

406 A Additional Experiment Results

407 A.1 Result on Digit Bias Benchmark of across Models

408 We plot the digit distribution in generated outputs across all models on the Digit Bias Benchmark as
 409 shown in Figure 9. All models show significant overgeneration of small digits, with digit 1 being
 410 especially dominant. This consistent trend highlights the generality of digit-level generation bias
 411 across open-source LLMs.

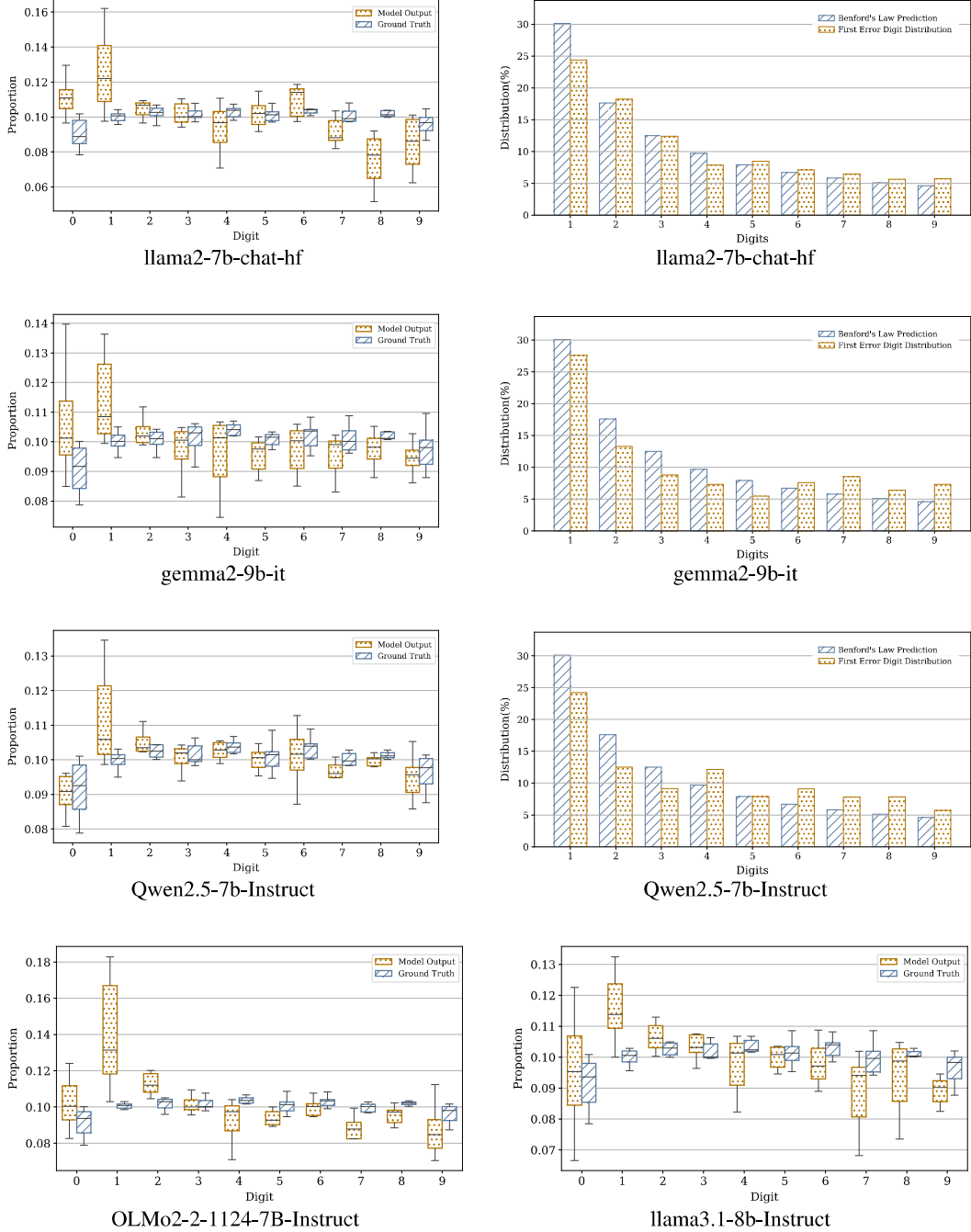


Figure 9: Digit generation bias across models on the Digit Bias Benchmark. Since OLMo and LLaMA3.18B employ multi-digit tokenization schemes, first-error-digit distribution analyses are not applicable and thus omitted for these models.

412 A.2 Digit Selectivity

413 Figure 10 visualizes the selectivity of FFN neurons across all digit tokens. The distributions are
 414 clearly skewed, indicating that more neurons are specialized toward frequent digits like '1', suggesting
 an uneven allocation of model capacity that may underlie observed generation biases.



Figure 10: FFN neuron-level selectivity distributions for digit '1' (blue) versus all other digits (orange) across four open-source LLMs: (a) LLaMA2-7B, (b) Qwen2.5-7B, (c) Gemma2-9B, and (d) Mistral-7B. For each digit, the top 1000 neurons with the highest selectivity scores are selected independently.

415

416 A.3 Additional Pruning Examples and Results

417 Figure 11 presents additional examples of samples corrected after pruning. Many of the original
 418 errors occur in surprisingly simple scenarios, such as repeated or trivial arithmetic, highlighting how
 419 digit bias can subtly but significantly affect digit generation and lead to failure even in straightforward
 420 cases.

Comparing Model Behavior With and Without Pruning

Prompt 1: Solve $-2005 = -1453p + 1977p + 3235$ for p .
Generation (no pruning): ... Simplify the left side: $\backslash[-5240 = 524p\backslash]$ 3. ****Solve for p **** by by dividing both sides by 524: $\backslash[p = \frac{-5241}{524}\backslash]$... **✗**
Generation (pruning): ... Simplify the left side: $\backslash[-5240 = 524p\backslash]$ 3. ****Solve for p **** by by dividing both sides by 524: $\backslash[p = \frac{-5240}{524}\backslash]$... **✓**

Prompt 2: Solve $38w = 75 - 75 - 1026$ for w .
Generation (no pruning): Let the equation given is: $\backslash[38w = 75 - 75 - 1126\backslash]$... **✗**
Generation (pruning): Let the equation given is: $\backslash[38w = 75 - 75 - 1026\backslash]$... **✓**

Prompt 3: Solve $78 + 491 + 817 = 28x + 266$ for x .
Generation (no pruning): ... 4. ****Perform the division:**** $\backslash[x = 40\backslash]$ So, the solution for x is is 41. **✗**
Generation (pruning): ... 4. ****Perform the division:**** $\backslash[x = 40\backslash]$ So, the solution for x is 40. **✓**

Prompt 4: Solve $70 = -1160m - 6852 - 3518$ for m .
Generation (no pruning): To solve the equation $71 = -1161m - 6852 - 3518$ for m ... **✗**
Generation (pruning): To solve the equation $71 = -1160m - 6852 - 3518$ for m ... **✓**

Prompt 5: Solve $38a - 1664 = -36a - 69a + 39a$ for a .
Generation (no pruning): ... Simplify the left side: $\backslash[104a - 1664 = 0\backslash]$ 3. ****Isolate the term with a :** Add 1664 to both sides to move the constant to the right side: $\backslash[104a = 1164\backslash]$... **✗**
Generation (pruning): ... Simplify the left side: $\backslash[104a - 1664 = 0\backslash]$ 3. ****Isolate the term with a :** Add 1664 to both sides to move the constant to the right side: $\backslash[104a = 1664\backslash]$...**✓**

Prompt 6: Solve $-47613w = -47331w + 14100$ for w .
Generation (no pruning): To solve the equation $-47613w = -47331w + 14110$ for w ... **✗**
Generation (pruning): To solve the equation $-47613w = -47331w + 14100$ for w ...**✓**

Figure 11: Examples showing model responses before and after pruning the top 0.01% of neurons biased toward digit 1. In each cases, pruning corrects an originally erroneous sample by rectifying an intermediate step, demonstrating a causal relationship between digit bias and numerical hallucination.

421 Table 3 reports the change in the generation frequency of digit '1' and the proportion of corrected
 422 samples after pruning the top 0.05% of neurons with the highest selectivity toward digit '1'. Table 4
 423 presents the change in task accuracy before and after neuron pruning.

Table 3: Effect of pruning top 0.005% most digit-1-selective neurons. **Coreccted - Prop.** denotes the proportion of all test samples that were originally incorrect but become correct after pruning. **Original - Prop.** denotes the original frequency of digit 1 in model outputs. **Pruned - Prop.** denotes the digit 1 frequency after pruning.

Model		Multistep Reasoning Tasks			
		Evaluate	Linear_1d	Nearest Integer root	Sequence Next term
Llama2-7B	Coreccted - Prop.	0.58 %	0.78 %	0.23 %	0.27 %
	Original - Prop.	16.26 %	16.21 %	11.91 %	11.70 %
	Pruned - Prop.	14.43 %	15.28 %	9.81 %	10.89 %
Mistral-7B	Coreccted - Prop.	1.11 %	1.55 %	0.16 %	0.80 %
	Original - Prop.	15.63 %	14.49 %	21.64 %	11.90 %
	Pruned - Prop.	12.43 %	12.20 %	12.71 %	11.04 %
Qwen2.5-7B	Coreccted - Prop.	3.90 %	4.08 %	6.90 %	4.91 %
	Original - Prop.	16.45 %	15.85 %	16.25 %	14.06 %
	Pruned - Prop.	15.64 %	15.10 %	14.39 %	12.00 %
Gemma2-9B	Coreccted - Prop.	0.87 %	1.55 %	1.86 %	1.78 %
	Original - Prop.	13.66 %	11.04 %	17.49 %	11.36 %
	Pruned - Prop.	12.30 %	10.88 %	15.02 %	11.11 %

Table 4: Effect of pruning most digit-1-selective neurons. **Original - Acc.** denotes the original accuracy of digit 1 in model outputs. **Pruned - Acc.(0.005%/0.01%)** denotes the accuracy after pruning 0.005%/0.01% neurons.

Model		Multistep Reasoning Tasks			
		Evaluate	Linear_1d	Nearest Integer root	Sequence Next term
Llama2-7B	Original - Acc.	12.39 %	3.30 %	0.08 %	3.75 %
	Pruned - Acc.(0.005%)	12.62 %(+0.23)	3.60 %(+0.30)	0.23 %(+0.15)	3.84 %(+0.09)
	Pruned - Acc.(0.01%)	12.62 %(+0.23)	3.89 %(+0.59)	0.19 %(+0.11)	3.66 %(-0.09)
Mistral-7B	Original - Acc.	27.17 %	8.36 %	0.93 %	13.02 %
	Pruned - Acc.(0.005%)	27.52 %(+0.35)	8.26 %(-0.10)	0.76 %(-0.17)	12.93 %(-0.09)
	Pruned - Acc.(0.01%)	27.28 %(+0.11)	8.36 % (0.00)	0.81 %(-0.12)	11.60 %(-1.42)
Qwen2.5-7B	Original - Acc.	38.28 %	33.82 %	19.81 %	14.72 %
	Pruned - Acc.(0.005%)	38.74 %(+0.46)	34.01 %(+0.19)	19.23 %(-0.58)	15.61 %(+0.89)
	Pruned - Acc.(0.01%)	39.62 %(+1.34)	33.43 %(-0.39)	18.26 %(-1.55)	16.41 %(+1.69)
Gemma2-9B	Original - Acc.	59.69 %	88.92 %	12.87 %	39.88 %
	Pruned - Acc.(0.005%)	59.86 %(+0.17)	89.21 %(+0.29)	12.41 %(-0.46)	40.41 %(+0.53)
	Pruned - Acc.(0.01%)	60.15 %(+0.46)	89.12 %(+0.20)	12.29 %(-0.58)	39.43 %(-0.45)

B Additional Implementation Details

B.1 Experiment Setup

To ensure the accuracy and reproducibility of all results, we employed greedy decoding for generation. Additionally, to achieve fully accurate statistical outcomes, we utilized the DeepSeek API to extract the answers model generated for each sample, thereby preventing potential omissions in script-based statistics due to variations in the position of answers within the responses.

B.2 Tokenization Method

LLMs divide numbers into segmented tokens rather than representing the entire number as a single token. Different LLMs employ various tokenization methods, including *one-digit tokenizers* and *multi-digit tokenizers*. Benford’s Law states that in many real-life sets of nu-

merical data, the leading digit is likely to be small. In other words, regardless of the type of tokenizer, the distribution of number tokens in pretraining data is likely to be long-tailed. Therefore, in this paper, we use six models with two different tokenizers to investigate this phenomenon of digit bias as shown in Figure 12. LLaMA2-7B, Mistral-7B, Qwen2.5-7B, and Gemma2-9B employ single-digit tokenizers, whereas LLaMA3.1-8B and OLMo2-7B utilize multi-digit tokenization schemes.

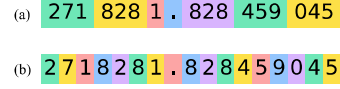


Figure 12: (a) multi-digit tokenizer. (b) single-digit tokenizer.

B.3 Prompt Templates

We provide the exact prompt templates used for Identification task and Digit Bias Benchmark in table 5 and table 6.

Table 5: Prompt Templates Used in Identification Task

Identification Prompt Template
1. What is the result when the last term of the sequence is multiplied by two? [...]
2. What is the outcome when the final term of the sequence is doubled? [...]
3. What is the product of the sequence’s last term and two? [...]
4. What is the result of multiplying the sequence’s last term by two? [...]

C Use of existing assets

C.1 Models

Table 7: The list of models used in this work.

Model	Accessed via	License
Qwen2.5-7B-Instruct	Link	Apache license 2.0
gemma-2-9b-it	Link	Gemma Terms of Use
Mistral-7B-Instruct-v0.3	Link	Apache license 2.0
Llama-3.1-8B-Instruct	Link	Llama 3.1 Community License Agreement
Llama-2-7b-chat-hf	Link	Llama 2 Community License Agreement
OLMo-2-1124-7B-Instruct	Link	Apache license 2.0

C.2 Dataset

Table 8: The list of datasets used in this work.

Dataset	Accessed via	License
olmo-mix-1124	Link	Open Data Commons License Attribution
mathematics_dataset	Link	Apache license 2.0

D Compute statement

All experiments presented in this paper were run on a cluster of four NVIDIA GeForce RTX 3090 GPUs with 24GB of memory and using a single 24GB memory GPU. Each model requires an average of 50 hours to complete a full run across the entire benchmark.

Table 6: Prompt Templates Used in Digit Bias Benchmark

Addition	Division
<p> $\{p\} + \{q\}$ $\{p\} + \{q\}$ Work out $\{p\} + \{q\}$. Add $\{p\}$ and $\{q\}$. Put together $\{p\}$ and $\{q\}$. Sum $\{p\}$ and $\{q\}$. Total of $\{p\}$ and $\{q\}$. Add together $\{p\}$ and $\{q\}$. What is $\{p\}$ plus $\{q\}$? Calculate $\{p\} + \{q\}$. What is $\{p\} + \{q\}$? </p>	<p> Calculate the division of $\{q\}$ by $\{p\}$. Divide $\{q\}$ by $\{p\}$. What is the quotient of $\{q\}$ divided by $\{p\}$? What is $\{q\}$ divided by $\{p\}$? Find $\{q\}$ divided by $\{p\}$. Compute $\{q\} \div \{p\}$. Solve $\{q\}$ divided by $\{p\}$. </p>
Subtraction	Multiplication
<p> $\{p\} - \{q\}$ Work out $\{p\} - \{q\}$. What is $\{p\}$ minus $\{q\}$? What is $\{p\}$ take away $\{q\}$? What is $\{q\}$ less than $\{p\}$? Subtract $\{q\}$ from $\{p\}$. Calculate $\{p\} - \{q\}$. What is $\{p\} - \{q\}$? </p>	<p> $\{p\} \times \{q\}$ Calculate $\{p\} \times \{q\}$. Work out $\{p\} \times \{q\}$. Multiply $\{p\}$ and $\{q\}$. Product of $\{p\}$ and $\{q\}$. What is the product of $\{p\}$ and $\{q\}$? $\{p\}$ times $\{q\}$ What is $\{p\}$ times $\{q\}$? </p>
Evaluate	
<p> Let $\{c(x) = f(x)\}$. What is $\{c(a)\}$? Let $\{c(x) = f(x)\}$. Determine $\{c(a)\}$. Let $\{c(x) = f(x)\}$. Give $\{c(a)\}$. Let $\{c(x) = f(x)\}$. Calculate $\{c(a)\}$. </p>	
Nearest Integer Root	
<p> What is the $\{\text{num-th}\}$ root of $\{p\}$ to the nearest integer? What is $\{p\}$ to the power of $\{1/\text{num-th}\}$, to the nearest integer? </p>	
Linear_1d	
Solve {equation} for {r}.	
Sequence Next Term	
<p> What comes next: {sequence}? What is next in {sequence}? What is the next term in {sequence}? </p>	