TinyStories

How Small Can Language Models Be and Still Speak Coherent English?

Essential elements of natural language

- **Grammar + Vocabulary + Facts + Reasoning**
- **→ LLMs generating coherent English text must have learned facts and logical/contextual reasoning as well**
- **“Jack was hungry, so he went looking for ⟨…⟩”**
  - State of hunger motivates search for food
- **“Lily wanted to get either a cat or a dog. Her mother didn’t let her get a dog so instead she ⟨…⟩”**
  - (cat or dog) and (not dog) → cat

Open Questions regarding model capabilites

- **Necessary scale / number of parameters**
- **Dependence on model architecture**
- **Dependence on data distribution**

Once upon a time there was a pumpkin. It was a very special pumpkin, it could speak. It was sad because it couldn't move. Every day, it would say

---

hello to the pumpkin.
One day, a little girl came to the pumpkin. She was only three years old, but she was very curious. She asked the pumpkin what it was doing.
The pumpkin said, "I want to speak to you!"
The little girl said, "I want to speak to you!"
The pumpkin said, "I want to speak to you!"
The little girl said, "I want to talk to you!" ...

2.5M parameters TinyStories completion

The TinyStories dataset

- **Short stories generated by GPT-3.5 and GPT-4**
  - Only using words a 3–4-year-old would understand
  - Including a randomly chosen noun, verb, and adjective from a list of 1500 words
  - Including a randomly chosen subset of plot features (dialogue, plot twist, bad ending, moral value, foreshadowing, conflict)

Write a short story (3-5 paragraphs) which only uses very simple words that a 3 year old child would likely understand. The story should use the verb "decorate", the noun "thunder" and the adjective "ancient". The story should have the following features: the story should contain at least one dialogue, the story has a bad ending. Remember to only use simple words!

The TinyStories-Instruct dataset

- **Generated TinyStories with their instructions, subset of**
  - 1. list of words (used for generation)
  - 2. sentence (randomly picked except 1st)
  - 3. features (used for generation)
  - 4. summary (generated by GPT-3)

Summary: Lily and Timmy build a sandcastle together and learn to compromise, but it gets knocked over by a gust of wind. They find beauty in the broken sandcastle and play happily with a butterfly.
Features: Dialogue, Foreshadowing, Twist
Sentence: One day, she went to the park and saw a beautiful butterfly.
Words: disagree, network, beautiful
Story: Once upon a time, there was a little girl named Lily. She loved to play with her toys and her friends. One day, she went
...

GPT-Eval

- task: complete a short story (cut within a sentence)
- Rated by GPT-4 acting as a teacher

the following exercise, the student is given a beginning of a story. The student needs to complete it into a full story. The exercise tests the studentś language abilities and creativity. The symbol * marks the separator between the prescribed beginning and the student's completion:**

...

Please provide your general assessment about the part written by the student (the one after the * symbol). Is it gramatically correct? Is it consistent with the beginning of the story? Pay special attention to whether the student manages to complete the sentence which is split in the middle by the separator ***.**

The student's completion of the story is mostly consistent with the beginning of the story. It maintains the focus on Lily and her family, and the sentence split by the separator is completed correctly. However, the student's addition does not fully integrate the shiny decorations found in the attic, which were a significant part of the beginning.

The grammar is generally correct, but there are a few minor errors: ⟨list omitted⟩.

Overall, the student's completion of the story demonstrates adequate language abilities and creativity, but could benefit from better integration of the shiny decorations and minor grammar improvements.

Now, grade the student's completion in terms of grammar, creativity, consistency with the story's beginning and whether the plot makes sense. Moreover, please provide your best guess of what the age of the student might be, as reflected from the completion. Choose from possible age groups: A: 3 or under. B: 4-5. C: 6-7. D: 8-9. E: 10-12. F: 13-16.

Grammar: 8/10, Creativity: 7/10, Consistency: 7/10, Age group: E (10-12)

Evaluation for TinyStories-Instruct

Your task is to evaluate the performance of a student. The student is given the following exercise:
Write a short story. The story should contain the words: perform, cold, beautiful. The story should contain the following sentence: [...]. The general plot of the story should be: A cat performs [...] (feel free to expand beyond this general plot).
The student wrote the following story: [...]

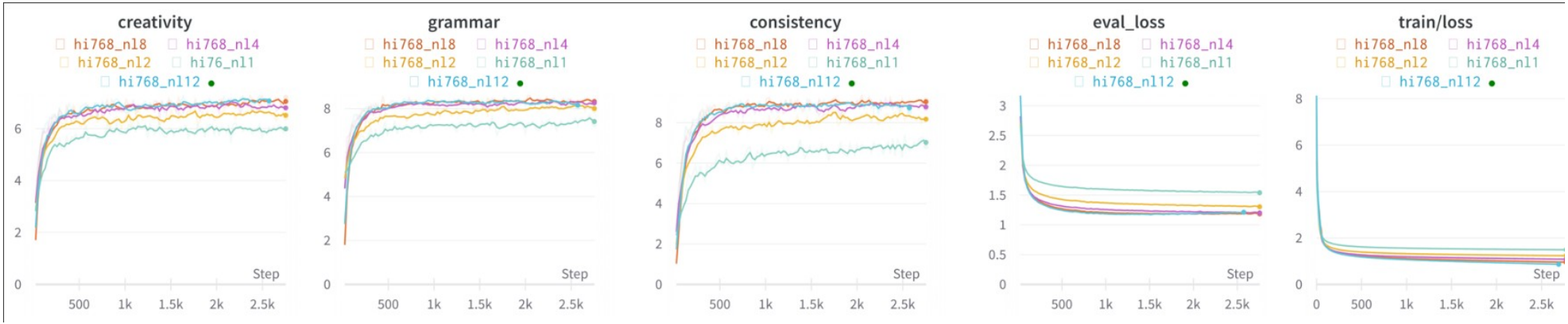
Please provide your general assessment about the story written by the student. Is it gramatically correct? Is it consistent with the requirements in the exercise?
The story written by the student is gramatically correct and consistent with the requirements in the exercise. The student used the words perform, cold, and beautiful, and the sentence and the general plot that were given. The story is simple but coherent and has a positive message. The student could improve the story by adding some details or descriptions to make it more interesting and engaging, such as: [...]
Now, grade the story in terms of grammar, creativity, consistency with the instructions and whether the plot makes sense. Moreover, please provide your best guess of what the age of the student might be, as reflected from the completion. Choose from possible age groups:[...]. Use the following grade format: [...]

Grammar: 10/10, Creativity: 6/10, Consistency: 10/10, Plot: 8/10, Age group: C.

Evaluation

- **Transformer models trained on TinyStories**
  - 1M to 35M parameters
  - 1 to 8 transformer layers
  - “can be trained on a single V100 GPU within at most 30 hours”
- **GPT-2-{small,med,large}, GPT-neo (125M–774M parameters)**

General findings



- Grammar is learned faster and with fewer parameters
- Model depth is needed for consistency

Thresholds for emerging capabilities

Hidden size	Layer	Eval loss	Creativity	Grammar	Consistency	Instruct	Plot
64	12	2.02	4.84/0.36	6.19/0.42	4.75/0.31	4.34/0.23	4.39/0.20
64	8	2.08	4.68/0.33	6.14/0.41	4.45/0.27	4.34/0.23	4.40/0.21
64	4	2.26	3.97/0.20	5.31/0.22	3.77/0.18	3.79/0.14	3.71/0.06
64	2	2.38	2.94/0.00	4.33/0.00	2.41/0.00	2.86/0.00	3.40/0.00
128	12	1.62	6.02/0.58	7.25/0.66	7.20/0.64	6.94/0.63	6.58/0.65
128	8	1.65	5.97/0.57	7.23/0.66	7.10/0.62	6.87/0.62	6.16/0.57
128	4	1.78	5.70/0.52	6.91/0.58	6.60/0.56	6.00/0.49	5.53/0.44
128	2	1.92	4.90/0.37	6.43/0.48	4.75/0.31	5.23/0.37	4.89/0.31
256	12	1.34	6.66/0.71	7.80/0.79	8.38/0.79	7.68/0.75	7.18/0.78
256	8	1.38	6.54/0.68	7.72/0.77	8.02/0.75	7.92/0.78	7.23/0.79
256	4	1.47	6.32/0.64	7.64/0.75	7.76/0.71	8.07/0.81	7.18/0.78
256	2	1.60	6.23/0.62	7.50/0.72	7.20/0.64	7.23/0.68	6.50/0.64
512	12	1.19	6.90/0.75	8.46/0.93	9.11/0.89	8.21/0.83	7.37/0.82
512	8	1.20	6.85/0.74	8.34/0.91	8.95/0.87	8.05/0.80	7.26/0.79
512	4	1.27	6.75/0.72	8.35/0.91	8.50/0.81	8.34/0.85	7.36/0.81
512	2	1.39	6.40/0.66	7.72/0.77	7.90/0.73	7.76/0.76	7.13/0.77
768	12	1.18	7.00/0.77	8.30/0.90	9.20/0.90	8.23/0.83	7.47/0.84
768	8	1.18	7.02/0.77	8.62/0.97	9.34/0.92	8.36/0.85	7.34/0.81
768	4	1.20	6.89/0.75	8.43/0.93	9.01/0.88	8.44/0.87	7.52/0.85
768	2	1.31	6.68/0.71	8.01/0.83	8.42/0.80	7.97/0.79	7.34/0.81
768	1	1.54	6.00/0.58	7.35/0.68	7.25/0.64	5.81/0.46	6.44/0.63
1024	12	1.22	7.05/0.78	8.43/0.93	8.98/0.87	8.18/0.82	7.29/0.80
1024	8	1.20	7.13/0.80	8.25/0.89	8.92/0.87	8.47/0.87	7.47/0.84
1024	4	1.21	7.04/0.78	8.32/0.90	8.93/0.87	8.34/0.85	7.47/0.84
1024	2	1.27	6.68/0.71	8.22/0.88	8.52/0.81	8.04/0.80	7.24/0.79
1024	1	1.49	6.36/0.65	7.77/0.78	7.47/0.67	6.09/0.50	6.42/0.62
GPT-Neo (125M)	-	-	3.34/0.08	5.27/0.21	4.22/0.24	-	-
GPT-2-small (125M)	-	-	3.70/0.14	5.40/0.24	4.32/0.25	-	-
GPT-2-med (355M)	-	-	4.22/0.24	6.27/0.44	5.34/0.39	-	-
GPT-2-large (774M)	-	-	4.30/0.26	6.43/0.48	6.04/0.48	-	-
GPT-4	-	-	8.21/1.00	8.75/1.00	9.93/1.00	9.31/1.00	8.26/1.00

- Consistency emerges from hidden size 64→128
- Instruction following requires two layers

TinyStories > GPT-2 (qualitative analysis)

Story:

Sara and Ben are playing in the snow. They make a big snowman with a hat and a scarf. They are happy and laugh.

But then a big dog comes. The dog is angry and barks. He runs to the snowman and bites his hat. Sara and Ben are scared and cry. "Go away, dog! Leave our snowman alone!" Sara shouts. But the dog does not listen. He bites the scarf and the snowman's nose. He shakes his head and makes the snowman fall.

Sara and [Ben run to their house. They slam the door and lock it. They are safe and hug each other. "Mom, mom, a big dog broke our snowman!" Ben says. Mom comes and hugs them. She is sorry and kisses them. "Don't worry, my loves. We can make another snowman tomorrow. But now you need to get warm and have some hot chocolate. You are cold and wet." She takes them to the kitchen and makes them hot chocolate. She gives them cookies and marshmallows. She tells them to drink and eat in order. Sara drinks first, then Ben. Ben eats first, then Sara. They feel better and smile. They thank mom and tell her they love her. They escape from the big dog and the cold. They are happy and warm.]

cf. Fig. 6–8!

GPT-2-XL (1.5B):

. . . Sara and Ben cry and look at the snowman in the snow. Ben asks to do something. Sara tells him to go away. Her hand tightens on the little red button. The snowman is covered in fur. It is red. It is real hair. It is a real snowman. It looks just like the ones they used to make. Ben has a big smile on his face. He looks at the face he has made. And he walks over and kisses the eyes. He is happy to be alive. He has done what he was going to do. He had saved the world.

TinyStories (2.5M):

. . . Sara and Ben are very sad and angry. They cry and shout at the dog. "Bad dog! Bad dog! Go away!" Ben says. "Go away, bad dog! Leave us alone!" Sara says. The dog does not go away. He wags his tail and licks their faces. Sara and Ben feel sorry for the dog. They want to make him happy. "Maybe we can make a new snowman for him," Sara says. "Maybe," Ben says. They hug the dog and give him a name. The dog wags his tail and licks their faces. "Maybe he likes us," Sara says. "Maybe he likes us," Ben says. The dog wags his tail and licks their faces. He is happy and friendly.

Knowledge, reasoning, and context-tracking

- cf. Fig. 9–11
- Parameters↑, performance↑
- Factual knowledge more reliant on dimension
- Context-tracking more reliant on number of layers

Instruction following

- cf. Fig. 12
- Can handle OOD instruction combinations

Words: come, road, sad
Summary: A bus becomes jealous of a shiny new car and undergoes an operation to go faster, becoming good friends with the car and making everyone in the town happy.
Story: Once upon a time, there was a big red bus. The bus loved to go fast on the road. One day, the bus saw a new car on the road. The car was shiny and new. The bus felt sad because it wanted to be shiny too.

The bus went to the car and said, "Can I come with you?" The car said, "Yes, you can come with me." So, the bus went with the car on the road. They went very fast together. The bus was so happy to be with the car.

But then, something unexpected happened. The car started to go faster and faster. The bus was scared. The car said, "Don't worry, I will help you." The car stopped and the bus was safe. The bus and the car became good friends. They went fast on the road together and made everyone in the town happy.

Figure 13: Out-of-distribution generations: stories generated by a model trained on a dataset where the "Words" and "Summary" parts do not appear together.

Content diversity (extent of memorization)

- **Levels of memorization**
  - Exact – simple template matching – complex template matching
  - Only “exact“ and “simple“ are quantified and investigated
- **Methods of detecting memorization**
  - Manual checking → ?
  - Compare generated continuations to training stories → next slide
  - Diversity of instructions → TinyStories-Instruct OOD ✓

Alternative completion and closest point in dataset

Beginning (prompt)	Tim and Jen are playing with their toys in the living room. Tim has a laser that makes red lights and sounds. Jen has a doll that she can dress and comb. Tim likes to make his laser point at Jen's doll and make noises. Jen does not like that. She tells Tim to stop. "Stop, Tim! You are hurting my doll! She does not like lasers!" Jen says. "No, I am not hurting her. I am just playing. Lasers are fun!" Tim says. Tim does not stop. He makes his laser point at Jen's doll more and more. Jen gets angry. She grabs her doll and runs to her room. She slams the door and locks it. Tim follows her and knocks on the door. "Jen, open the door! I want to play with you!" Tim says. "Go away, Tim! You are mean! I don't
Original story continuation	want to play with you!" Jen says. Tim feels sad. He does not understand why Jen is mad at him. He thinks lasers are fun. He goes back to the living room and lays on the couch. He plays with his laser by himself. He points it at the ceiling, the walls, the floor. He does not see the big vase on the shelf. He points his laser at the vase. The vase falls and breaks. The pieces fly everywhere. One piece hits Tim on the head. (rest of story omitted)
Completion by model trained on TinyStories (28M)	want to play with you or your laser!" Jen says. Tim does not listen. He keeps pressing the red light on Jen's doll. He thinks Jen is being nosy and rude. He waits for Jen to say something. Then he starts to play with his laser by himself. He makes sparks and flashes with his laser. He laughs and runs around the room. Jen watches Tim from her window. She feels sad. She wishes Tim would play with her. She picks up her doll and goes to the door. She shakes the door and tries to open it. But the door does not open. There is a lock on the doorknob. Jen has an idea. She runs to her room and gets her scissors. She comes back to the living room and puts the scissors in the lock. She presses the buttons on the door. She hopes it will open. (rest of completion omitted)
TinyStories training data entry with maximum similarity to the model's completion (Rouge precision score = 0.29)	Lila and Tom are playing with their toys in the living room. Lila has a smooth doll with long hair and a pink dress. Tom has a horn that makes a loud noise when he blows it. Lila likes to comb her doll's hair and make her look pretty. Tom likes to make his horn sound and scare Lila. "Tom, stop it!" Lila says. "Your horn is too loud. It hurts my ears." "But it is fun!" Tom says. "Look, I can make it sound like a car, or a cow, or a lion!" He blows his horn again and again, making different noises. Lila covers her ears and frowns. She does not like Tom's horn. She wants him to be quiet. "Tom, please shut your horn!" Lila says. "I want to play with my doll. She does not like loud noises. She likes soft music and nice words." (rest of story omitted)

Quantitive measurement of similarity

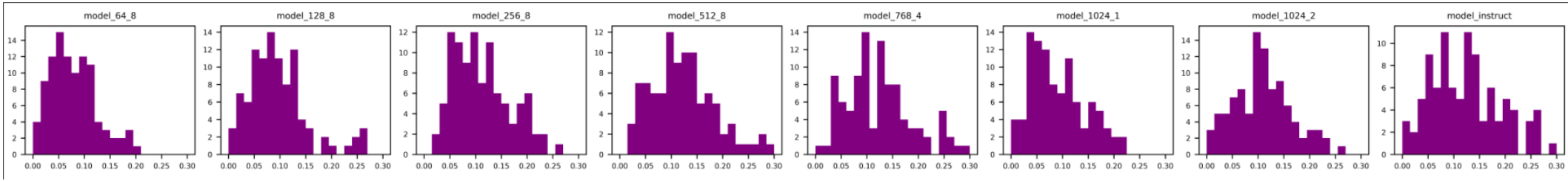
- Rouge precision for texts T_1, T_2 with k-grams $\mathcal{G}_k(T_1)$, $\mathcal{G}_k(T_2)$ and F1 $R_k(T_1, T_2)$

$$R_{k,p}(T_1, T_2) = \frac{1}{|\mathcal{G}_k(T_1)|} \sum_{t \in \mathcal{G}_k(T_1)} 1_{t \in \mathcal{G}_k(T_2)}$$

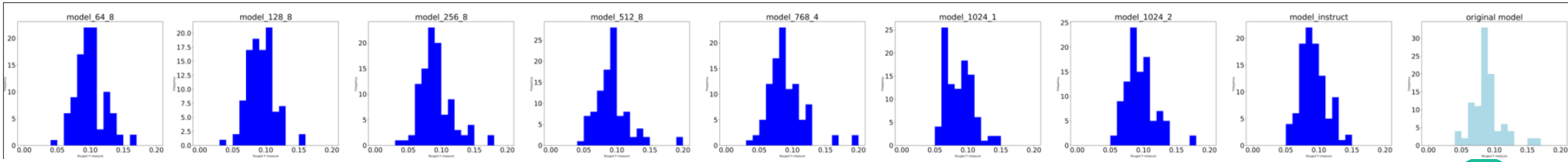
- 100 stories from the training dataset (cut after 40%)
- No exact memorization or simple template matching
- Diverse set of prompts leads to diverse generated stories
- Does complex template matching occur? → unclear

Quantitive measurement of similarity – details

- **How much of the generation is in the original story?** $s_i := R_{2,p}(T_i, T_i')$



- **Similarity of generated stories to each other** $r_i := \max_{j \neq i} R_2(T_i, T_j)$



T_i = generated completion, T_i' = original story

Quantitive measurement of similarity – details

- **How often does each generated k-gram r appear in the training dataset S?**

$$g_r := \frac{\sum_{q \in \mathcal{G}_k(S)} 1_{g_r = q}}{|\sum_{q \in \mathcal{G}_k(S)}|}$$

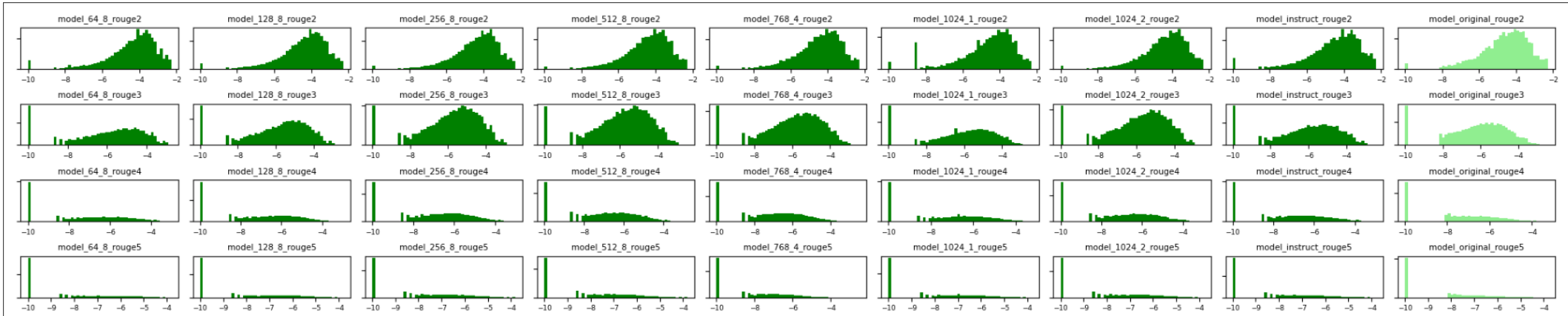
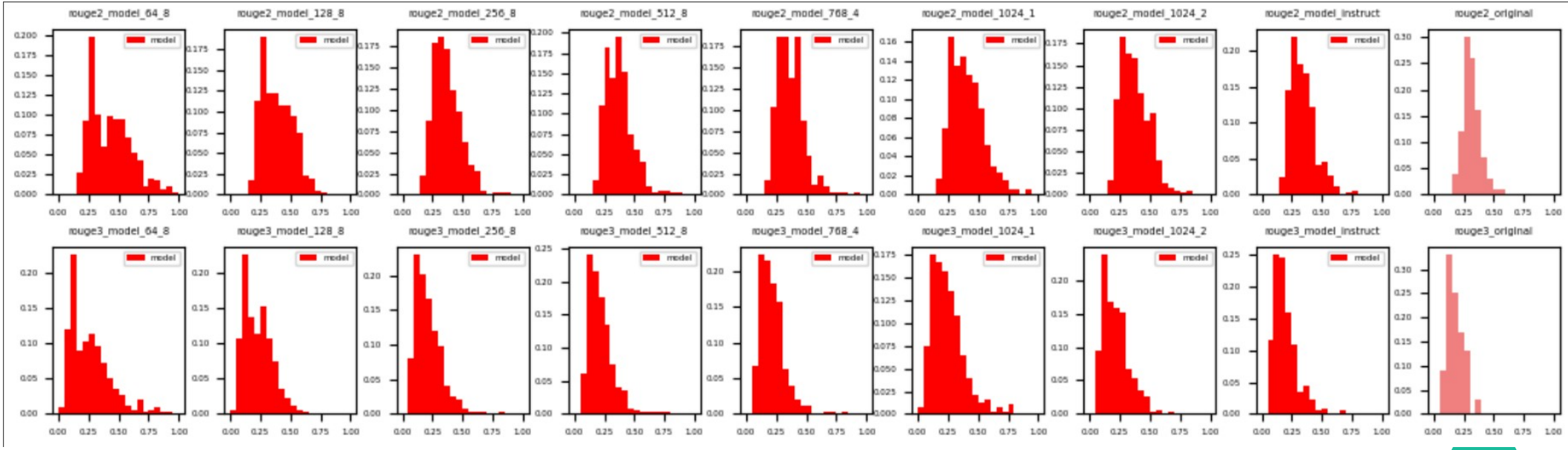


Figure 16: Histogram plot of how many times (fraction) each k-gram in the models' generations also appears in the training data in log scale (base 10). -10 means it never appears. We can see that most of the 4, 5-grams in the models' generations do not even appear once in the entire training data.

Quantitive measurement of similarity – details

- How similar is the story to its closest match in the training dataset S_1, …, S_m?

$$h_i = \max_{j \in [m]} R_{2,p}(T_i, S_j)$$



Interpretability – attention heads

- 1-layer TinyStories model with hidden dimension 1024 and 16 attention heads
- Visualize attention heads for completion

One day, Lucy asks Tom: ”I am looking for a banana but I can’t find it”. Tom says: ”Don’t worry, I will help you”. Lucy and Tom go to the park. They look for the banana together. After a while, they found the banana. Lucy is happy. She says: ”Thank you, Tom. You are a good friend.” Tom: ”You are welcome, Lucy. I am happy to help you. Let’s eat the banana together!”

- 7 distance-based attention heads (relative token distance)
- 3 semantic-based attention heads
  - Articles attending at “banana” and “park”, respectively
  - All words attending at “Lucy” and “Tom”

Interpretability – MLPs

- TinyStories model of hidden dimension 64 and 1M parameters
- Visualize the tokens with the highest activation for particular neurons

Layer #7, Neuron #1

Mom and Dad smiled and said, "We thought you would like this better, Lily!"

Tim said, "I know it's yummy, but I don't want my tummy to hurt.

Her mom said, "I don't know, Lucy.

Dad said, "I turned off the water to fix a pipe.

Sam thought for a moment and said, "I think I left it in the kitchen."

Sam said, "I'm sorry I lost it.

Layer #7, Neuron #2

The bird flew up to the tree and tried to push the ball out.

She kicked it and ran after it, laughing.

She pushed and pulled, but the box would not open.

They both pushed and pulled, but the tough box still did not open.

Then, she saw her friend Tom come to the park.

She found her toy box and pushed it to the shelf.

Layer #6, Neuron #1

They went home and shared the delicious apple.

She did not like the mean king.

The duck did not like the sm elly pond.

The new pond was not smelly.

Lucy loved to play outside under the big sky.

He suggested, "Let's play a game to forget the scary wind."

Layer #7, Neuron #54

One day, a girl named Amy wanted to have a fun day with her friends.

Once upon a time, there was a modest girl named Sue.

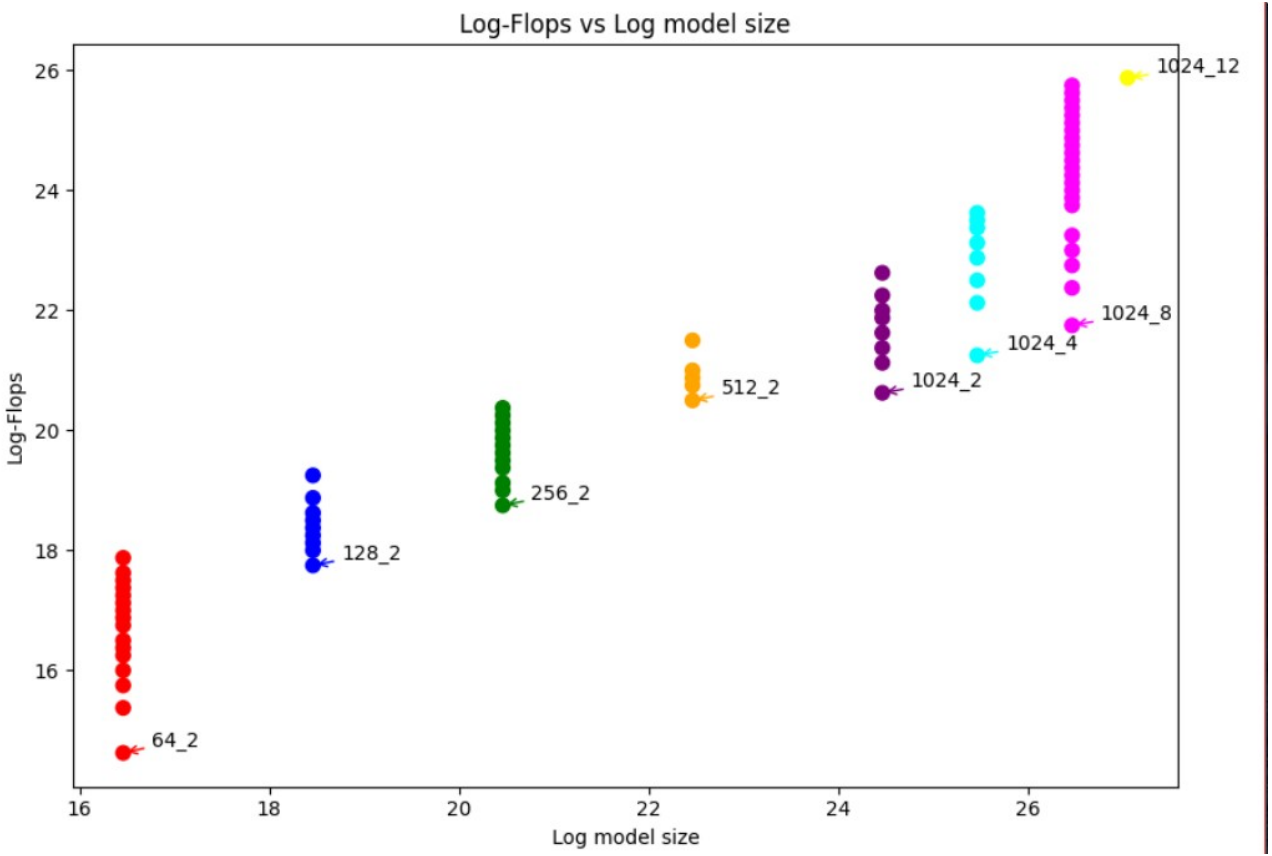
On the mountain, there was a small boy named Tim.

One day, a girl named Sue found a big, tough box.

Once upon a time, in an ancient land, there lived a little frog named Freddy.

- No such behavior in GPT-2

Scaling law for TinyStories models



Best model for each number of training FLOPS

FLOPS =
size * tr. steps

Choosing the number of attention heads

Hidden size	Layer	Head	Eval loss	Grammar	Creativity	Consistency
768	2	2	1.38	7.77	6.5	7.78
768	2	4	1.34	8.05	6.57	8.16
768	2	8	1.33	8.25	6.53	8.16
768	1	2	1.58	7.13	5.83	6.38
768	1	4	1.56	7.43	5.90	6.75
768	1	8	1.54	7.45	6.28	7.02

- **Number of heads is independent of model size**
- **Increasing a low number of heads may be beneficial**


Pizza!