# Factual Inconsistency in Data-to-Text Generation Scales Exponentially with LLM Size: A Statistical Validation

Anonymous ACL submission

## Abstract

Monitoring factual inconsistency is essential for ensuring trustworthiness in data-to-text generation (D2T). While large language models (LLMs) have demonstrated exceptional performance across various D2T tasks, previous studies on scaling laws have primarily focused on generalization error through power law scaling to LLM size (i.e., the number of model parameters). However, no research has examined the impact of LLM size on factual inconsistency in D2T. In this paper, we investigate how factual inconsistency in D2T scales with LLM size by exploring two scaling laws: power law and exponential scaling. To rigorously evaluate and compare these scaling laws, we employ a statistical validation framework consisting of three key stages: predictive performance estimation, goodness-of-fit assessment, and comparative analysis. For a comprehensive empirical study, we analyze three popular LLM families across five D2T datasets, measuring factual inconsistency inversely using four state-of-the-art consistency metrics. Our findings, based on exhaustive empirical results and validated through our framework, reveal that, contrary to the widely assumed power law scaling, factual inconsistency in D2T follows an exponential scaling with LLM size.

## 1 Introduction

Data-to-text (D2T) generation (Lin et al., 2024; Li et al., 2024) converts semi-structured data (e.g., tables) into natural language, with applications in conversation systems, automated journalism, and other fields. A key challenge in D2T is factual inconsistency (Li et al., 2022; Huang et al., 2023)—when generated text fails to entails with input facts—leading to hallucinations that undermine trust in D2T models (Figure 1). Therefore, it is essential to monitor and mitigate factual inconsistency in order to construct trustworthy D2T models.

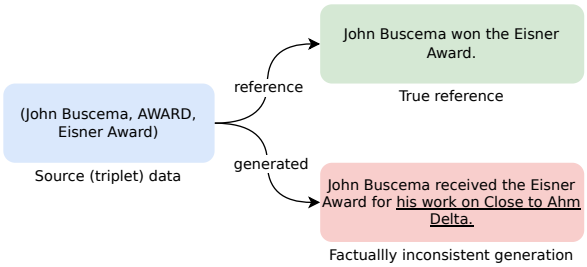


Figure 1: Example of data-to-text generation from the DART dataset, with a factually inconsistent output from the Pythia-1.4B model.

Large language models (LLMs) have achieved remarkable success in D2T, primarily due to their massive model sizes (parameter counts) and training on vast text corpora (Lorandi and Belz, 2024). Several studies shows that LLMs often adhere to scaling laws, typically power laws, governing generalization error or perplexity in relation to model size (Kaplan et al., 2020; Hoffmann et al., 2022). These scaling laws play a crucial role in predicting model performance, guiding hyperparameter tuning, estimating computational costs, and optimizing resource allocation (Hendrycks, forthcoming; Zhang et al., 2024). Existing LLM scaling laws in D2T focus on generalization loss or test perplexity (Bahri et al., 2024), overlooking factual inconsistency. Understanding how factual inconsistency scales with LLM model size can help researchers and practitioners optimize model selection and enhance trustworthiness in D2T, highlighting a key research gap.

In this paper, we address the research gap by examining scaling laws for factual inconsistency in D2T with respect to LLM size. Unlike prior studies that focus solely on power law scaling, we explore both power law and exponential scaling with a rigorous three-stage statistical validation framework. This framework comprises three key stages: predictive performance estimation (evaluating Huber loss on held-out data), goodness-of-fit assessment

(using an F-test to measure goodness-of-fit), and comparative analysis (utilizing Vuong's likelihood-ratio test to compare power law and exponential scaling). By integrating rigorous statistical validation, we ensure more reliable and robust insights, particularly in data-limited settings. Our study spans three widely used LLM families—Pythia, OPT, and BLOOM—and five well-established D2T datasets: E2E, ViGGO, WebNLG, DART, and WikiTableText. Factual inconsistency is quantified as the inverse of factual consistency, measured using four state-of-the-art metrics—ALIGNSCORE, QAFACTEVAL, SUMMAC-CONV, and UNIEVAL-FACT—which strongly correlate with human judgments. Our findings, validated through extensive empirical analysis and the rigorous validation framework, reveal that factual inconsistency in D2T follows exponential scaling with LLM size rather than power law scaling.

## 2 Related Work

**Data-to-Text Generation (D2T) and Factual Inconsistency.** Data-to-Text generation (D2T)(Lin et al., 2024) aims to transform non-textual, semi-structured data—such as tables, graphs, or slot-value pairs (meaning representation, MR)—into human-readable text. It can be categorized into three types based on source representation: graph-to-text(Gardent et al., 2017; Nan et al., 2021), table-to-text (Bao et al., 2018), and meaning representation (MR)-to-text (Novikova et al., 2017; Juraska et al., 2019). Recently, LLMs have become foundational models for D2T due to their extensive pre-training on large text datasets (Zhang et al., 2022) and their high model capacity (Scao et al., 2022). Moreover, with parameter-efficient fine-tuning techniques (Dettmers et al., 2023) and prompt-based learning (Lester et al., 2021), LLMs have gained widespread popularity for D2T tasks (Raffel et al., 2020; Lewis et al., 2020; Scao et al., 2022), often outperforming earlier models in generation quality and overall performance (Ge et al., 2023). In D2T, LLMs are often prone to generating factually inconsistent text, presenting a key research challenge. Factual inconsistency, defined as the lack of factual entailment between generated text and input data, contributes to hallucinations and undermines model reliability. Evaluation methods include human assessment (gold standard but costly) and automatic metrics (scalable but debated). Recently, trained automatic metrics (Fabbri et al., 2022; Zha et al., 2023) have shown strong correlations with human judgments, making them promising for factual inconsistency evaluation.

**Scaling Law for LLM.** Scaling laws for LLMs describe how their performance scales with key factors such as model size (number of parameters) and training data size. Hestness et al. (2017) demonstrated that deep language models follow a power law scaling, laying the foundation for scaling law research. Kaplan et al. (2020) expanded this by systematically analyzing model size, data size, and computational efficiency, reinforcing the dominance of power law scaling in LLM performance. As research on scaling laws has expanded, various studies have explored their applications across different task domains, including close-ended text generation (Bansal et al., 2022) and open-ended text generation (Kaplan et al., 2020). Recent investigations have further examined scaling in diverse paradigms, such as sparse modeling (Frantar et al., 2024) and parameter-efficient fine-tuning (Zhang et al., 2024). Additionally, joint scaling laws—such as additive and multiplicative formulations—are gaining prominence in multi-factor scaling setups (Hoffmann et al., 2022; Zhang et al., 2024). Scaling laws offer several key advantages, including optimizing hyperparameter tuning (Hendrycks, forthcoming), estimating training costs (Hägele et al., 2024), and setting realistic expectations for model performance (Hoffmann et al., 2022). A recent study by Bahri et al. (2024) further reinforces the theoretical foundations of scaling laws.

## 3 Scaling Law Models and Training

Moving beyond existing studies, we formulate the scaling law for factual inconsistency in D2T concerning LLM size by considering two models—power law scaling (following a power law function) and exponential scaling (following an exponential function). The power law scaling model ($\mathcal{M}_{pow}$) is defined as follows:

$$\mathcal{M}_{pow} : f(x) = \begin{cases} Ax^{\alpha} + B & x \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Where $A$ and $B$ are case-specific parameters, $\alpha$ is the power law exponent, $x$ represents LLM size, and $f(x)$ denotes factual inconsistency.

Similarly, the exponential scaling model ($\mathcal{M}_{exp}$) is defined as follows:

$$\mathcal{M}_{exp} : F(x) = \begin{cases} Ce^{\beta x} + D & x \geq 0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Where $C$ and $D$ are case-specific parameters, $\beta$ is the exponential scaling rate, $x$ represents LLM size, and $f(x)$ denotes factual inconsistency.

The parameters of both models are estimated using maximum likelihood estimation (MLE) on the factual inconsistency score data $\mathcal{D}$, optimized through the standard Huber loss ($\delta = 1$), denoted as $\mathcal{L}_{\text{Huber}}$, due to its robust estimation capability.

$$\hat{A}, \hat{B}, \hat{\alpha} \leftarrow \text{MLE}(M_{power}, \mathcal{D}, \mathcal{L}_{\text{Huber}}) \quad (3)$$

$$\hat{C}, \hat{D}, \hat{\beta} \leftarrow \text{MLE}(M_{exp}, \mathcal{D}, \mathcal{L}_{\text{Huber}}) \quad (4)$$

## 4 Statistical Validation Framework

To empirically study the two scaling models under limited data, we employ a structured three-stage statistical validation, consisting of predictive performance estimation, goodness-of-fit assessment, and comparative analysis, as detailed below.

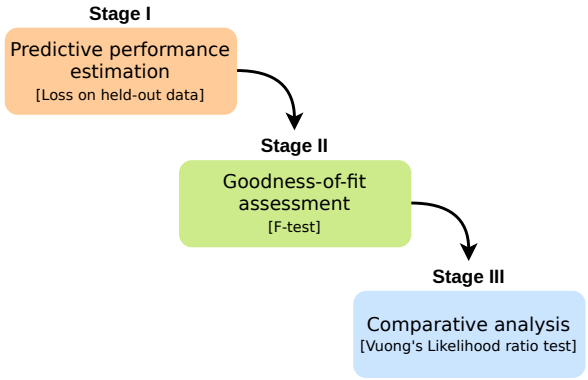


Figure 2: All three stages of our statistical validation framework.

- **Stage I: Predictive performance estimation.** This validation stage ensures how well the scaling laws generalize in terms of their predictive ability on unseen data. To achieve this, we evaluate the scaling laws on held-out data using Huber loss. Given the limited data availability, we employ five-fold cross-validation for predictive performance assessment.

- **Stage II: Goodness-of-fit assessment.** Predictive performance alone is not sufficient to validate a scaling law; assessing its goodness-of-fit is also crucial for its acceptance. Therefore, in this stage, we evaluate the goodness-of-fit of the scaling law models using a goodness-of-fit test—specifically, an F-test for regression (Siegel, 2016). The test statistic for the F-test is calculated as follows:

$$F_{\text{stat}} = \frac{SSR_{\mathfrak{R}} - SSR_{\mathfrak{E}}}{df_{\mathfrak{R}} - df_{\mathfrak{E}}} \times \frac{SSR_{\mathfrak{E}}}{df_{\mathfrak{E}}} \quad (5)$$

$$F_{\text{stat}} \sim \text{F-distribution}(x) \quad (6)$$

Here, $SSR_{\mathfrak{R}}$ and $SSR_{\mathfrak{E}}$ represent the sum of squared residuals for the reduced and exact models, respectively. Similarly, $df_{\mathfrak{R}}$ and $df_{\mathfrak{E}}$ denote the degrees of freedom for the reduced and exact models, respectively. We consider our scaling models ($\mathcal{M}_{pow}$ and $\mathcal{M}_{exp}$) as exact models, while the reduced model is represented by a simple mean-response model. Since the F-test applies only to linear regression models, we use a log transformation to convert our scaling models into their linear forms. We perform the F-test with a significance level of $p < 0.05$, which is often considered a moderate range. If both scaling models qualify in the goodness-of-fit assessment, we proceed to Stage III.

- **Stage III: Comparative analysis.** In this final stage of validation, we compare the two scaling law models, $\mathcal{M}pow$ and $\mathcal{M}exp$, through hypothesis testing to determine which better explains the data. Since power law and exponential scaling models are not nested hypotheses, the standard likelihood-ratio test is not applicable. Instead, we employ Vuong's likelihood-ratio test for comparison. The test statistic for Vuong's likelihood-ratio test is computed as follows:

$$V_{\text{stat}} = \frac{\sqrt{n} \cdot \text{mean}(d)}{\sqrt{\text{Var}(d)}} \quad (7)$$

$$V_{\text{stat}} \sim \text{normal}(0, 1) \quad (8)$$

Where $n$ represents the sample size, and $d$ denotes the $n$-sized sample of the log-likelihood differences between the two scaling law models. We conduct Vuong's likelihood ratio test at a stringent significance level of $p < 0.005$ to provide highly compelling evidence for our conclusion.

**Assumptions verification.** In the second and third stages of our validation framework, we incorporate

the F-test and Vuong's likelihood-ratio test. Both tests rely, directly or indirectly, on the assumption that the residuals of our scaling law models, $\mathcal{M}_{pow}$ and $\mathcal{M}_{exp}$, follow a normal distribution. Therefore, validating this assumption is essential. To assess the normality of residuals for both $\mathcal{M}_{pow}$ and $\mathcal{M}_{exp}$, we employ the Shapiro–Wilk test in our experiments.

## 5 Experiment Setup

### 5.1 Dataset

We utilize five well-known D2T datasets, covering three major D2T types: DART and WebNLG for graph-to-text, WikiTableText for table-to-text, and E2E and ViGGO for MR-to-text. All datasets are sourced from (Kasner et al., 2023) and (Wolf et al., 2020). The E2E dataset (Novikova et al., 2017; Dusek et al., 2018) contains over 37K MR-text pairs from the restaurant domain, with an average text length of approximately 21 words. ViGGO (Juraska et al., 2018) includes 7K MR-to-text instances spanning nine dialogue acts in the video game domain, with an average text length of around 14 words. Both E2E and ViGGO are closed-domain datasets. WikiTableText (Bao et al., 2018) is an open-domain D2T dataset consisting of approximately 13K table-to-text pairs extracted from Wikipedia tables. DART (Nan et al., 2021) contains nearly 70K knowledge graph triplets, with an average text length of 34 words. WebNLG (Gardent et al., 2017) focuses on RDF-to-text generation, comprising around 38K samples with an average text length of 30 words. Both DART and WebNLG are open-domain datasets.

| family | model counts | parameters of each models |
|---|---|---|
| Pythia | 8 | 70M, 160M, 410M, 1B, 1.4B, 2.8B, 6.9B, 12B |
| OPT | 6 | 130M, 350M, 1.3B, 2.7B, 6.7B, 13B |
| BLOOM | 5 | 0.56M, 1.1B, 1.7B, 3B, 7B |

Table 1: The three LLM families along with their models and corresponding sizes (M=million, B=billion).

### 5.2 Models

We incorporate three widely used LLM families in our experiments: Pythia, OPT, and BLOOM. Examining multiple families offers broader insights than focusing on a single family. Pythia is a suite of eight decoder-only autoregressive models (70M–12B parameters), following a GPT-style (Brown et al., 2020) architecture with flash attention. All Pythia models are trained on the Pile dataset in the same order. We consider OPT (Zhang et al., 2022), which includes six models, each being a decoder-only transformer (130M–13B parameters), trained on datasets including Reddit, the Pile, and RoBERTa, following the training details outlined in (Brown et al., 2020). BLOOM (Scao et al., 2022) is another decoder-only LLM trained on the ROOT dataset. We include six BLOOM models in our study. A summary of these LLM families, their models, and corresponding sizes is provided in Table 1.

### 5.3 Fine-tuning for D2T

All LLMs are fine-tuned separately on each of the five D2T datasets. Given the large model sizes, full fine-tuning is computationally expensive. To mitigate this, we use QLoRA (Quantized Low-Rank Adapter) (Dettmers et al., 2023), a parameter-efficient fine-tuning method, with a learning rate of 1.00e-04 and $r = 16$ (reduced rank) for the attention module.

### 5.4 Quantification for Factual Inconsistency

We define factual inconsistency as the inverse of factual consistency, computed as $1 - z$ (where $z$ is the factual consistency score ranging from 0 to 1). To evaluate factual inconsistency in LLMs for D2T, we use four state-of-the-art automatic metrics that strongly correlate with human judgments: ALIGNSCORE (measures consistency through information alignment) (Zha et al., 2023), QAFACTEVAL (assesses consistency via question generation and answering) (Fabbri et al., 2022), SUMMAC-CONV (leverages natural language inference) (Laban et al., 2022), and UNIEVAL-FACT (employs unified training) (Zhong et al., 2022). Given their high agreement with human annotations, these metrics provide a strong foundation for our study.

### 5.5 Decoding Strategies

Given the importance of decoding strategies in D2T, we include both deterministic (greedy and beam search) and stochastic (nucleus and top-k sampling) methods for a comprehensive analysis. However, due to space constraints, we present here the results using nucleus sampling, while results for other strategies are provided in the appendix (Appendix A).

## 6 Results

This section presents our empirical results and the validation framework's evaluation of factual incon-

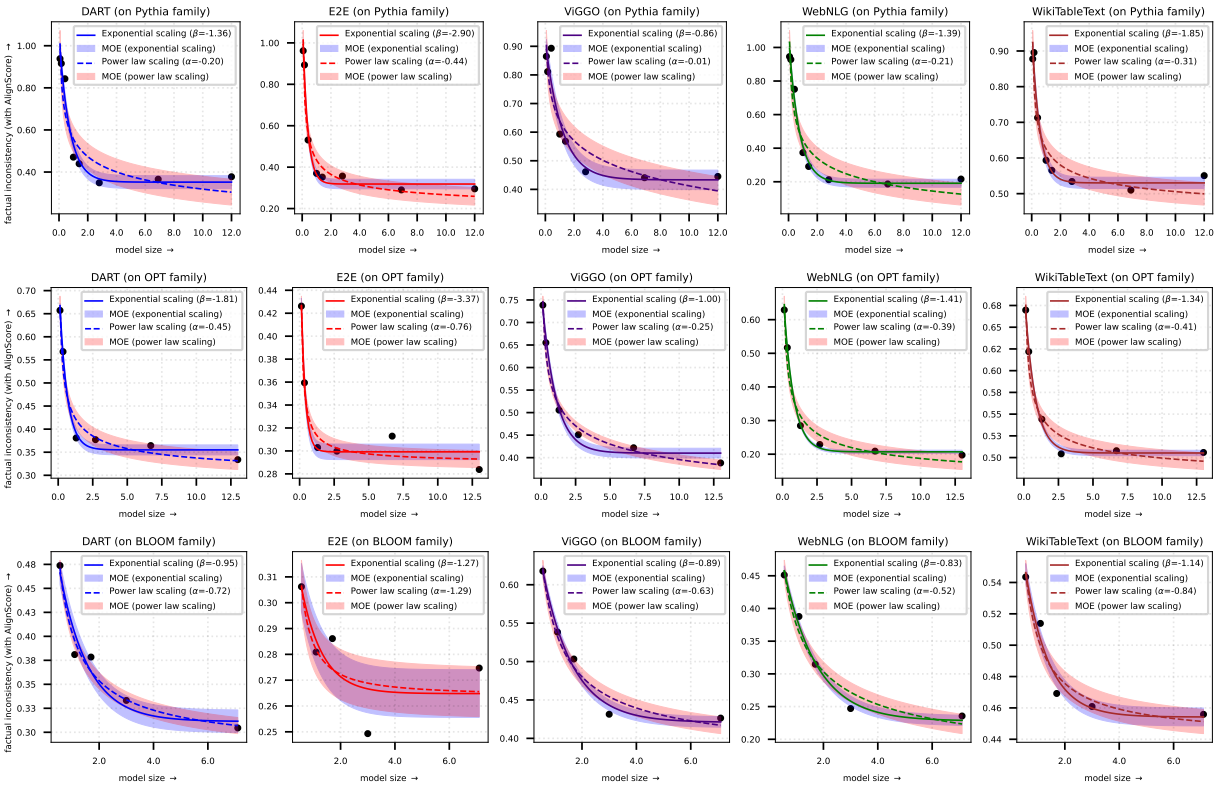


Figure 3: Visualization of exponential and power law scaling of factual inconsistency (ALIGNSCORE) across datasets and LLM families, with margin of error (MOE) and 95% confidence intervals on residuals.

| | | Results of stage I | | | | | Results of stage II and III | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LLM family | Scaling law | DART | E2E | ViGGO | WebNLG | WikiTableText | DART | E2E | ViGGO | WebNLG | WikiTableText |
| BLOOM | Exponential | 6.56e-04 | 2.49e-03 | 2.13e-04 | 1.33e-03 | 4.42e-03 | ✗ | ✗ | ✓(✿) | ✓ | ✓ |
| | Power law | 8.42e-04 | 3.07e-04 | 2.44e-03 | 1.30e-02 | 2.50e-02 | ✓ | ✗ | ✓ | ✗ | ✗ |
| OPT | Exponential | 4.15e-04 | 9.15e-04 | 4.16e-04 | 1.30e-04 | 2.82e+03 | ✓(✿) | ✓(✿) | ✓ | ✓(✿) | ✓(✿) |
| | Power law | 2.27e-03 | 1.38e-03 | 1.94e-03 | 2.08e-02 | 6.37e+01 | ✓ | ✓ | ✓(✿) | ✓ | ✓ |
| Pythia | Exponential | 1.89e-03 | 4.33e-02 | 2.20e-03 | 1.86e-03 | 2.70e-03 | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) |
| | Power law | 1.47e-02 | 2.71e-01 | 1.58e-02 | 5.74e-03 | 1.83e-02 | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 2: Results of the validation framework (all three stages) for both scaling laws of factual inconsistency (ALIGNSCORE). High held-out losses (Stage I) are highlighted in red. ✓/✗ indicates pass/fail (also marked in red) in the goodness-of-fit test (Stage II), while ✿ denotes the effective scaling law from Stage III.

sistency scaling in D2T based on the two scaling laws. We report findings from the standpoint of automatic metrics used to assess factual inconsistency.

### 6.1 Findings from ALIGNSCORE

Figure 3 illustrates both fitted scaling laws for factual inconsistency measured by ALIGNSCORE. Exponential scaling generally outperforms power law scaling, except for minor deviations, such as larger margins of error (MOE) in the BLOOM family for the E2E dataset. Table 2 presents our statistical validation results. Note that we successfully verified the normality assumption on residuals before applying our validation framework. Most Huber loss values are low, confirming the predictive reliability of both scaling laws. However, in several cases within the BLOOM family, one or both scaling models fail the goodness-of-fit test, while both pass for OPT and Pythia. This confirms that reliable predictive performance alone is not always sufficient to pass the goodness-of-fit test. Stage III results (Table 2) indicate that exponential scaling is generally preferred over power law scaling, except for the ViGGO dataset with OPT (marked in yellow). The stringent significance level of the final test further strengthens our conclusion, providing compelling evidence that factual inconsistency in D2T, when measured using ALIGNSCORE, consistently follows an exponential scaling pattern with respect to LLM size, rather than a power law trend.

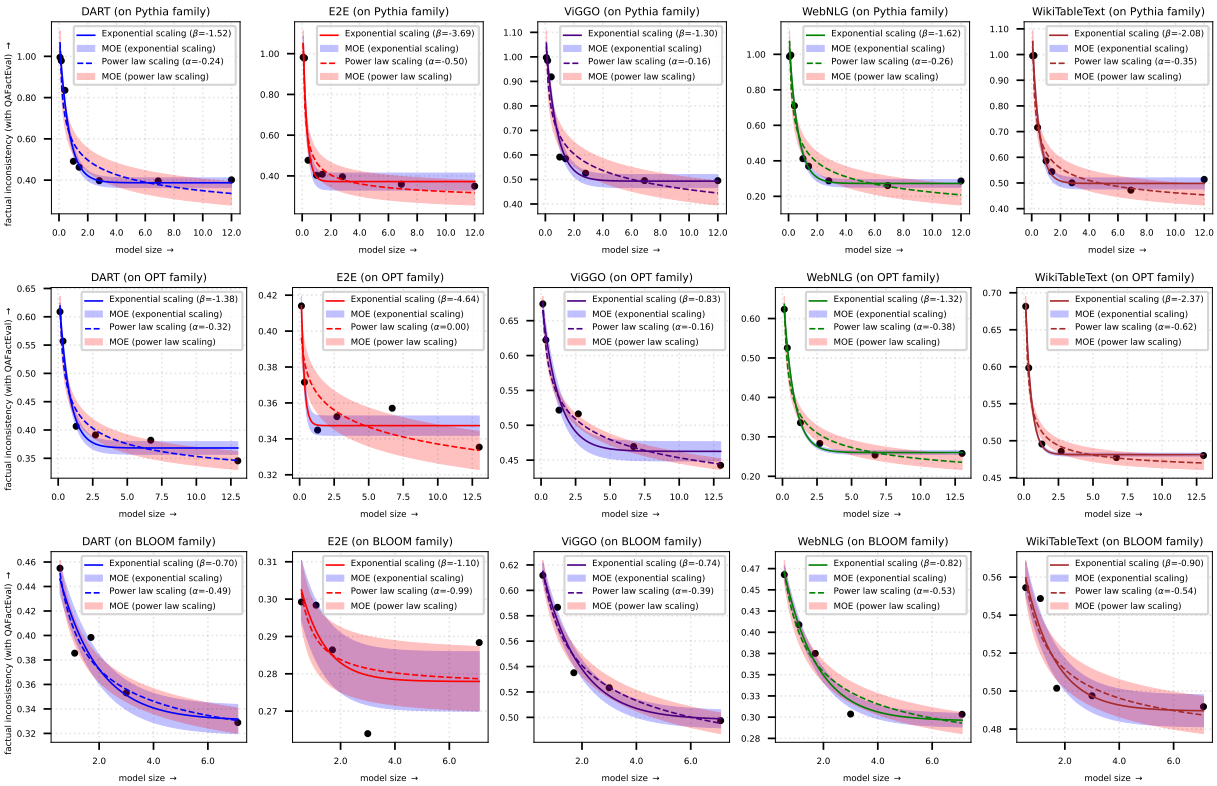


Figure 4: Visualization of exponential and power law scaling of factual inconsistency (QAFACTEVAL) across datasets and LLM families, with margin of error (MOE) and 95% confidence intervals on residuals.

| | | Results of stage I | | | | | Results of stage II and III | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LLM family | Scaling law | DART | E2E | ViGGO | WebNLG | WikiTableText | DART | E2E | ViGGO | WebNLG | WikiTableText |
| BLOOM | Exponential | 6.26e-04 | 1.73e-03 | 3.10e-04 | 6.15e-04 | 1.59e-02 | ✗ | ✗ | ✓(✿) | ✓ | ✗ |
| | Power law | 4.11e-04 | 4.59e-03 | 2.04e-03 | 1.87e-03 | 3.23e-01 | ✗ | ✗ | ✓ | ✗ | ✗ |
| OPT | Exponential | 3.57e-03 | 1.28e+01 | 6.56e-04 | 2.12e-03 | 1.67e-05 | ✓(✿) | ✓ | ✓(✿) | ✓(✿) | ✓(✿) |
| | Power law | 1.41e-02 | 2.81e-04 | 4.43e-04 | 8.45e-03 | 5.99e-03 | ✓ | ✗ | ✓ | ✓ | ✓ |
| Pythia | Exponential | 1.89e-03 | 4.33e-02 | 2.20e-03 | 1.86e-03 | 2.70e-03 | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) |
| | Power law | 1.47e-02 | 2.71e-01 | 1.58e-02 | 5.74e-03 | 1.83e-02 | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 3: Results of the validation framework (all three stages) for exponential and power law scaling of factual inconsistency (QAFACTEVAL). High held-out losses (Stage I) are highlighted in red. ✓/✗ indicates pass/fail (also marked in red) in the goodness-of-fit test (Stage II), while ✿ denotes the effective scaling law from Stage III.

### 6.2 Findings from QAFACTEVAL

Figure 4 suggests that when factual inconsistency is measured using QAFACTEVAL, both scaling laws fit well across most datasets and LLM families. Power law scaling shows a larger margin of error (MOE) compared to exponential scaling, with extremely high MOE for OPT and BLOOM families on the E2E dataset. Here we also verified the normality assumption before applying our validation framework. Low losses in stage I results (Table 3) indicate strong predictive performance for both scaling laws. Stage II and III results (Table 3) show that both scaling law not qualify for goodness-of-fit in the BLOOM family, while both laws are qualified for goodness-of-fit in Pythia and OPT, with exponential scaling outperforming power law scaling. Thus, based on Pythia and OPT, we observe that exponential scaling appears more suitable when factual inconsistency is measured using QAFACTEVAL in D2T.

### 6.3 Findings from SUMMAC-CONV

Figure 5 shows both fitted scaling laws when factual inconsistency is measured using SUMMAC-CONV. In most cases, exponential scaling provides a better fit than power law scaling. Some datasets, like E2E, exhibit a slightly larger margin of error (MOE), particularly in the BLOOM model family, which can be considered a minor exception. From Table 4, we observe that, except for a few cases in WikiTableText, the Huber loss values re-

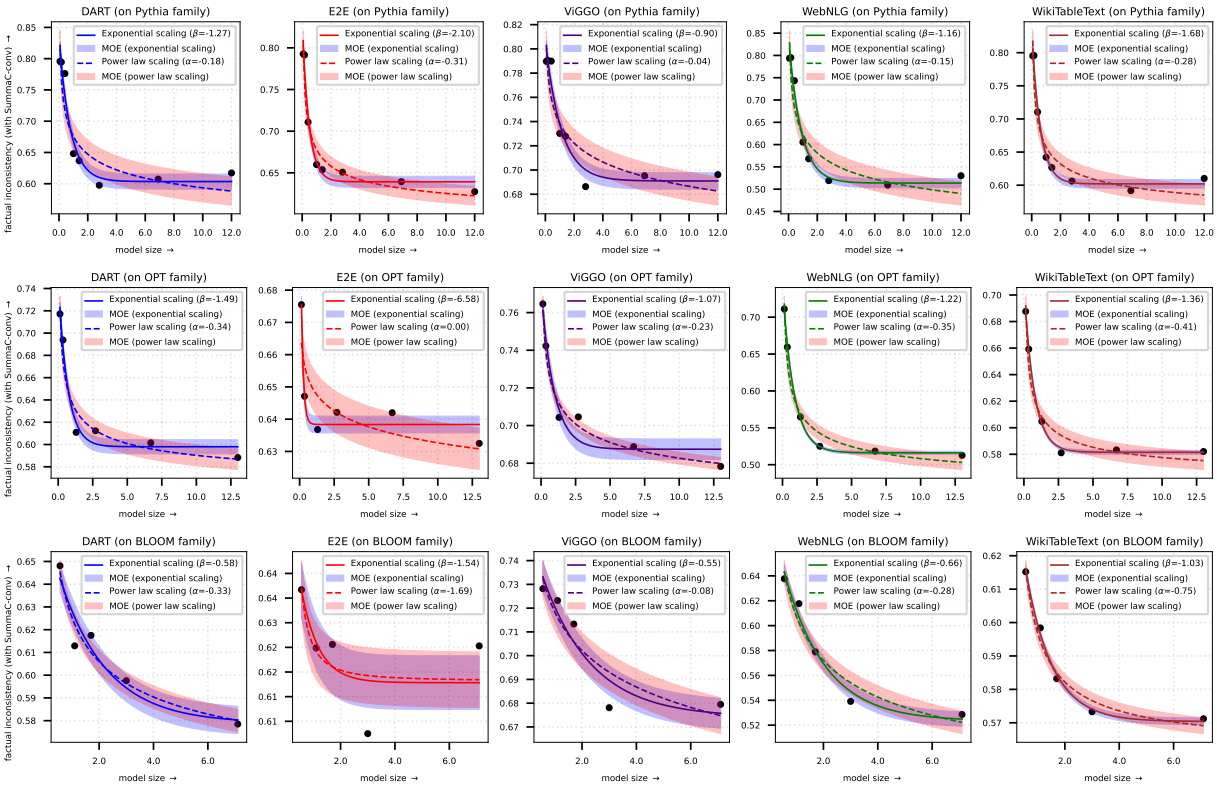


Figure 5: Visualization of exponential and power law scaling of factual inconsistency (SUMMAC-CONV) across datasets and LLM families, with margin of error (MOE) and 95% confidence intervals on residuals.

| | | Results of stage I | | | | | Results of stage II and III | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LLM family | Scaling law | DART | E2E | ViGGO | WebNLG | WikiTableText | DART | E2E | ViGGO | WebNLG | WikiTableText |
| BLOOM | Exponential | 1.70e-04 | 4.25e-04 | 7.60e-04 | 1.00e-03 | 1.94e-05 | ✗ | ✗ | ✗ | ✓ | ✓(✿) |
| | Power law | 7.00e-05 | 1.28e-04 | 4.27e-04 | 1.77e-03 | 9.60e-04 | ✗ | ✗ | ✗ | ✗ | ✓ |
| OPT | Exponential | 1.71e-04 | 9.87e-05 | 8.72e-05 | 2.83e-05 | 1.01e+03 | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) |
| | Power law | 1.60e-03 | 1.36e-04 | 9.33e-05 | 5.39e-03 | 5.48e+19 | ✓ | ✓ | ✓ | ✓ | ✓ |
| Pythia | Exponential | 4.17e-04 | 3.36e-04 | 5.00e-04 | 3.66e-04 | 1.18e-04 | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) |
| | Power law | 1.43e-03 | 1.79e-04 | 1.40e-02 | 2.23e-03 | 2.19e-02 | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 4: Results of the validation framework (all three stages) for exponential and power law scaling of factual inconsistency (SUMMAC-CONV). High held-out losses (Stage I) are highlighted in red. ✓/✗ indicates pass/fail (also marked in red) in the goodness-of-fit test (Stage II), while ✿ denotes the effective scaling law from Stage III.

main low, ensuring the predictive quality of both scaling laws. Additionally, both scaling law fails to qualify in several goodness-of-fit tests across multiple datasets in BLOOOM family. However, in the other two LLM families, where both scaling laws pass the test, exponential scaling consistently outperforms power law scaling.

### 6.4 Findings from UNIEVAL-FACT

Figure 6 illustrates how both scaling laws perform across all LLM families and D2T datasets when measuring factual inconsistency with UNIEVAL-FACT. Table 5 presents the validation framework results, consistently showing that exponential scaling captures factual consistency better than power law scaling. The only exception is WikiTableText in the BLOOM family (highlighted in yellow), where power law scaling surpasses exponential scaling.

## 7 Discussion

Our results demonstrate that when factual inconsistency is measured using the four automatic metrics, exponential scaling consistently outperforms power law scaling in most cases. While a few exceptions arise, particularly within the BLOOM family, we consider these anomalies to be outliers, likely due to the limited number of models in the family (only five). Beyond this, both scaling laws exhibit minimal margins of error across all plots, reinforcing their predictive reliability. The acceptance of exponential scaling for factual inconsistency in D2T

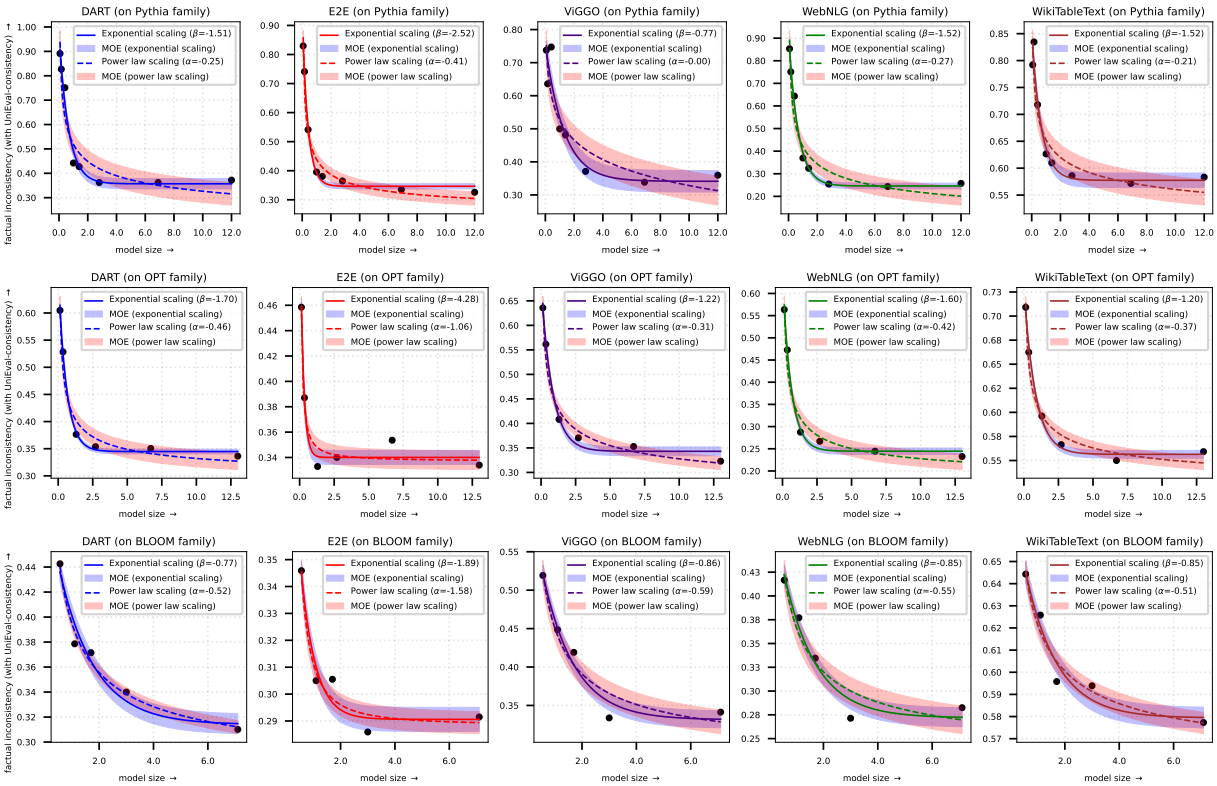


Figure 6: Visualization of exponential and power law scaling of factual inconsistency (UNIEVAL-FACT) across datasets and LLM families, with margin of error (MOE) and 95% confidence intervals on residuals.

| | | Results of stage I | | | | | Results of stage II and III | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LLM family | Scaling law | DART | E2E | ViGGO | WebNLG | WikiTableText | DART | E2E | ViGGO | WebNLG | WikiTableText |
| BLOOM | Exponential | 2.87E-04 | 1.16E-04 | 2.01E-04 | 3.66E-03 | 2.13E-04 | ✓(✿) | ✗ | ✓ | ✗ | ✓ |
| | Power law | 3.71E-04 | 6.11E-05 | 1.57E-03 | 8.66E-03 | 6.57E-04 | ✓ | ✗ | ✗ | ✗ | ✓(✿) |
| OPT | Exponential | 7.78E-05 | 6.45E-04 | 3.32E-03 | 1.62E-04 | 8.35E-05 | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) |
| | Power law | 6.25E-03 | 5.70E-04 | 3.93E-03 | 5.47E-03 | 3.18E-03 | ✓ | ✓ | ✓ | ✓ | ✓ |
| Pythia | Exponential | 6.34E-04 | 2.63E-04 | 2.08E-03 | 4.58E-04 | 4.54E-05 | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) |
| | Power law | 7.29E-03 | 8.28E-03 | 1.83E-01 | 1.13E-02 | 4.19E-03 | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 5: Results of the validation framework (all three stages) for exponential and power law scaling of factual inconsistency (UNIEVAL-FACT). High held-out losses (Stage I) are highlighted in red. ✓/✗ indicates pass/fail (also marked in red) in the goodness-of-fit test (Stage II), while ✿ denotes the effective scaling law from Stage III.

relative to LLM size suggests a rapid initial decline in factual inconsistency up to approximately 3–4 billion parameters, after which it stabilizes. Furthermore, in cases where exponential scaling does not show high margins of error, the scaling rate ($\beta$) consistently falls within the range of $-1.8$ to $-0.6$. Understanding this range of exponential scaling rates could be crucial for predicting the performance of LLMs in D2T tasks concerning factual inconsistency, providing valuable insights for future model development and evaluation.

## 8 Conclusion

This paper shows that factual inconsistency in D2T generally follows exponential scaling with respect to LLM size, rather than the commonly assumed power law scaling. Our findings are validated through a structured three-stage statistical framework, ensuring robustness in our conclusions. Moreover, we conduct a comprehensive empirical study using three major LLM families across five D2T datasets, measuring factual inconsistency inversely with four state-of-the-art consistency metrics. We believe these findings will help researchers and practitioners select appropriate model sizes to achieve specific levels of factual consistency. All results in this paper are based on a parameter-efficient fine-tuning approach (QLoRA) for LLMs. However, in-context learning and prompting strategies are not considered in this study, which we leave as future work.

## 9 Limitations

While our study provides a thorough empirical analysis of scaling laws for factual consistency in LLMs, validated through a structured three-stage framework, it is important to acknowledge the following limitations:

1. **Empirical basis without theoretical guarantee.** Our findings are entirely based on empirical observations, relying on the datasets and LLM families incorporated in this study. We do not provide a formal theoretical guarantee for the observed scaling behavior, making our conclusions inherently dependent on the data and models used.

2. **Non-universality of scaling law parameters.** Scaling law parameters are not universally applicable across different datasets, models, and task domains. While our results indicate a strong preference for exponential scaling, this does not guarantee that the same trend will persist across all datasets or model architectures, even when using the same set of parameters. Therefore, applying these scaling laws— including our own findings—requires careful consideration and validation within the specific context of use.

3. **Reliance on automated metrics without human evaluation.** In this study, factual inconsistency is estimated inversely from automated factual consistency metrics. While these metrics have demonstrated strong correlations with human judgments, we do not incorporate direct human evaluations of factual consistency. This remains a limitation of our work, though it presents a clear direction for future research to further validate and refine our findings with human annotation studies.

## References


Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. 2024. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121.

Yamini Bansal, Behrooz Ghorbani, Ankush Garg, Biao Zhang, Colin Cherry, Behnam Neyshabur, and Orhan Firat. 2022. Data scaling laws in NMT: the effect of noise and architecture. In *ICML*, volume 162, pages 1466–1482.

Junwei Bao, Duyu Tang, Nan Duan, Zhao Yan, Yuanhua Lv, Ming Zhou, and Tiejun Zhao. 2018. Table-to-text: Describing table region with natural language. In *Proceedings of the AAAI*, pages 5020–5027.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Proceedings of the NeurIPS*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Proceedings of the NeurIPS*.

Ondrej Dusek, Jekaterina Novikova, and Verena Rieser. 2018. Findings of the E2E NLG challenge. In *Proceedings of the INLG*, pages 322–328.

Alexander R. Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. Qafacteval: Improved qa-based factual consistency evaluation for summarization. In *Proceedings of the NAACL-HLT*, pages 2587–2601.

Elias Frantar, Carlos Riquelme Ruiz, Neil Houlsby, Dan Alistarh, and Utku Evci. 2024. Scaling laws for sparsely-connected foundation models. In *ICLR*.

Claire Gardent, Anastasia Shimorina, Shashi Narayan, and Laura Perez-Beltrachini. 2017. The webnlg challenge: Generating text from RDF data. In *Proceedings of the INLG*, pages 124–133.

Yingqiang Ge, Wenyue Hua, Kai Mei, Jianchao Ji, Juntao Tan, Shuyuan Xu, Zelong Li, and Yongfeng Zhang. 2023. Openagi: When LLM meets domain experts. In *Proceedings of the NeurIPS*.

Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna Ben Allal, Leandro von Werra, and Martin Jaggi. 2024. Scaling laws and compute-optimal training beyond fixed training durations. *CoRR*, abs/2405.18392.

Dan Hendrycks. forthcoming. Introduction to ai safety, ethics, and society. *Taylor & Francis*.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory F. Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. 2017. Deep learning scaling is predictable, empirically. *CoRR*, abs/1712.00409.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan

Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. Training compute-optimal large language models. *CoRR*, abs/2203.15556.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *CoRR*, abs/2311.05232.

Juraj Juraska, Kevin Bowden, and Marilyn A. Walker. 2019. Viggo: A video game corpus for data-to-text generation in open-domain conversation. In *Proceedings of the INLG*, pages 164–172.

Juraj Juraska, Panagiotis Karagiannis, Kevin Bowden, and Marilyn A. Walker. 2018. A deep ensemble model with slot alignment for sequence-to-sequence natural language generation. In *Proceedings of the NAACL-HLT*, pages 152–162.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Zdenek Kasner, Ekaterina Garanina, Ondrej Plátek, and Ondrej Dusek. 2023. Tabgenie: A toolkit for table-to-text generation. In *Proceedings of the ACL*, pages 444–455.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the EMNLP*, pages 3045–3059.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the ACL*, pages 7871–7880.

Shujie Li, Liang Li, Ruiying Geng, Min Yang, Binhua Li, Guanghu Yuan, Wanwei He, Shao Yuan, Can Ma, Fei Huang, et al. 2024. Unifying structured data as graph for data-to-text pre-training. *Transactions of the Association for Computational Linguistics*, 12:210–228.

Wei Li, Wenhao Wu, Moye Chen, Jiachen Liu, Xinyan Xiao, and Hua Wu. 2022. Faithfulness in natural language generation: A systematic survey of analysis, evaluation and optimization methods. *CoRR*, abs/2203.05227.

Yupian Lin, Tong Ruan, Jingping Liu, and Haofen Wang. 2024. A survey on neural data-to-text generation. *IEEE Transactions on Knowledge and Data Engineering*, 36(4):1431–1449.

Michela Lorandi and Anya Belz. 2024. High-quality data-to-text generation for severely under-resourced languages with out-of-the-box large language models. In *Proceedings of the EACL Findings*, pages 1451–1461.

Linyong Nan, Dragomir R. Radev, Rui Zhang, Amrit Rau, Abhinand Sivaprasad, Chiachun Hsieh, Xiangru Tang, Aadit Vyas, Neha Verma, Pranav Krishna, Yangxiaokang Liu, Nadia Irwanto, Jessica Pan, Faiaz Rahman, Ahmad Zaidi, Mutethia Mutuma, Yasin Tarabar, Ankit Gupta, Tao Yu, Yi Chern Tan, Xi Victoria Lin, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. 2021. DART: open-domain structured data record to text generation. In *Proceedings of the NAACL-HLT*, pages 432–447.

Jekaterina Novikova, Ondrej Dusek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the SIGDIAL*, pages 201–206.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Andrew F Siegel. 2016. *Practical Business Statistics*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the EMNLP*, pages 38–45.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with A unified alignment function. In *Proceedings of the ACL*, pages 11328–11348.

Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. 2024. When scaling meets LLM finetuning: The effect of data, model and finetuning method. In *ICLR*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the EMNLP*, pages 2023–2038.

# A Appendix

In the main paper, we discussed the crucial role of decoding strategies in data-to-text generation (D2T) and presented results based on nucleus sampling. Here, we extend our analysis by presenting empirical results from our validation framework for both power law and exponential scaling, using three additional decoding strategies—greedy, beam search, and top-k decoding. Among these, greedy and beam search are deterministic, while top-k decoding falls under stochastic methods. For our experiments, we use beam search with a beam size of 3 and top-k decoding with $k = 640$ (sample size).

## A.1 Discussion

Across all three decoding strategies, we consistently observe that **exponential scaling outperforms power law scaling** in nearly all cases for the Pythia and OPT LLM families across the four factual inconsistency metrics. While this trend is dominant, we identify a few noteworthy exceptions:

1. **Goodness-of-fit test failures in BLOOM and OPT.** In the BLOOM and OPT families, we frequently find cases where the scaling laws fail to qualify the goodness-of-fit test (Stage II), despite demonstrating low predictive loss in Stage I. This indicates that strong predictive performance alone is not always sufficient for a model to align well with the expected scaling trend. In other words, higher predictive performance does not necessarily imply goodness-of-fit.

2. **High margin of error in E2E dataset.** The margin of error (with a $95\%$ confidence interval) tends to be significantly higher in the E2E dataset, particularly for the BLOOM and OPT model families. This suggests a higher variance in factual inconsistency measurements, potentially due to dataset-specific characteristics or the way these models generalize.

3. **Aberrant behavior in E2E and ViGGO with deterministic decoding.** A particularly intriguing anomaly is observed in the E2E and ViGGO datasets (Figures 7 to 14), where factual inconsistency increases with LLM model size under deterministic decoding strategies (greedy search and beam search). This contradicts the general trend seen with stochastic decoding strategies (nucleus and top-k sampling), where inconsistency decreases with model size. We hypothesize that this aberrant behavior may be attributed to one or both of the following factors:

   - **Deterministic Decoding Bias.** Since greedy search and beam search select high-likelihood tokens, they might reinforce factual errors present in the training data rather than mitigating them.
   - **Closed-Domain Nature of E2E and ViGGO.** These datasets focus on highly structured, domain-specific content, which may lead to overfitting in larger models when using deterministic decoding.

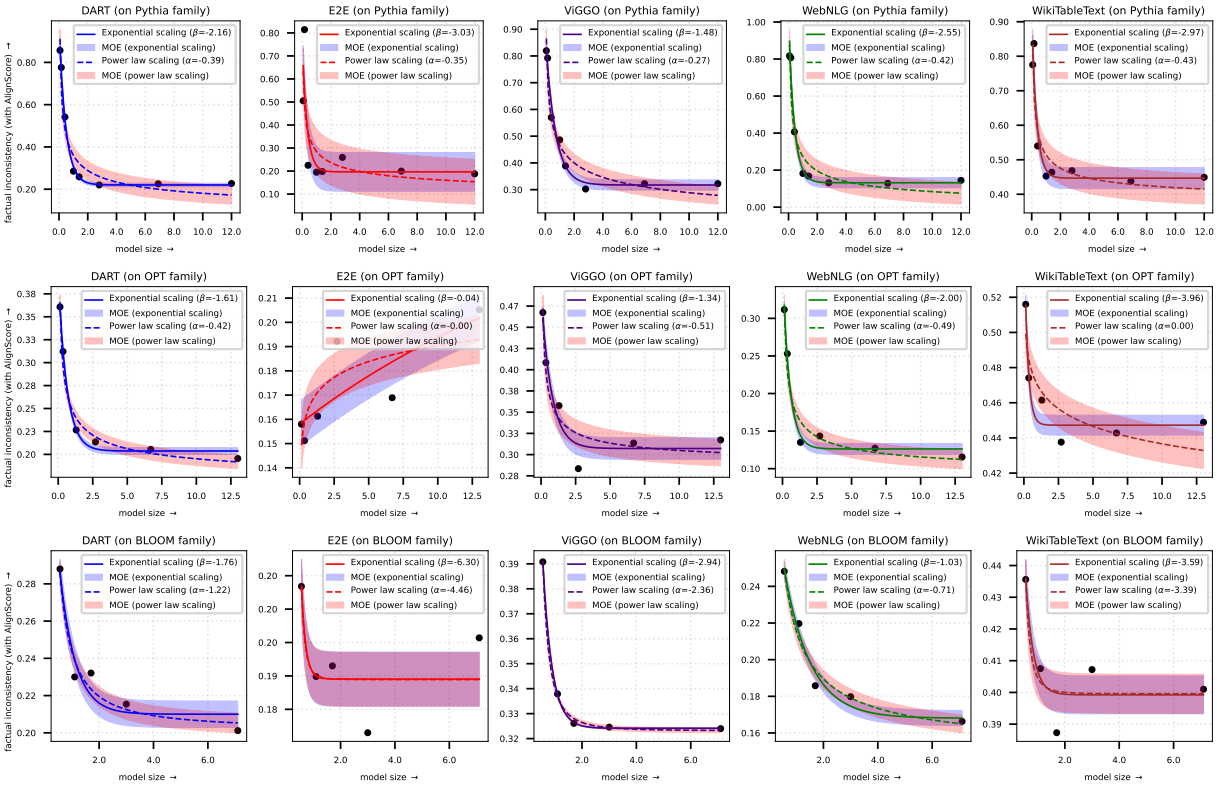


Figure 7: Visualization of exponential and power law scaling of factual inconsistency (ALIGNSCORE) across datasets and LLM families, with the margin of error (MOE) and 95% confidence intervals on residuals. Texts generated using the greedy search decoding algorithm. Aberrant behavior of the E2E dataset with the OPT family, as discussed in A.1.

| | | Results of stage I | | | | | Results of stage II and III | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LLM family | Scaling law | DART | E2E | ViGGO | WebNLG | WikiTableText | DART | E2E | ViGGO | WebNLG | WikiTableText |
| BLOOM | Exponential | 4.36e-04 | 5.32e-06 | 5.41e-05 | 2.69e-04 | 4.75e+02 | ✗ | ✗ | ✓(✿) | ✓(✿) | ✗ |
| | Power law | 2.75e-04 | 4.85e-05 | 2.19e-06 | 2.88e-05 | 3.30e-04 | ✓ | ✗ | ✓ | ✓ | ✗ |
| OPT | Exponential | 3.12e-05 | 1.15e+02 | 7.46e-03 | 2.37e-01 | 2.00e-04 | ✓(✿) | ✗ | ✓(✿) | ✓(✿) | ✓ |
| | Power law | 2.17e-03 | 1.11e-04 | 7.62e+19 | 6.96e-03 | 1.87e-04 | ✓ | ✗ | ✓ | ✓ | ✗ |
| Pythia | Exponential | 2.60e-04 | 1.96e-02 | 1.05e-03 | 6.05e-03 | 1.80e+01 | ✓(✿) | ✓ | ✓(✿) | ✓(✿) | ✓(✿) |
| | Power law | 2.22e-02 | 3.19e+00 | 2.02e-03 | 1.37e-01 | 1.15e-01 | ✓ | ✗ | ✓ | ✓ | ✓ |

Table 6: [Case: greedy decoding] Results of the validation framework (all three stages) for exponential and power law scaling of factual inconsistency (ALIGNSCORE). High held-out losses (Stage I) are highlighted in red. ✓/✗ indicates pass/fail (also marked in red) in the goodness-of-fit test (Stage II), while ✿ denotes the effective scaling law from Stage III.

| | | Results of stage I | | | | | Results of stage II and III | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LLM family | Scaling law | DART | E2E | ViGGO | WebNLG | WikiTableText | DART | E2E | ViGGO | WebNLG | WikiTableText |
| BLOOM | Exponential | 1.47e-04 | 4.07e-05 | 1.60e-04 | 1.11e-04 | 2.35e-04 | ✗ | ✗ | ✗ | ✓(✿) | ✗ |
| | Power law | 1.96e-04 | 4.17e-05 | 1.07e-04 | 3.86e-04 | 6.78e-05 | ✗ | ✗ | ✗ | ✓ | ✗ |
| OPT | Exponential | 2.59e-05 | 3.98e-03 | 2.99e-04 | 1.88e-03 | 1.54e-04 | ✓(✿) | ✓(✿) | ✓ | ✓ | ✓(✿) |
| | Power law | 1.90e-04 | 4.07e-04 | 1.48e-04 | 3.44e-03 | 5.28e-04 | ✓ | ✓ | ✓(✿) | ✓(✿) | ✓ |
| Pythia | Exponential | 2.89e-03 | 2.06e+01 | 1.47e-03 | 1.91e-02 | 8.07e-03 | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) |
| | Power law | 4.84e-02 | 3.79e-02 | 4.50e-02 | 2.83e-01 | 2.88e-01 | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 7: [Case: greedy decoding] Results of the validation framework (all three stages) for exponential and power law scaling of factual inconsistency (QAFACTEVAL). High held-out losses (Stage I) are highlighted in red. ✓/✗ indicates pass/fail (also marked in red) in the goodness-of-fit test (Stage II), while ✿ denotes the effective scaling law from Stage III.

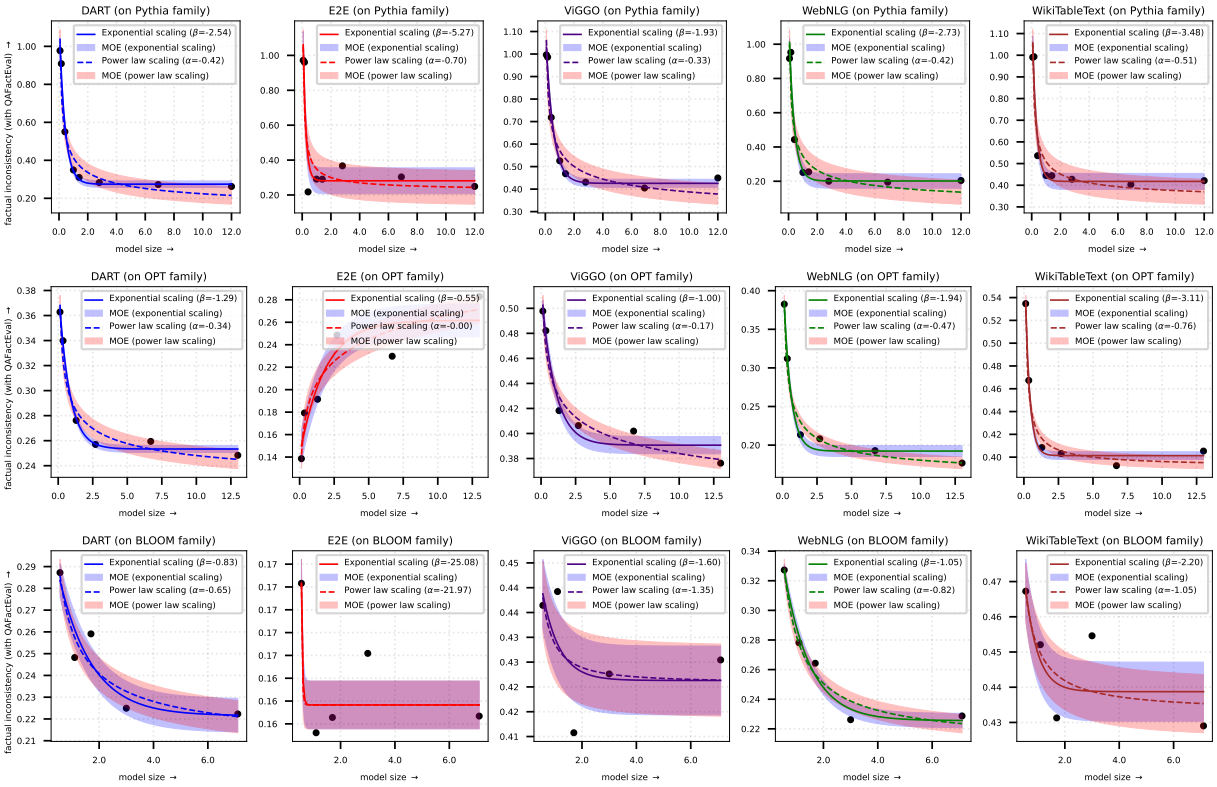


Figure 8: Visualization of exponential and power law scaling of factual inconsistency (QAFACTEVAL) across datasets and LLM families, with margin of error (MOE) and 95% confidence intervals on residuals. Texts generated using the greedy search decoding algorithm. Aberrant behavior of the E2E dataset with the OPT family, as discussed in A.1.

| | | Results of stage I | | | | | Results of stage II and III | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LLM family | Scaling law | DART | E2E | ViGGO | WebNLG | WikiTableText | DART | E2E | ViGGO | WebNLG | WikiTableText |
| BLOOM | Exponential | 4.46e-05 | 6.49e-06 | 2.80e-04 | 1.49e-04 | 2.24e-03 | ✓(✿) | ✗ | ✓(✿) | ✓ | ✗ |
| | Power law | 2.44e-04 | 8.29e-07 | 1.57e-04 | 2.83e-05 | 1.41e-04 | ✓ | ✗ | ✓ | ✗ | ✗ |
| OPT | Exponential | 1.14e-03 | 1.81e-02 | 6.09e-03 | 6.51e-03 | 5.25e-03 | ✓(✿) | ✓ | ✓(✿) | ✓ | ✓(✿) |
| | Power law | 1.12e-02 | 1.91e-02 | 3.07e-02 | 4.41e-03 | 1.12e-01 | ✓ | ✓(✿) | ✓ | ✓(✿) | ✗ |
| Pythia | Exponential | 2.88e-04 | 5.56e+00 | 9.95e-05 | 1.16e-03 | 1.31e-03 | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) |
| | Power law | 2.39e-01 | 2.61e+03 | 4.16e-04 | 3.87e-02 | 1.80e-03 | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 8: [Case: greedy decoding] Results of the validation framework (all three stages) for exponential and power law scaling of factual inconsistency (SUMMAC-CONV). High held-out losses (Stage I) are highlighted in red. ✓/✗ indicates pass/fail (also marked in red) in the goodness-of-fit test (Stage II), while ✿ denotes the effective scaling law from Stage III.

| | | Results of stage I | | | | | Results of stage II and III | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LLM family | Scaling law | DART | E2E | ViGGO | WebNLG | WikiTableText | DART | E2E | ViGGO | WebNLG | WikiTableText |
| BLOOM | Exponential | 6.76e-04 | 4.41e-05 | 7.84e-05 | 6.20e-05 | 5.08e-05 | ✗ | ✗ | ✗ | ✓(✿) | ✗ |
| | Power law | 2.40e-04 | 3.14e-05 | 8.77e-05 | 1.62e-03 | 1.65e-04 | ✗ | ✗ | ✗ | ✓ | ✗ |
| OPT | Exponential | 1.19e-05 | 1.30e-04 | 6.18e-05 | 7.44e-04 | 1.40e-04 | ✓(✿) | ✗ | ✓(✿) | ✓(✿) | ✓ |
| | Power law | 7.95e-04 | 5.41e-05 | 2.21e-04 | 2.73e-04 | 4.62e-04 | ✓ | ✗ | ✓ | ✓ | ✗ |
| Pythia | Exponential | 1.35e-03 | 4.14e-02 | 6.82e-04 | 9.60e-04 | 5.77e-03 | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) |
| | Power law | 3.32e-03 | 6.94e-01 | 4.18e-04 | 8.26e-04 | 1.08e-02 | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 9: [Case: greedy decoding] Results of the validation framework (all three stages) for exponential and power law scaling of factual inconsistency (UNIEVAL-FACT). High held-out losses (Stage I) are highlighted in red. ✓/✗ indicates pass/fail (also marked in red) in the goodness-of-fit test (Stage II), while ✿ denotes the effective scaling law from Stage III.

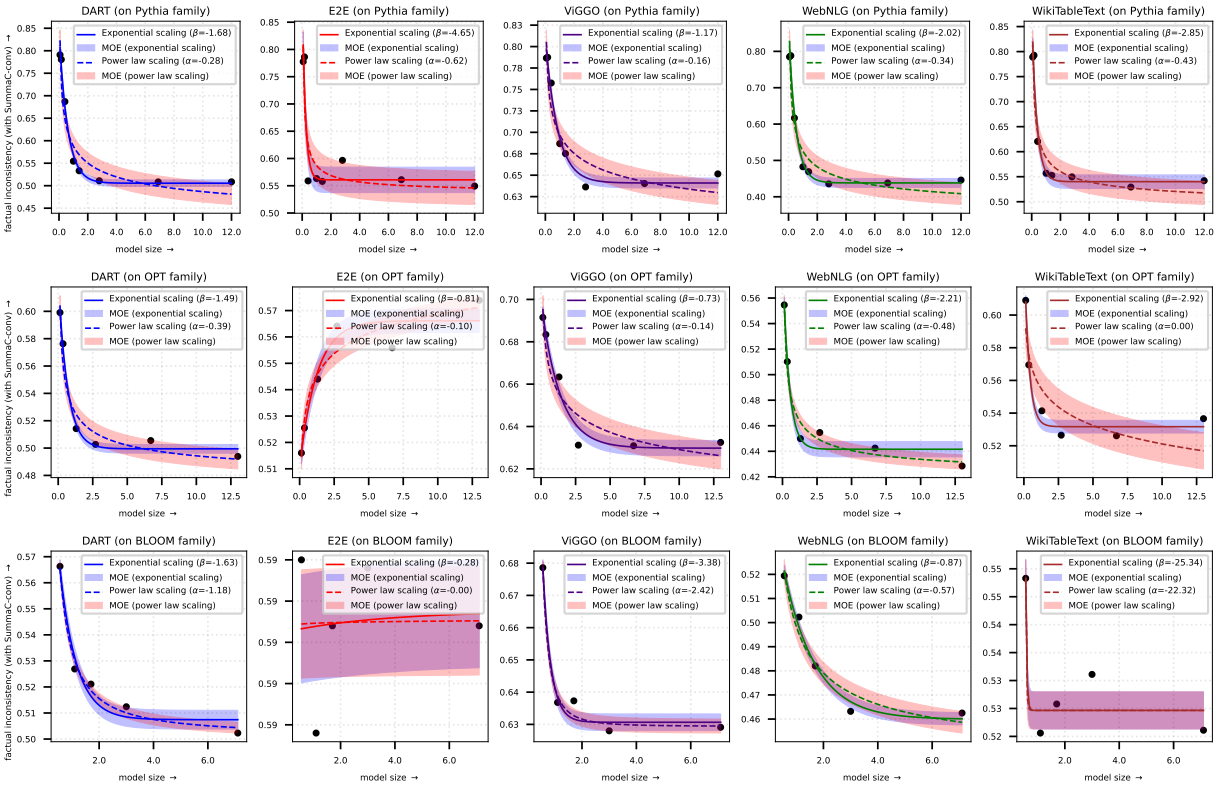


Figure 9: Visualization of exponential and power law scaling of factual inconsistency (SUMMAC-CONV) across datasets and LLM families, with margin of error (MOE) and 95% confidence intervals on residuals. Texts generated using the greedy search decoding algorithm. The unusual behavior of the E2E dataset with the OPT and BLOOM families is discussed in A.1.

| | | Results of stage I | | | | | Results of stage II and III | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LLM family | Scaling law | DART | E2E | ViGGO | WebNLG | WikiTableText | DART | E2E | ViGGO | WebNLG | WikiTableText |
| BLOOM | Exponential | 2.30e-04 | 4.46e-05 | 1.20e-03 | 2.35e+02 | 1.05e-04 | ✗ | ✗ | ✗ | ✗ | ✗ |
| | Power law | 1.69e-04 | 9.06e-05 | 9.50e-04 | 5.79e-05 | 1.94e-04 | ✓ | ✗ | ✗ | ✗ | ✗ |
| OPT | Exponential | 5.75e-06 | 6.66e-05 | 7.95e+02 | 1.57e-03 | 8.62e-03 | ✓(✿) | ✓ | ✓ | ✓ | ✓ |
| | Power law | 4.15e-04 | 1.20e-04 | 1.47e-03 | 3.25e-02 | 4.00e-04 | ✓ | ✗ | ✗ | ✗ | ✗ |
| Pythia | Exponential | 2.67e-03 | 2.74e-01 | 4.12e-03 | 1.48e-02 | 1.00e-02 | ✓(✿) | ✓ | ✓(✿) | ✓(✿) | ✓(✿) |
| | Power law | 1.79e+00 | 3.62e+00 | 3.52e-03 | 6.17e-02 | 1.90e-03 | ✓ | ✗ | ✓ | ✓ | ✓ |

Table 10: [Case: beam search decoding] Results of the validation framework (all three stages) for exponential and power law scaling of factual inconsistency (ALIGNSCORE). High held-out losses (Stage I) are highlighted in red. ✓/✗ indicates pass/fail (also marked in red) in the goodness-of-fit test (Stage II), while ✿ denotes the effective scaling law from Stage III.

| | | Results of stage I | | | | | Results of stage II and III | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LLM family | Scaling law | DART | E2E | ViGGO | WebNLG | WikiTableText | DART | E2E | ViGGO | WebNLG | WikiTableText |
| BLOOM | Exponential | 1.91e-04 | 5.37e-06 | 2.93e-04 | 9.19e-05 | 2.14e-04 | ✗ | ✗ | ✗ | ✗ | ✗ |
| | Power law | 1.37e-04 | 8.13e-06 | 4.54e-04 | 1.52e-03 | 9.32e-05 | ✗ | ✗ | ✗ | ✗ | ✗ |
| OPT | Exponential | 1.39e+02 | 2.02e+02 | 1.40e-03 | 2.24e-04 | 7.31e-02 | ✓(✿) | ✗ | ✗ | ✓(✿) | ✓(✿) |
| | Power law | 1.09e-03 | 6.17e-03 | 7.27e-03 | 7.60e-03 | 1.25e+02 | ✓ | ✗ | ✗ | ✓ | ✓ |
| Pythia | Exponential | 2.39e-03 | 2.10e+01 | 2.87e-04 | 2.11e-02 | 1.51e-02 | ✓(✿) | ✓ | ✓(✿) | ✓(✿) | ✓(✿) |
| | Power law | 3.79e-01 | 3.64e+04 | 9.37e-02 | 1.87e-01 | 1.46e-01 | ✓ | ✗ | ✓ | ✓ | ✓ |

Table 11: [Case: beam search decoding] Results of the validation framework (all three stages) for exponential and power law scaling of factual inconsistency (QAFACTEVAL). High held-out losses (Stage I) are highlighted in red. ✓/✗ indicates pass/fail (also marked in red) in the goodness-of-fit test (Stage II), while ✿ denotes the effective scaling law from Stage III.

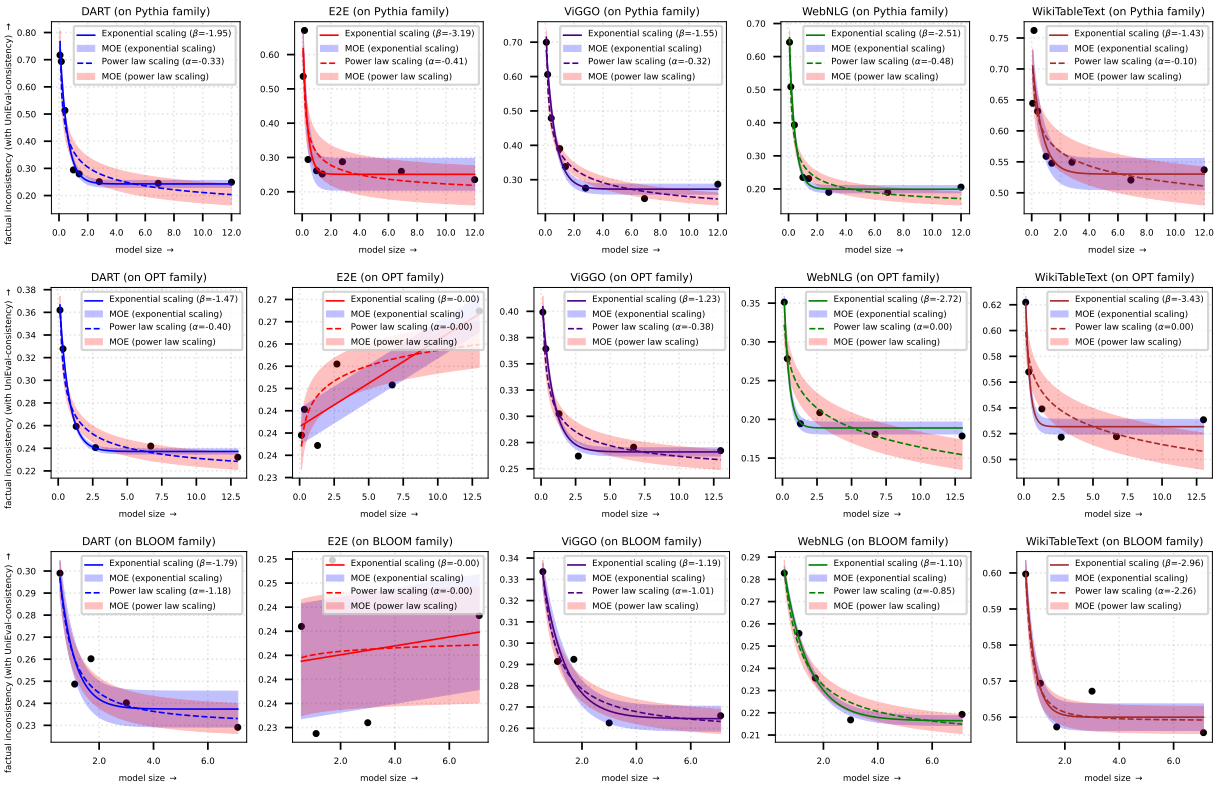


Figure 10: Visualization of exponential and power law scaling of factual inconsistency (UNIEVAL-FACT) across datasets and LLM families, with margin of error (MOE) and 95% confidence intervals on residuals. Texts generated using the greedy search decoding algorithm. The unusual behavior of the E2E dataset with the OPT family is discussed in A.1.

| | | Results of stage I | | | | | Results of stage II and III | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LLM family | Scaling law | DART | E2E | ViGGO | WebNLG | WikiTableText | DART | E2E | ViGGO | WebNLG | WikiTableText |
| BLOOM | Exponential | 1.55e-04 | 2.03e-05 | 1.53e+01 | 3.51e-04 | 3.38e-01 | ✗ | ✗ | ✓(✿) | ✗ | ✓ |
| | Power law | 2.65e-05 | 1.08e-04 | 6.76e-04 | 5.33e-05 | 1.49e-05 | ✓ | ✗ | ✓ | ✗ | ✗ |
| OPT | Exponential | 5.36e+01 | 2.00e-04 | 1.62e-04 | 3.13e-05 | 2.67e+01 | ✓(✿) | ✓ | ✓ | ✓(✿) | ✓ |
| | Power law | 5.59e-04 | 5.69e-05 | 3.19e-03 | 6.87e-05 | 1.64e-04 | ✓ | ✗ | ✗ | ✓ | ✗ |
| Pythia | Exponential | 1.11e-04 | 1.20e+01 | 1.34e-04 | 7.35e-04 | 1.36e-03 | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) |
| | Power law | 1.23e-03 | 1.10e+04 | 9.79e-04 | 1.91e-02 | 7.29e-04 | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 12: [Case: beam search decoding] Results of the validation framework (all three stages) for exponential and power law scaling of factual inconsistency (SUMMAC-CONV). High held-out losses (Stage I) are highlighted in red. ✓/✗ indicates pass/fail (also marked in red) in the goodness-of-fit test (Stage II), while ✿ denotes the effective scaling law from Stage III.

| | | Results of stage I | | | | | Results of stage II and III | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LLM family | Scaling law | DART | E2E | ViGGO | WebNLG | WikiTableText | DART | E2E | ViGGO | WebNLG | WikiTableText |
| BLOOM | Exponential | 6.54e-05 | 4.84e+02 | 7.41e-04 | 2.27e-04 | 8.25e-05 | ✗ | ✗ | ✗ | ✗ | ✓ |
| | Power law | 5.75e-05 | 3.88e-05 | 5.41e-05 | 3.64e-05 | 2.83e-06 | ✗ | ✗ | ✗ | ✗ | ✓(✿) |
| OPT | Exponential | 9.29e+01 | 6.84e-05 | 5.15e-05 | 1.05e-03 | 1.82e-04 | ✓(✿) | ✗ | ✓(✿) | ✓ | ✓(✿) |
| | Power law | 4.10e-03 | 1.15e-04 | 1.91e+01 | 7.27e-03 | 3.50e-05 | ✓ | ✗ | ✓ | ✗ | ✓ |
| Pythia | Exponential | 7.15e-05 | 3.72e-02 | 2.53e-03 | 9.19e-04 | 3.21e-03 | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) | ✓ |
| | Power law | 9.70e-01 | 4.10e+01 | 2.56e-04 | 5.47e-01 | 5.11e-03 | ✓ | ✓ | ✓ | ✓ | ✗ |

Table 13: [Case: beam search decoding] Results of the validation framework (all three stages) for exponential and power law scaling of factual inconsistency (UNIEVAL-FACT). High held-out losses (Stage I) are highlighted in red. ✓/✗ indicates pass/fail (also marked in red) in the goodness-of-fit test (Stage II), while ✿ denotes the effective scaling law from Stage III.

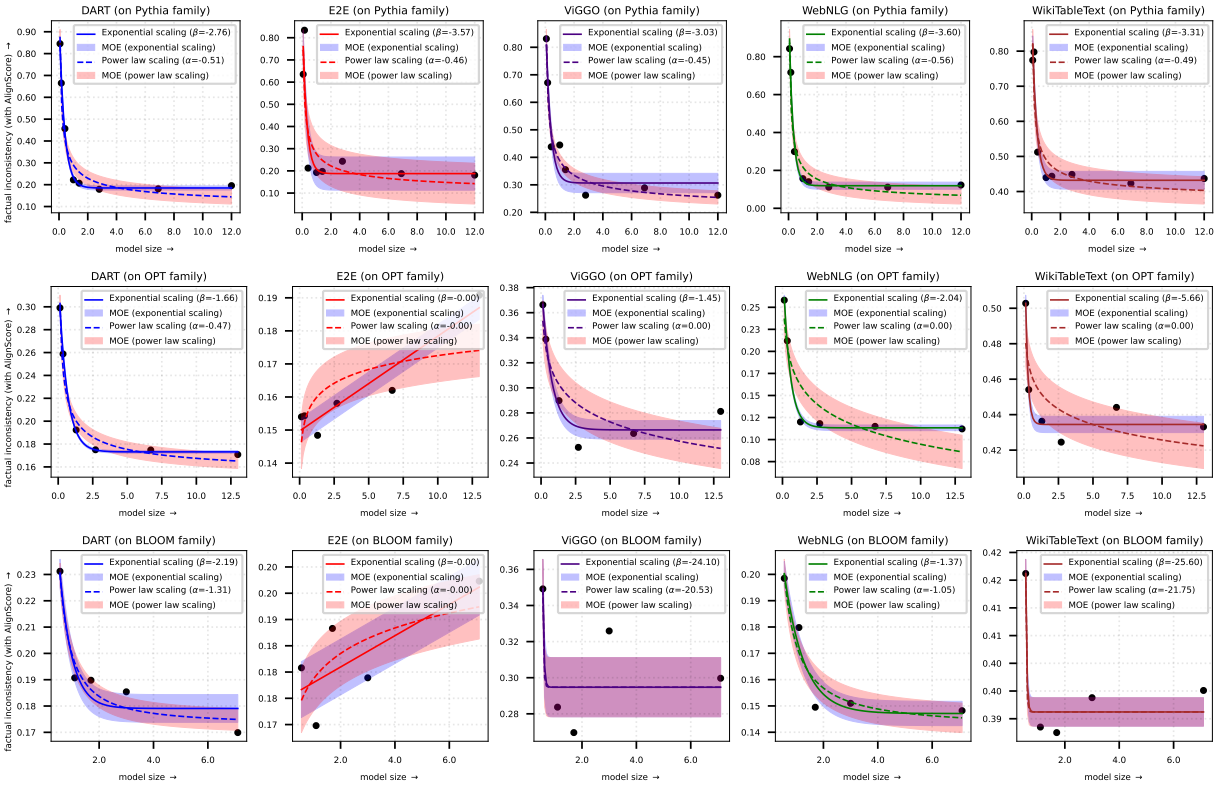


Figure 11: Visualization of exponential and power law scaling of factual inconsistency (ALIGNSCORE) across datasets and LLM families, with the margin of error (MOE) and 95% confidence intervals on residuals. Texts generated using the beam search decoding algorithm. The unusual behavior of the E2E dataset with the OPT and BLOOM families is discussed in A.1.

| | | Results of stage I | | | | | Results of stage II and III | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LLM family | Scaling law | DART | E2E | ViGGO | WebNLG | WikiTableText | DART | E2E | ViGGO | WebNLG | WikiTableText |
| BLOOM | Exponential | 1.70e-03 | 1.74e-03 | 4.01e-04 | 2.90e-04 | 2.39e-03 | ✗ | ✗ | ✓ | ✓(✿) | ✓ |
| | Power law | 4.85e-04 | 1.25e-04 | 3.89e-04 | 4.50e-03 | 3.91e-04 | ✓ | ✗ | ✓(✿) | ✓ | ✗ |
| OPT | Exponential | 8.14e-04 | 3.97e-03 | 1.15e-03 | 6.12e-04 | 3.80e+02 | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) |
| | Power law | 4.47e-03 | 2.84e-03 | 9.17e-04 | 3.69e-03 | 4.28e-03 | ✓ | ✓ | ✓ | ✓ | ✓ |
| Pythia | Exponential | 5.76e-04 | 6.57e-04 | 2.31e-03 | 1.28e-03 | 8.77e-04 | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) |
| | Power law | 9.31e-03 | 1.31e-02 | 7.78e-03 | 1.36e-02 | 1.52e-03 | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 14: [Case: top-k decoding] Results of the validation framework (all three stages) for exponential and power law scaling of factual inconsistency (ALIGNSCORE). High held-out losses (Stage I) are highlighted in red. ✓/✗ indicates pass/fail (also marked in red) in the goodness-of-fit test (Stage II), while ✿ denotes the effective scaling law from Stage III.

| | | Results of stage I | | | | | Results of stage II and III | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LLM family | Scaling law | DART | E2E | ViGGO | WebNLG | WikiTableText | DART | E2E | ViGGO | WebNLG | WikiTableText |
| BLOOM | Exponential | 8.47e-04 | 2.77e-04 | 2.86e-05 | 1.80e-04 | 1.34e-03 | ✗ | ✗ | ✓(✿) | ✓(✿) | ✗ |
| | Power law | 3.52e-04 | 3.60e-04 | 3.73e-05 | 5.46e-04 | 7.83e-04 | ✓ | ✗ | ✓ | ✓ | ✗ |
| OPT | Exponential | 1.54e-03 | 5.44e-04 | 3.86e-04 | 1.17e-03 | 6.48e-05 | ✓(✿) | ✓ | ✓(✿) | ✓ | ✓(✿) |
| | Power law | 1.37e-03 | 8.75e-05 | 1.72e-03 | 6.90e-03 | 2.49e+01 | ✓ | ✗ | ✓ | ✓(✿) | ✓ |
| Pythia | Exponential | 1.50e-03 | 1.48e-02 | 1.55e-03 | 2.76e-03 | 1.48e-03 | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) |
| | Power law | 2.30e-02 | 1.84e-01 | 5.73e-03 | 4.43e-02 | 3.10e-03 | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 15: [Case: top-k decoding] Results of the validation framework (all three stages) for exponential and power law scaling of factual inconsistency (QAFACTEVAL). High held-out losses (Stage I) are highlighted in red. ✓/✗ indicates pass/fail (also marked in red) in the goodness-of-fit test (Stage II), while ✿ denotes the effective scaling law from Stage III.

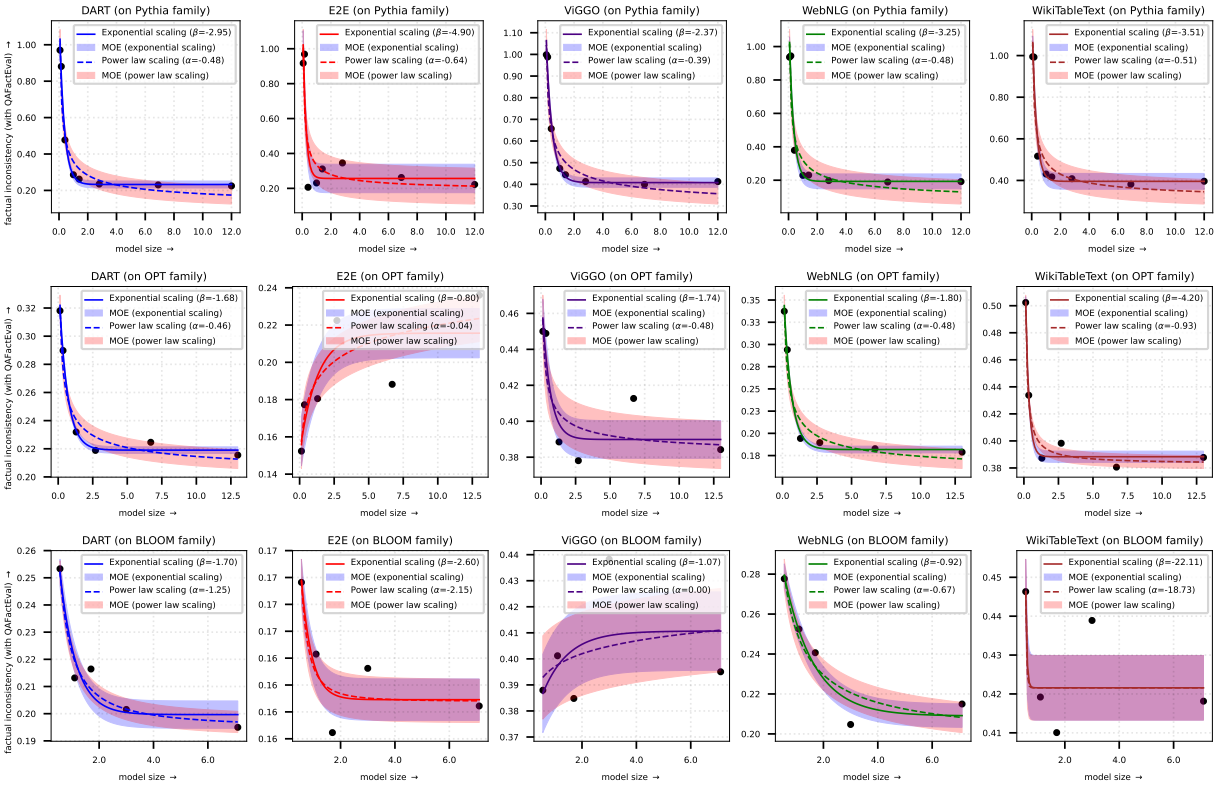


Figure 12: Visualization of exponential and power law scaling of factual inconsistency (QAFACTEVAL) across datasets and LLM families, with margin of error (MOE) and 95% confidence intervals on residuals. Texts generated using the beam search decoding algorithm. The unusual behavior of the E2E and ViGGO dataset with the OPT family is discussed in A.1.

| | | Results of stage I | | | | | Results of stage II and III | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LLM family | Scaling law | DART | E2E | ViGGO | WebNLG | WikiTableText | DART | E2E | ViGGO | WebNLG | WikiTableText |
| BLOOM | Exponential | 2.13e-04 | 4.52e-05 | 7.56e-05 | 1.68e-03 | 8.72e-04 | ✗ | ✗ | ✗ | ✓ | ✗ |
| | Power law | 4.90e-05 | 2.76e-05 | 8.40e-05 | 1.48e-04 | 1.60e-04 | ✓ | ✗ | ✗ | ✗ | ✗ |
| OPT | Exponential | 3.00e-05 | 4.87e-05 | 1.82e-04 | 1.34e-04 | 4.44e+02 | ✓(✿) | ✓ | ✓(✿) | ✓(✿) | ✓(✿) |
| | Power law | 1.11e-03 | 7.14e-05 | 6.31e-05 | 1.72e-04 | 8.68e-04 | ✓ | ✗ | ✓ | ✓ | ✓ |
| Pythia | Exponential | 9.16e-04 | 7.38e-05 | 7.46e-05 | 5.39e-05 | 1.48e-04 | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) |
| | Power law | 1.08e-03 | 2.50e-04 | 7.98e-05 | 2.99e-02 | 3.12e-04 | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 16: [Case: top-k decoding] Results of the validation framework (all three stages) for exponential and power law scaling of factual inconsistency (SUMMAC-CONV). High held-out losses (Stage I) are highlighted in red. ✓/✗ indicates pass/fail (also marked in red) in the goodness-of-fit test (Stage II), while ✿ denotes the effective scaling law from Stage III.

| | | Results of stage I | | | | | Results of stage II and III | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LLM family | Scaling law | DART | E2E | ViGGO | WebNLG | WikiTableText | DART | E2E | ViGGO | WebNLG | WikiTableText |
| BLOOM | Exponential | 4.84e-04 | 1.94e-03 | 1.99e-04 | 5.07e-05 | 1.03e-05 | ✗ | ✗ | ✓ | ✓(✿) | ✓(✿) |
| | Power law | 7.37e-04 | 9.46e-05 | 1.55e-04 | 2.52e-04 | 1.09e-03 | ✓ | ✗ | ✓(✿) | ✓ | ✓ |
| OPT | Exponential | 2.05e-04 | 5.95e-05 | 5.23e-04 | 7.91e-04 | 8.86e-04 | ✓(✿) | ✓(✿) | ✓ | ✓(✿) | ✓(✿) |
| | Power law | 7.80e-03 | 3.85e-04 | 2.05e-03 | 2.98e-03 | 5.55e-04 | ✓ | ✓ | ✓(✿) | ✓ | ✓ |
| Pythia | Exponential | 5.08e-04 | 7.40e-05 | 6.71e-03 | 1.47e-04 | 5.26e-04 | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) | ✓(✿) |
| | Power law | 7.34e-03 | 6.73e-04 | 1.54e-03 | 3.71e-03 | 5.66e-03 | ✓ | ✓ | ✓ | ✓ | ✓ |

Table 17: [Case: top-k decoding] Results of the validation framework (all three stages) for exponential and power law scaling of factual inconsistency (UNIEVAL-FACT). High held-out losses (Stage I) are highlighted in red. ✓/✗ indicates pass/fail (also marked in red) in the goodness-of-fit test (Stage II), while ✿ denotes the effective scaling law from Stage III.

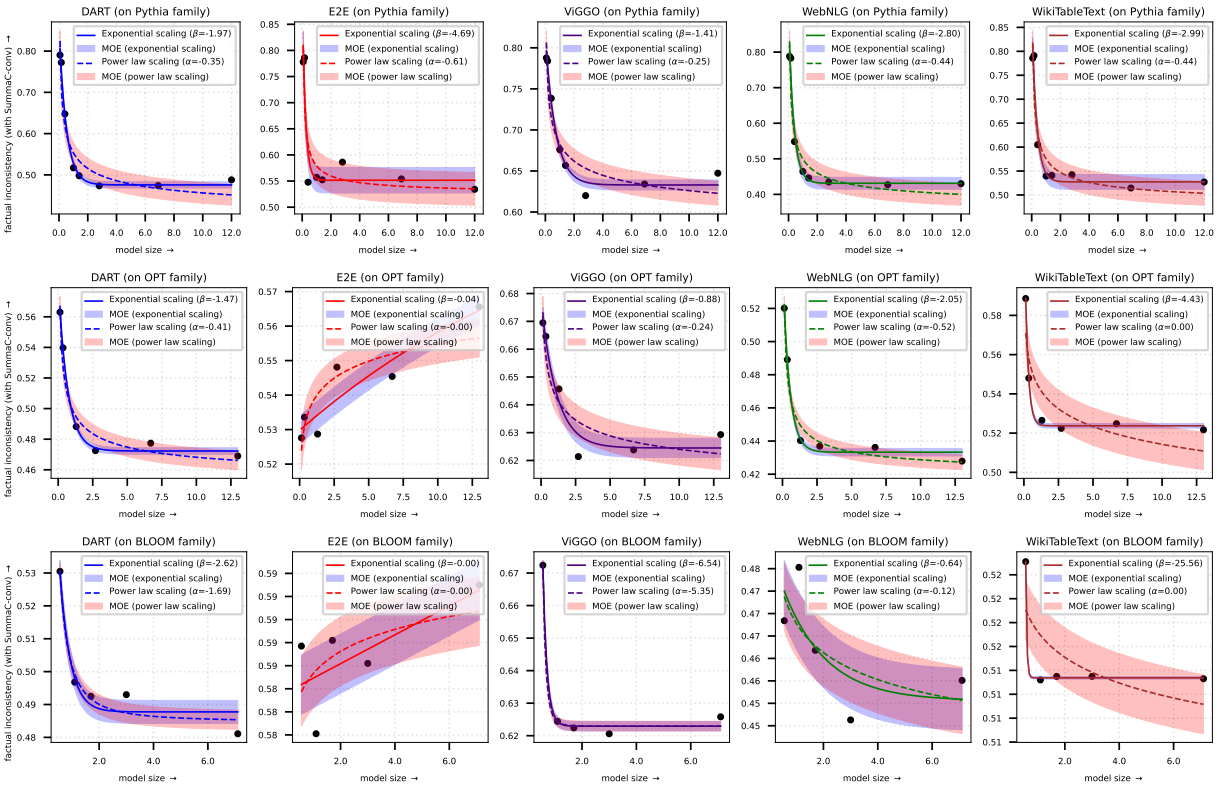


Figure 13: Visualization of exponential and power law scaling of factual inconsistency (SUMMAC-CONV) across datasets and LLM families, with margin of error (MOE) and 95% confidence intervals on residuals. Texts generated using the beam search decoding algorithm. The unusual behavior of the E2E dataset with the OPT and BLLOM families is discussed in A.1.

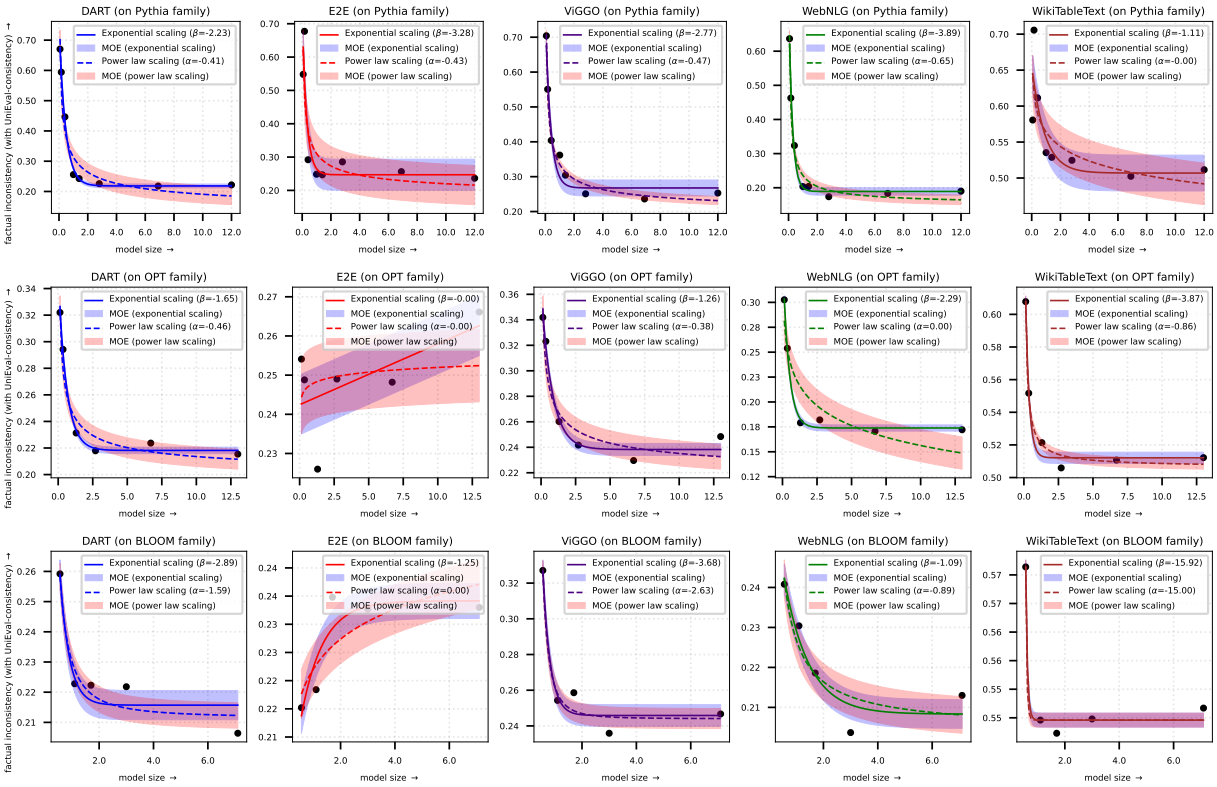


Figure 14: Visualization of exponential and power law scaling of factual inconsistency (UNIEVAL-FACT) across datasets and LLM families, with margin of error (MOE) and 95% confidence intervals on residuals. Texts generated using the beam search decoding algorithm. The unusual behavior of the E2E dataset with the OPT and BLOOM families is discussed in A.1.

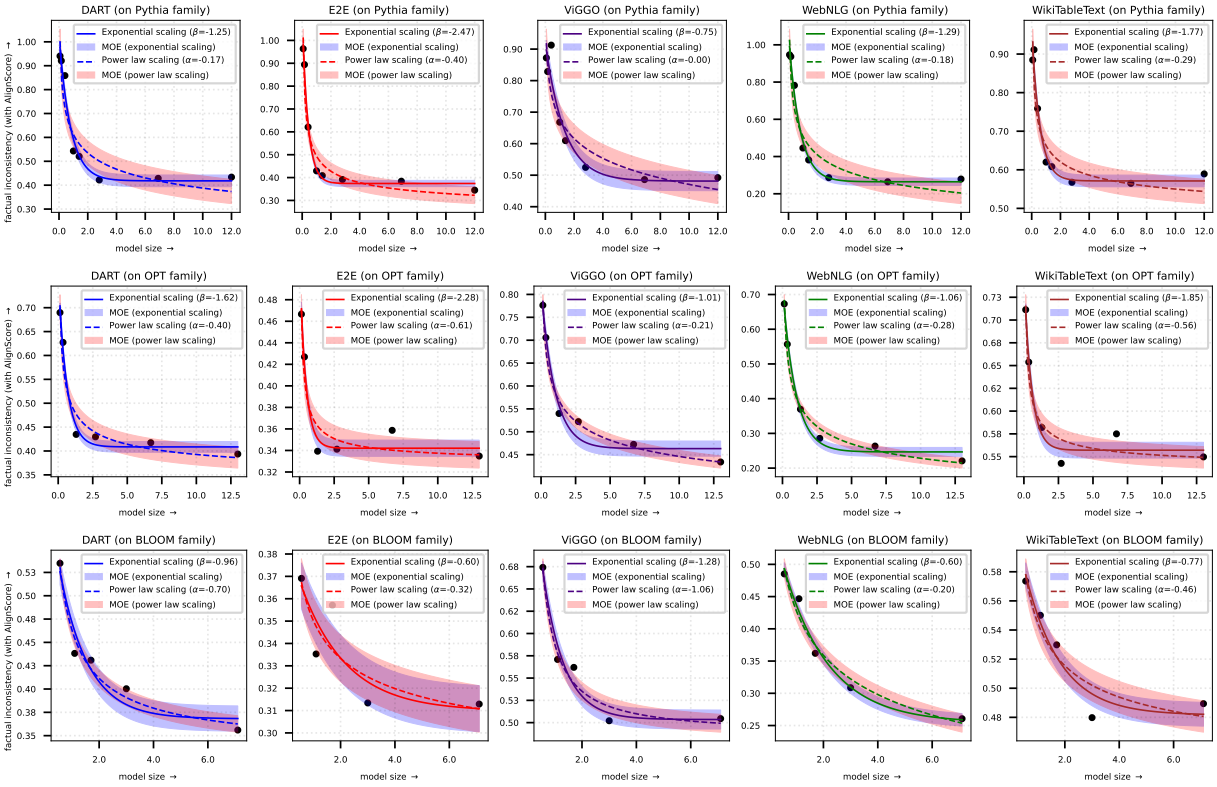


Figure 15: Visualization of exponential and power law scaling of factual inconsistency (ALIGNSCORE) across datasets and LLM families, with the margin of error (MOE) and 95% confidence intervals on residuals. Texts generated using the topk search decoding algorithm.

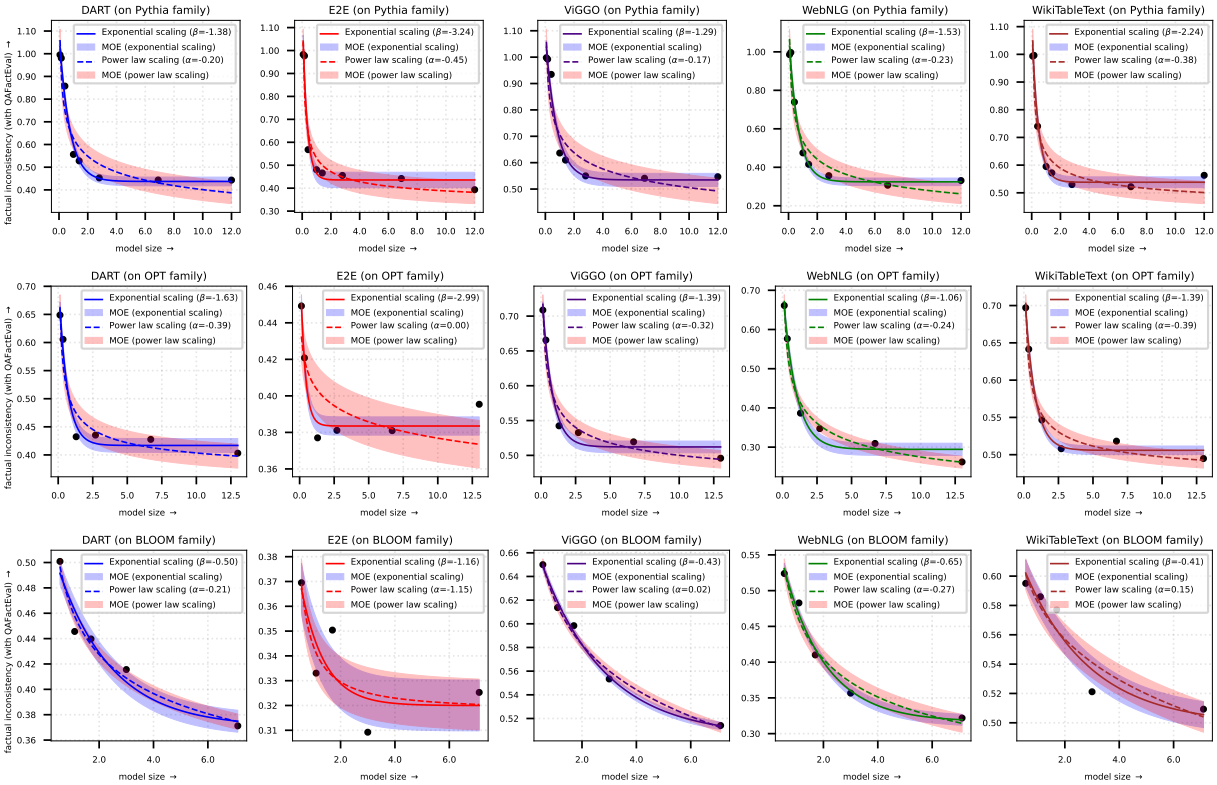


Figure 16: Visualization of exponential and power law scaling of factual inconsistency (QAFACTEVAL) across datasets and LLM families, with margin of error (MOE) and 95% confidence intervals on residuals. Texts generated using the topk search decoding algorithm.

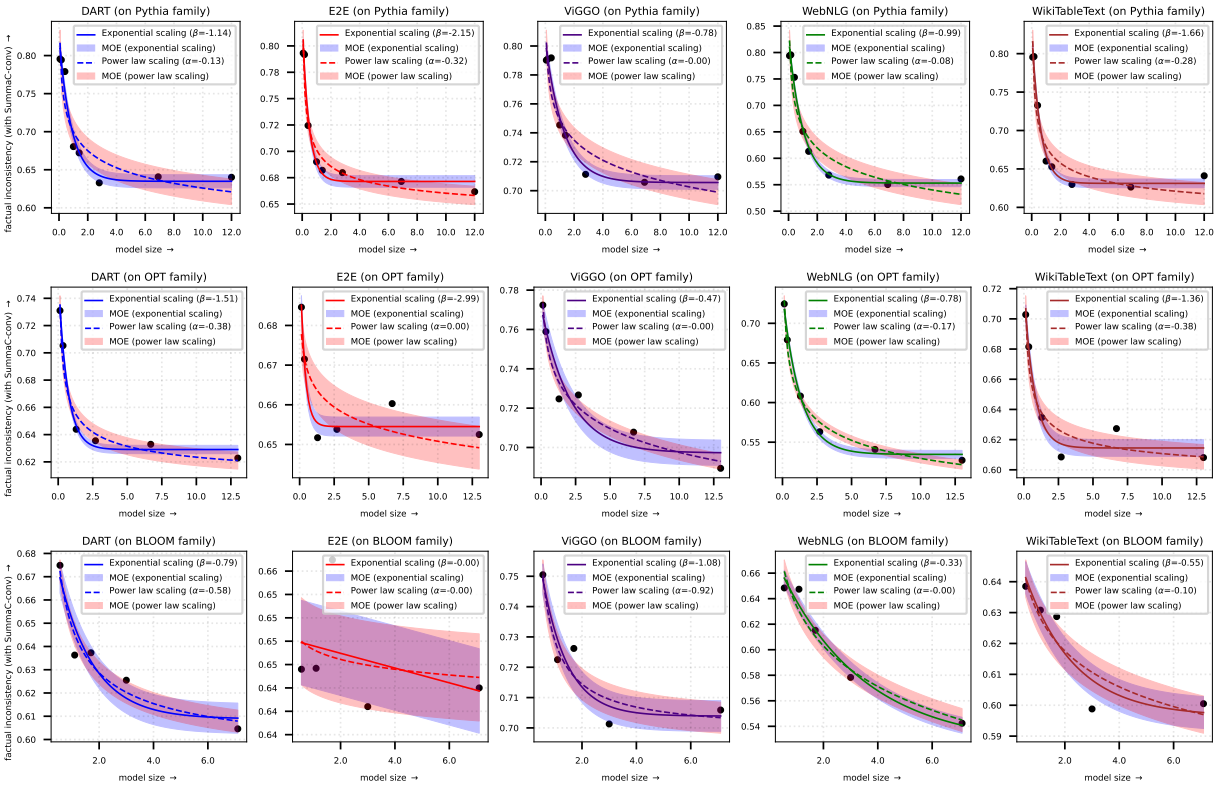


Figure 17: Visualization of exponential and power law scaling of factual inconsistency (SUMMAC-CONV) across datasets and LLM families, with margin of error (MOE) and 95% confidence intervals on residuals. Texts generated using the topk search decoding algorithm.

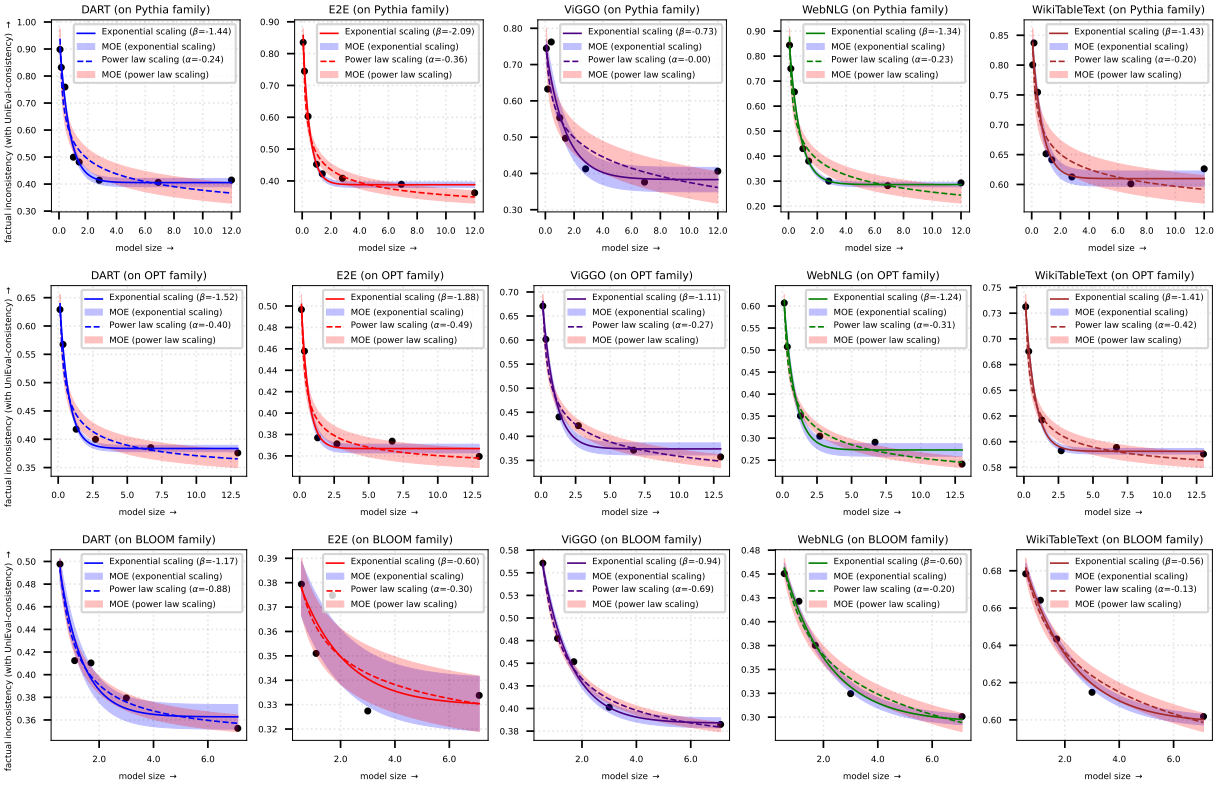


Figure 18: Visualization of exponential and power law scaling of factual inconsistency (UNIEVAL-FACT) across datasets and LLM families, with margin of error (MOE) and 95% confidence intervals on residuals. Texts generated using the topk search decoding algorithm.