributional robustness, larger cross-model samples could further strengthen statistical confidence, and the GEV structure requires validation across alternative architectures beyond standard Transformers, including both non-Transformer paradigms and Transformer variants like Mixture-of-Experts models.

Acknowledgements

We are grateful to Yufang Wang (Assistant Researcher at BNRist, Tsinghua University) for providing access to the OpenAI API. We thank Mengjiao Li for administrative support, which included facilitating access to the SiliconFlow API. This work was supported by these resources from Tsinghua University.

We thank OpenAI for the GPT-4o API and SiliconFlow for their services, which were instrumental for our experiments. We also thank the creators of the HotpotQA dataset and the anonymous reviewers for their insightful comments.

References

Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. 2022. Exploring length generalization in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 38546–38556.

Julian Arnold, Flemming Holtorf, Frank Schäfer, and Niels Lörch. 2024. Phase transitions in the output distribution of large language models. *arXiv preprint*.

Gábor Borbély and András Kornai. 2019. Sentence length. In *Proceedings of the 16th Meeting on the Mathematics of Language*, pages 114–125.

Bradley Butcher, Michael O'Keefe, and James Titchener. 2025. Precise length control for large language models. *Natural Language Processing Journal*, 11:100143.

Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024. The impact of reasoning step length on large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1830–1842.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. 2024. Same task, more tokens: the impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353.

Jiaxu Lou and Yifan Sun. 2024. Anchoring bias in large language models: An experimental study. *arXiv preprint*.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and LLM-generated news text. *Artificial Intelligence Review*, 57(10):265.

Sania Nayab, Giulio Rossolini, Marco Simoni, Andrea Saracino, Giorgio Buttazzo, Nicolamaria Manes, and Fabrizio Giacomelli. 2024. Concise thoughts: Impact of output length on LLM reasoning and cost. *arXiv preprint*.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? *arXiv preprint*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Matthew Renze. 2024. The effect of sampling temperature on problem solving in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7346–7356.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2023. A long way to go: Investigating length correlations in RLHF. *arXiv preprint*.

Natalia L. Tsizhmovska and Leonid M. Martyushev. 2021. Principle of least effort and sentence length in public speaking. *Entropy*, 23(8):1023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Pengkun Wang, Zhe Zhao, Haibin Wen, Fanfu Wang, Binwu Wang, Qingfu Zhang, and Yang Wang. 2024. LLM-AutoDA: Large language model-driven automatic data augmentation for long-tailed problems. In *Advances in Neural Information Processing Systems*, volume 37, pages 64915–64941.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Zangwei Zheng, Xiaozhe Ren, Fuzhao Xue, Yang Luo, Xin Jiang, and Yang You. 2023. Response length perception and sequence scheduling: An LLM-empowered LLM inference pipeline. In *Advances in Neural Information Processing Systems*, volume 36, pages 65517–65530.

Xin Zhou, Kisub Kim, Bowen Xu, Jiakun Liu, Dong-Gyun Han, and David Lo. 2023. The devil is in the tails: How long-tailed code distributions impact large language models. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 40–52.

A Data Statement

The experimental data in this work derives from the HotpotQA dataset (Yang et al., 2018), licensed under Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0). HotpotQA contains 112,779 English question-answer pairs collected from Wikipedia, covering diverse domains including history, science, and culture. Each question requires multi-hop reasoning with annotated supporting facts, though demographic information about annotators is not available.

While HotpotQA was originally designed for multi-hop question answering research, our repurposing for LLM response length analysis constitutes a valid research use under the license terms. Our derived GEV-GPD model specifically analyzes whitespace-delimited word counts in English LLM responses, with applicability to transformer-based models (tested on GPT-4o, Qwen, and DeepSeek architectures). The model assumes stationary length distributions and may not generalize to character-level or multilingual settings.

All derived annotations will be shared under the same CC BY-SA 4.0 license with research-only restrictions. Our implementation code will be released under the MIT License, including documentation of model assumptions and usage examples.

The HotpotQA dataset is derived from Wikipedia. While we did not independently verify content due to the dataset's scale and established academic usage, three factors mitigate risks: (1) Wikipedia's public editing policies inherently filter explicit PII, and (2) our analysis exclusively used whitespace-delimited word counts which discard raw text semantics.