TinyStories : Pushing the Limits? A Critical Analysis of the Dataset, Evaluation & Model Architecture

ANLP | Monsoon 2024

nvm. Nanda, Vaishnavi & Monish

02

background & motivation

background & motivation

The Original Paper | Eldan and Li (2023)

- Introduced **TinyStories** , a synthetic **dataset** of short stories that can be understood by 3- to 4-year olds (simple language & sentence structure, limited vocabulary).
- Explores how very **models much smaller than SOTA** produce coherent, consistent and diverse stories with near-perfect grammar.
- Employs **GPT-4 eval**, proposes that it overcomes flaws of traditional benchmarks that demand structure and are not multidimensional.
- Authors hope this can facilitate research into LMs for low-resource or highly specialised domains.

background & motivation

Our Project | Questions We Explored

- How complex does a model have to be to generate coherent language?
- Specific characteristics of the dataset that very small models when exposed to, can serve as sufficient for generating fluent text
- Is the transformer architecture even necessary, and can certain language tasks be reduced even further for simpler models to perform
- How do more traditional eval techniques hold up? Do they merit the use of GPT-eval?

table of contents

02

what we set out to do (and why)

what we set out to do (and why)

- We first set out to **recreate the baselines** , by training a few versions of a standard transformer architecture. How does sentence generation change with epoch? What does the loss look like?
- Then, we set out to see **how simple** a model can truly be? And trained a linear model on TinyStories.
- Is the **dataset** really that good as they claim? We analyzed and cleaned the data.
- The paper relies on GPT-eval. is this as robust? Can we find alternative **evaluation metrics** ? What is required for analyzing an eval metric?
- **Can we actually use TinyStories as a dataset for training story telling?**

03.

models & training

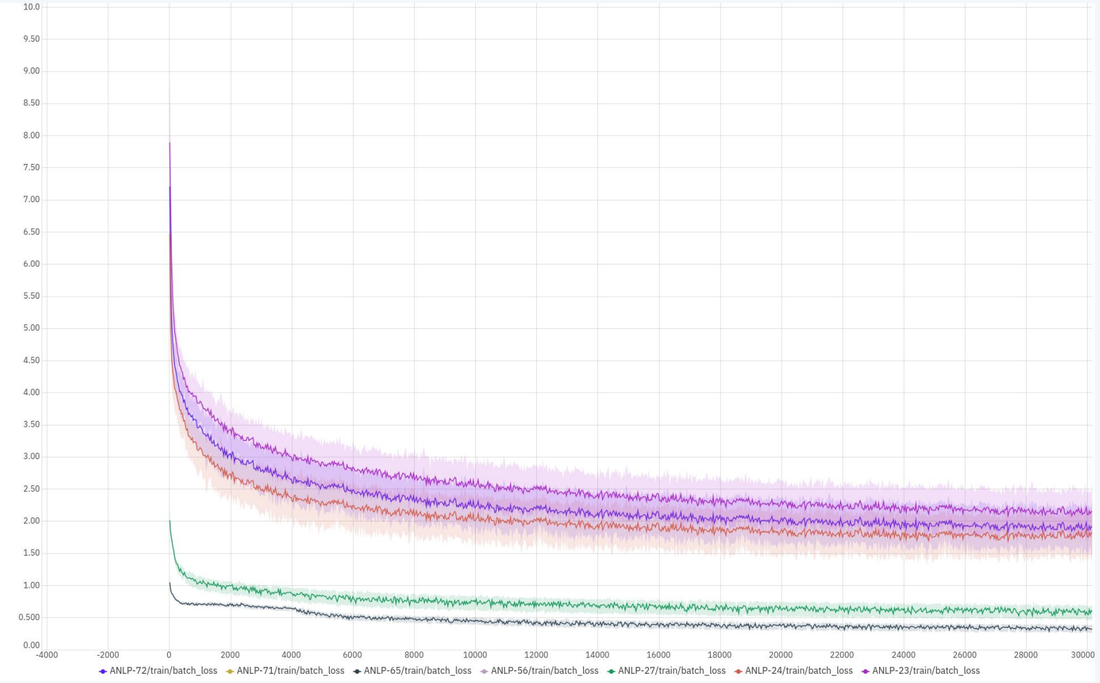
Model & Training

Like everyone else, we too started off with a Transformer. Because, duh.

We trained on several ablations:

1. Varying model dim from 256 to 1024 dimensions, keeping everything constant, training on 10% of the data
2. Training on 100% of the data for 1 epoch, 512 dimensions and everything else constant
3. Training on 100% of the data for 2 epochs, 768 dimensions, 8 heads, 12 layers

Model & Training

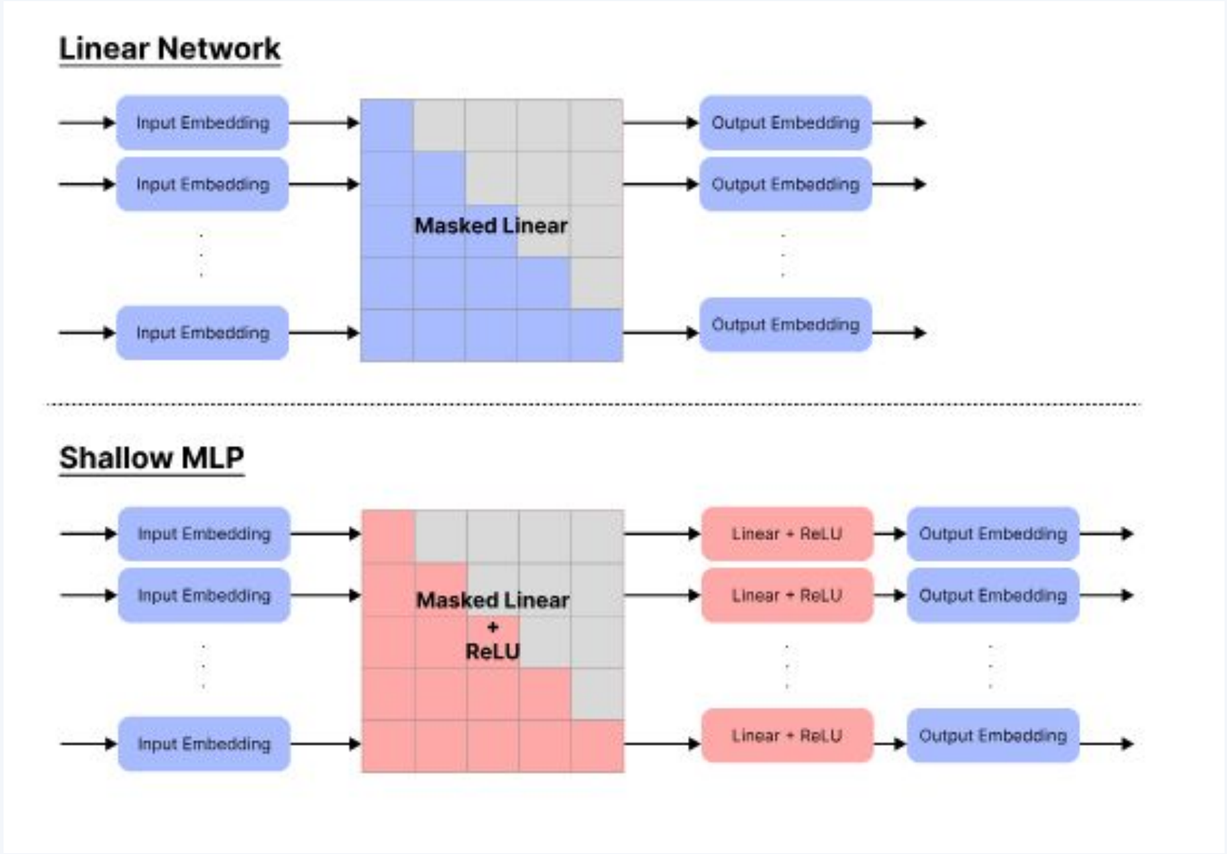


Loss curves for training

It was observed that there were no *significant* differences while training with more epochs. The evaluations also back up this observation.

model & training

but then, we stumbled upon this...



Prompt #1: He was very scared

Output: *He was very scared. He cried out loud, but then he heard a voice. He was afraid of the dark monster. He ran to the bear. He had a pain in his hand. He screamed and ran away. He was never seen again. He was scared of the monster and he went on. He never...*

Prompt #2: Adam was hungry, so

Output: *Adam was hungry, so he wanted to eat the dessert. But he knew it was not a good at his sister . He thought, "Maybe I can eat this food, but I don't want to share with you."*

Prompt #3: Alice was tired, so

Output #1: *Alice was tired, so she decided to take a nap . She put a blanket on the ground and started to cry. Then, she heard a noise. It was a nearby tree.*

Output #2: *Alice was tired, so she decided to go on an adventure . She hopped on the way to go home and look for her...*

MODEL	GRAMMAR GPT-4 / LT	CREATIVITY GPT-4	CONSISTENCY GPT-4	PLOT GPT-4
TS-33M	8.0 ± 0.8 / 62%	7.2 ± 0.5	7.0 ± 1.2	6.9 ± 0.8
TS-1M	6.9 ± 0.9 / 59%	6.7 ± 1.0	6.0 ± 1.5	5.6 ± 1.3
LINEAR	6.3 ± 2.0 / 64%	6.2 ± 1.8	5.9 ± 1.8	5.2 ± 1.8

Figure 2. **Top**: Example prompts and outputs for Linear model trained on TinyStories (grammatical/conceptual errors in red). **Bottom**: Comparison between Transformer-and Linear models, average grades from GPT-4.

04.

autoregression is all you need?*

autoregression is all you need?*

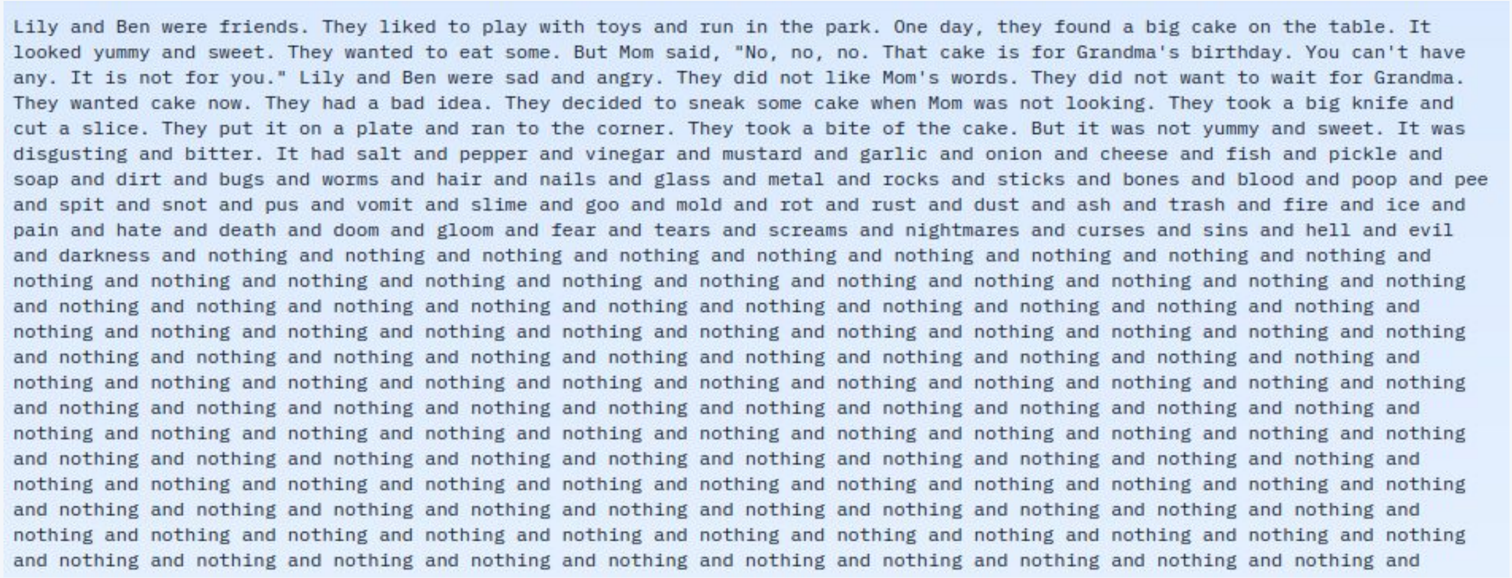
- Autoregression is a really simple way of analyzing nonlinear patterns in a dataset. The paper by ***Malach (2023)*** shows us that AR can **approximate** any function computable by a Turing Machine.
- With each iteration, we provide more information to our model. By constantly updating our input and label information, we can model complex nonlinear functions and patterns - through the simple task of autoregression!
- We can picture these *iterations* as **a form of chain-of-thought reasoning**.
- Example of a sentence generated from a linear model:
  Once upon a time, there was a little girl named Lily. She loved to play outside in the park. One day, she saw a big tree with lots of leaves. She wanted to climb it, but she couldn't climb it. Suddenly, she heard a voice from behind her…..

04.

game, dataset, match

game, **dataset** , match

- Dataset claimed to use ~1.5k unique words in order to mimic a child's vocabulary, but both a manual inspection and a simple word count failed this test.
- Dataset not cleaned properly
  - Very short or very long stories were not stories
- Weird characters show up randomly

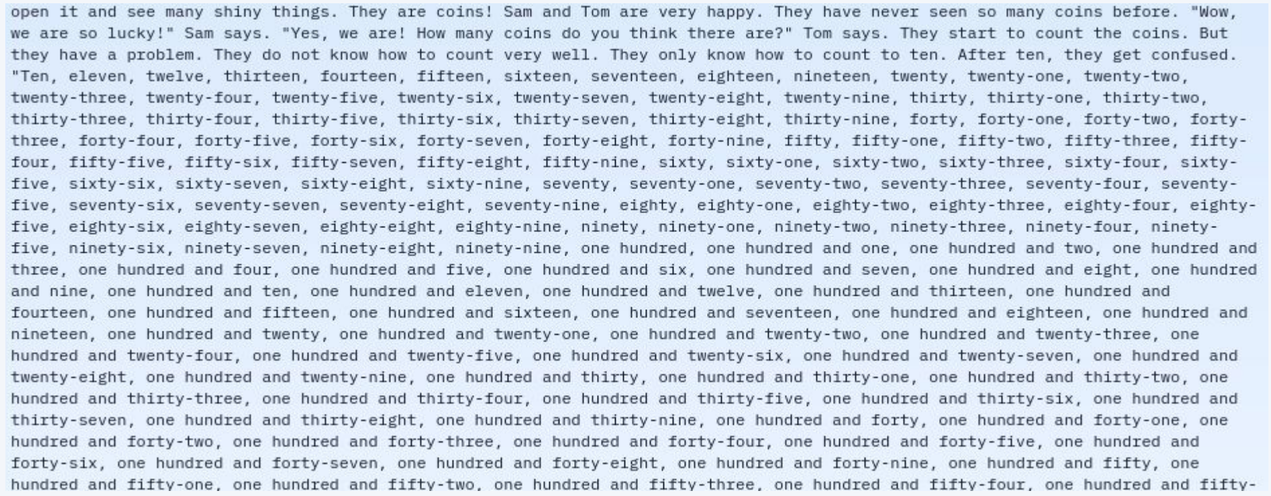


Lily and Ben were friends. They liked to play with toys and run in the park. One day, they found a big cake on the table. It looked yummy and sweet. They wanted to eat some. But Mom said, "No, no, no. That cake is for Grandma's birthday. You can't have any. It is not for you." Lily and Ben were sad and angry. They did not like Mom's words. They did not want to wait for Grandma. They wanted cake now. They had a bad idea. They decided to sneak some cake when Mom was not looking. They took a big knife and cut a slice. They put it on a plate and ran to the corner. They took a bite of the cake. But it was not yummy and sweet. It was disgusting and bitter. It had salt and pepper and vinegar and mustard and garlic and onion and cheese and fish and pickle and soap and dirt and bugs and worms and hair and nails and glass and metal and rocks and sticks and bones and blood and poop and pee and spit and snot and pus and vomit and slime and goo and mold and rot and rust and dust and ash and trash and fire and ice and pain and hate and death and doom and gloom and fear and tears and screams and nightmares and curses and sins and hell and evil and darkness and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and
nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing
and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and
nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing
and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and
nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing
and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and
nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing
and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and
nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing
and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and
nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing
and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and nothing and

I wonder how this ends...

r/redditsniper vibes

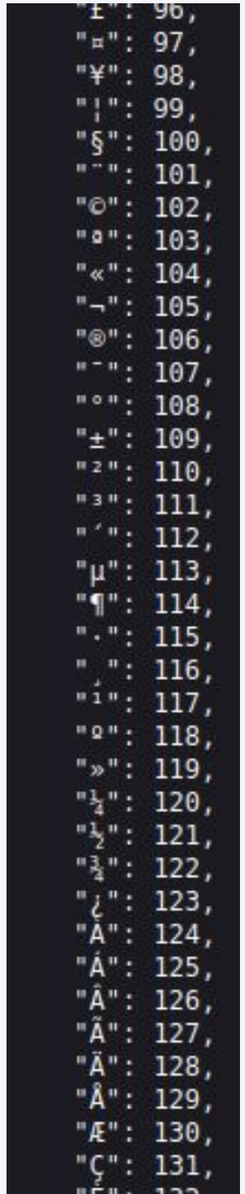
game, **dataset** , match



We learnt counting!

Once there was a little girl who loved to explore the outdoors. One day she went outside and observed a hose in her backyard. She was so excited to see the hose and decided to pick it up. Unfortunately, when she did, she noticed that the *hose was dead*. The girl felt very sad to see the hose not working and just sat there, looking at it for a few minutes. Then, after a few moments of **contemplation**, the girl decided to go back inside and find something else to play with.

Another gem from the dataset (randomly picked)



An actual snapshot from the tokenizer of the official reproduction of the TinyStories paper

(which, btw, has a vocab of 50k+)

some special mentions:

ħ

the symbol for the reduced Planck's constant

ãġ®æ

token number 27031

---

yes, it did not fit in a single line

game, **dataset** , match

One day, John was playing in his room when he felt a strange feeling in his mouth. He put his hand up to feel and noticed that he had a loose tooth in the back of his mouth. He was feeling a bit…
One day, John was playing in his room when he felt a strange feeling in his mouth. He put his hand up to feel and noticed that he had a loose tooth in the back of his mouth. He was feeling a bit…

The authors took great care in making the data and definitely did not include duplicate stories

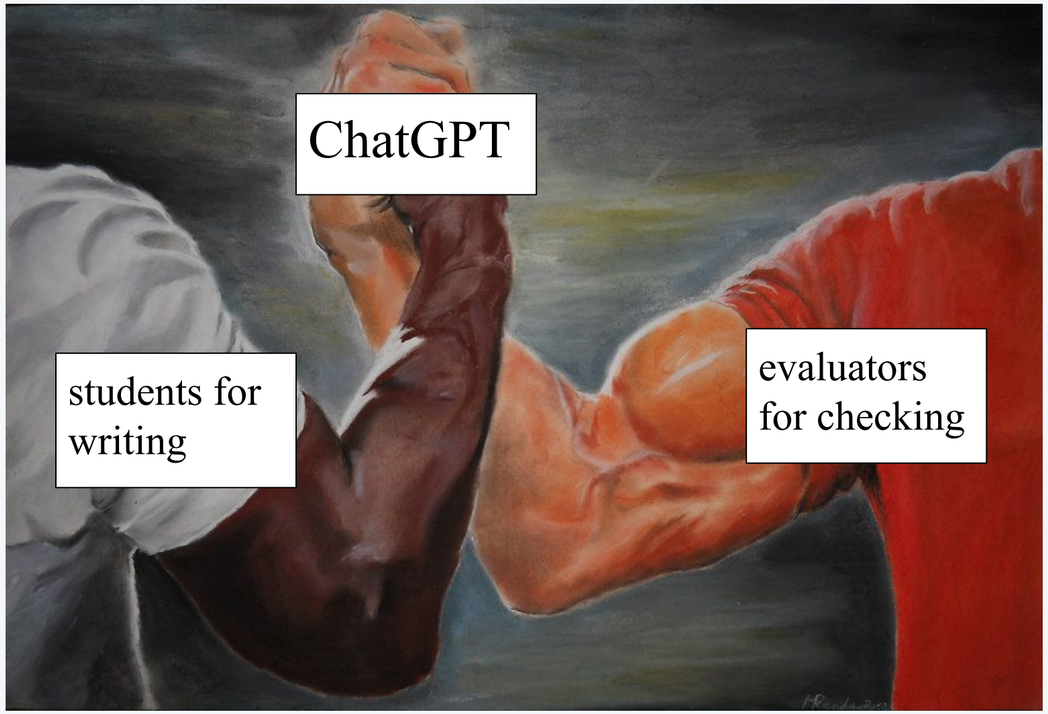
…intended to contain only words that most 3 to 4-year-old children would typically understand, generated by GPT-3.5 and GPT-4. TinyStories is designed to capture the essence of natural language, while reducing its breadth and diversity. Each story consists of 2-3 paragraphs that follow a simple plot and a consistent theme, **while the whole dataset aims to span the vocabulary and the factual knowledge base of a 3-4 year old child**…Moreover, such a dataset would facilitate the development and analysis of SLMs, especially for **low-resource or specialized domains, where large and diverse corpora are either unavailable or undesirable**.

04.

keeping score

keeping score

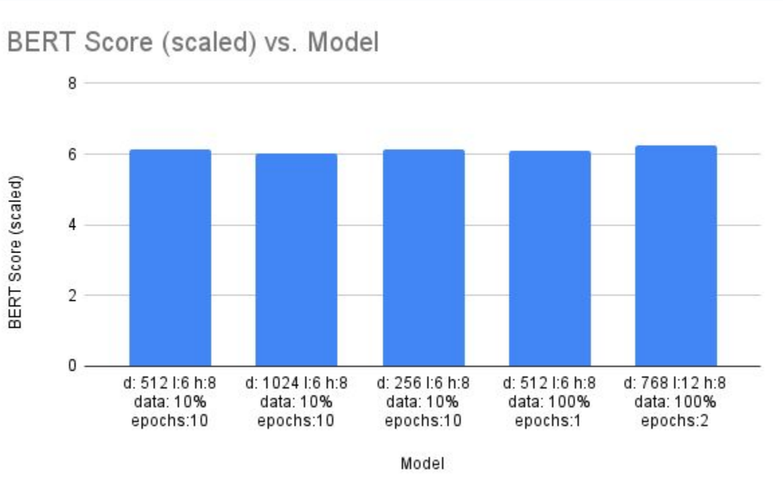
- What makes a good story?

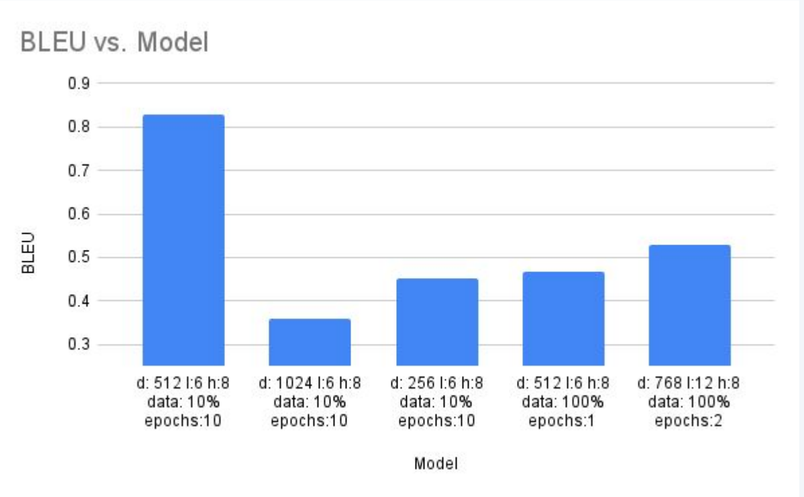


Using quantitative metrics for evaluation

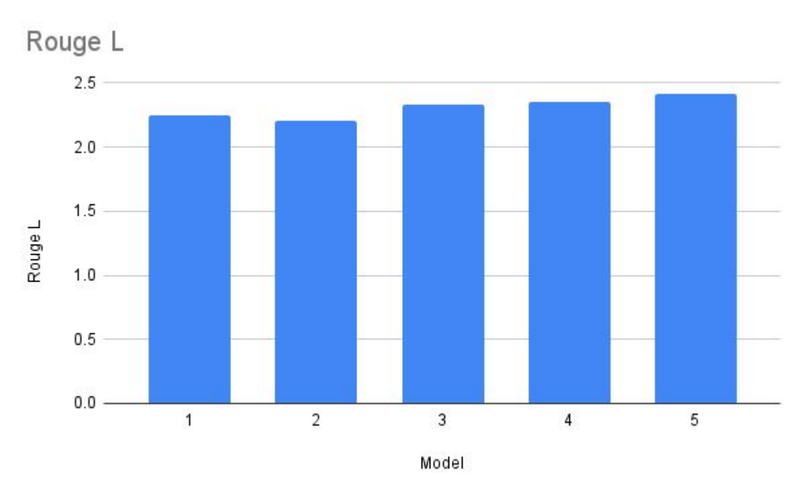
Using GPT

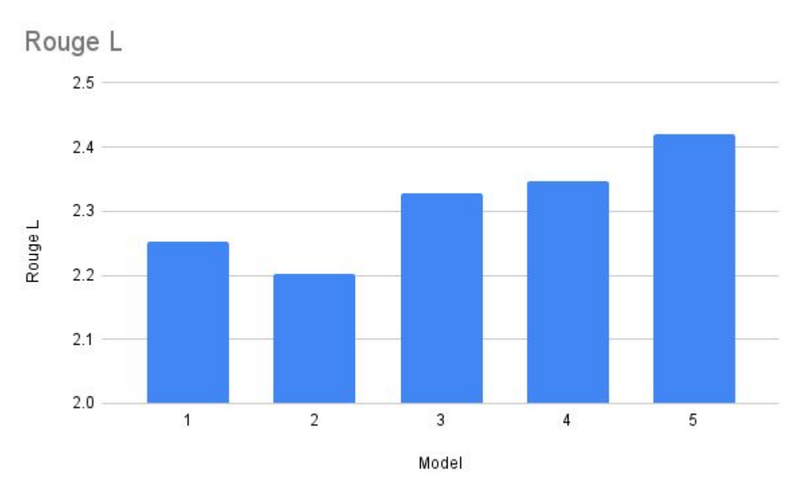
keeping score

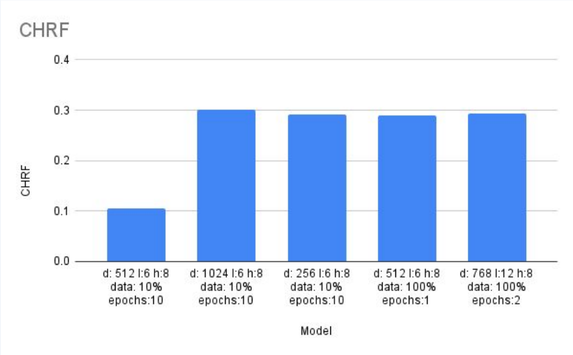


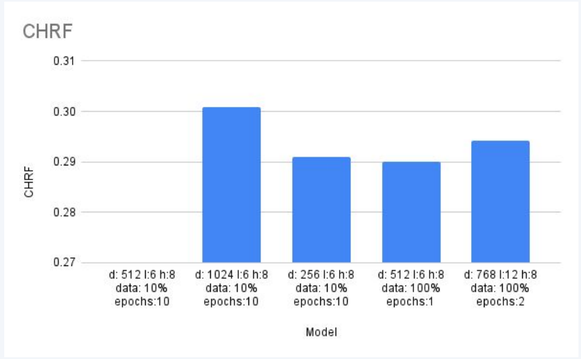


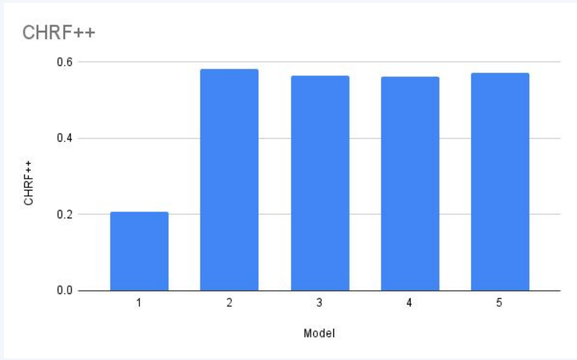
keeping score

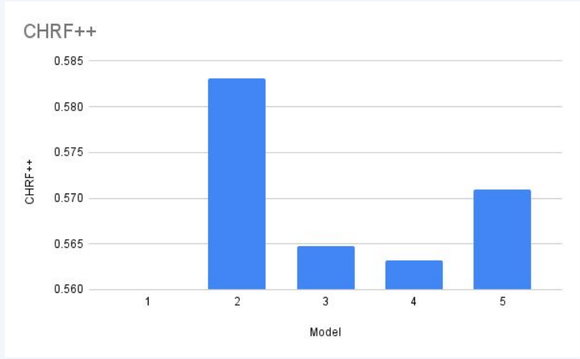




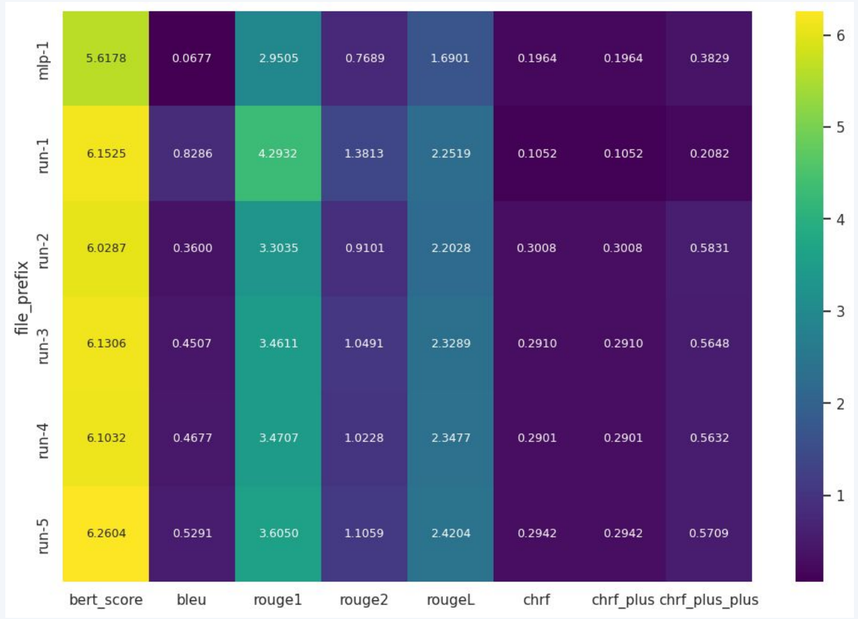


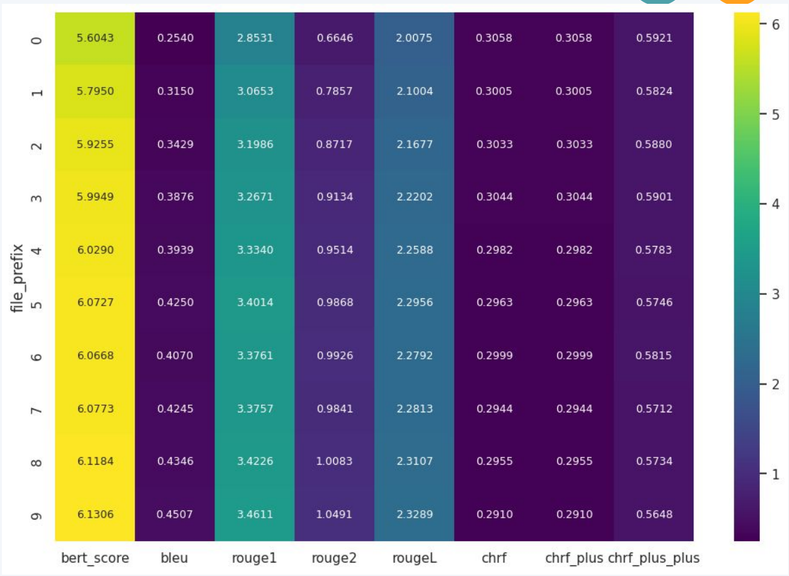


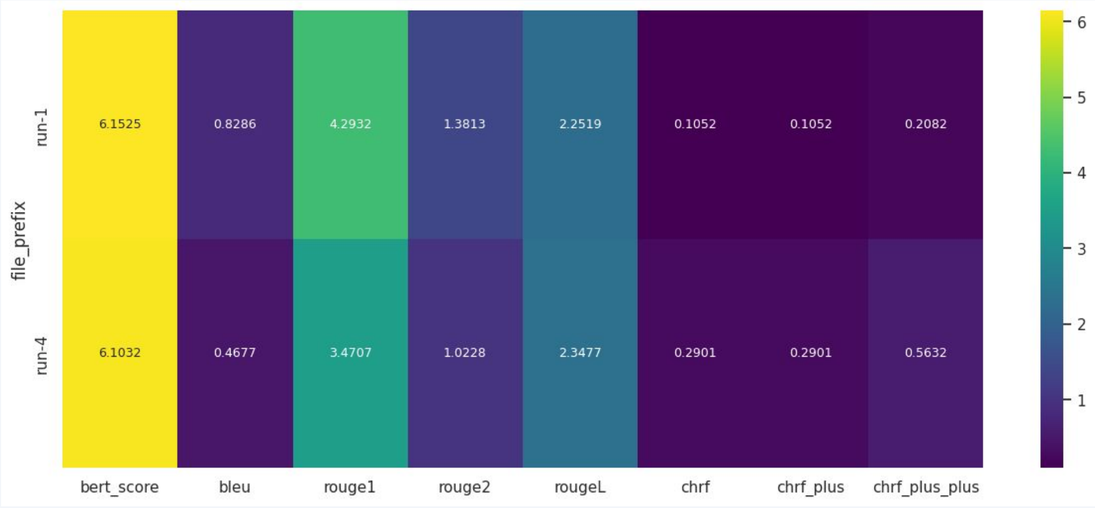




keeping score | visualising (trying)







keeping score

The following exercise tests the student’s language abilities and creativity.
The student is given a beginning of a story and is required to complete it.

*Please evaluate the part written by the student after the "***" symbol in terms of **grammar**, **creativity**, and **consistency** with the given prompt.*

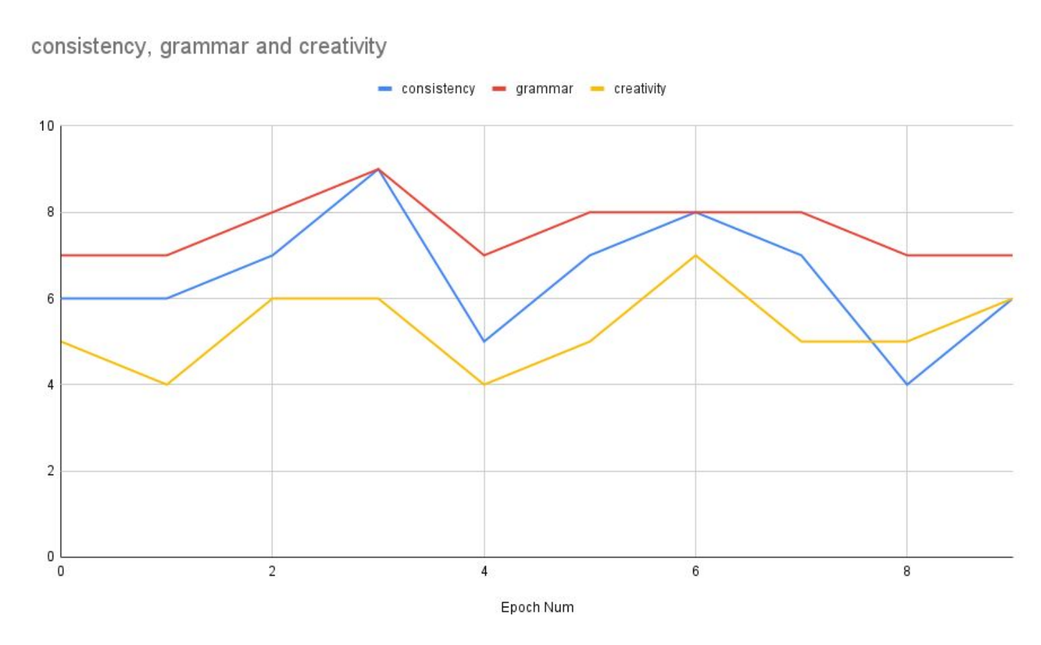
Here is the student's prompt and completion:
{}

Your task is to assess the following:
*- **Grammar**: Are there any grammatical errors?*
*- **Consistency**: Does the student’s completion logically follow from the beginning?*
*- **Creativity**: How creative or original is the student's addition to the story?*

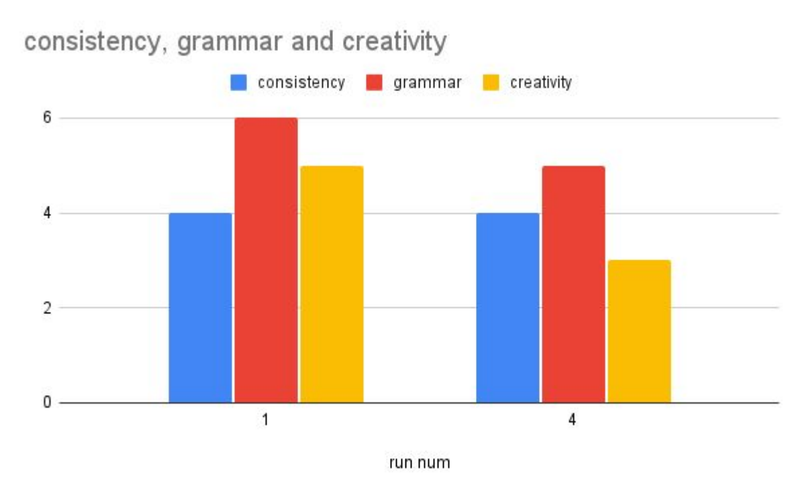
Give a score out of 10 for each:
- Consistency: X/10
- Grammar: X/10
- Creativity: X/10

Additionally, please provide an estimated age group based on the student’s completion (options: A: 3 or under, B: 4-5, C: 6-7, D: 8-9, E: 10-12):

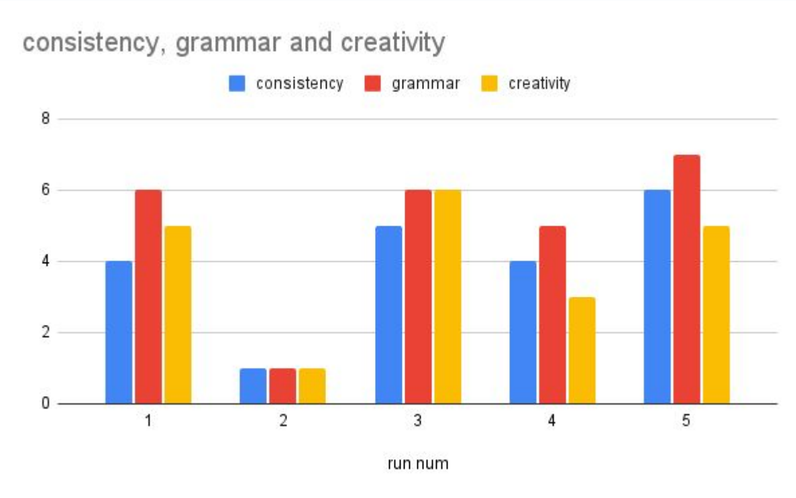
keeping score



epoch wise

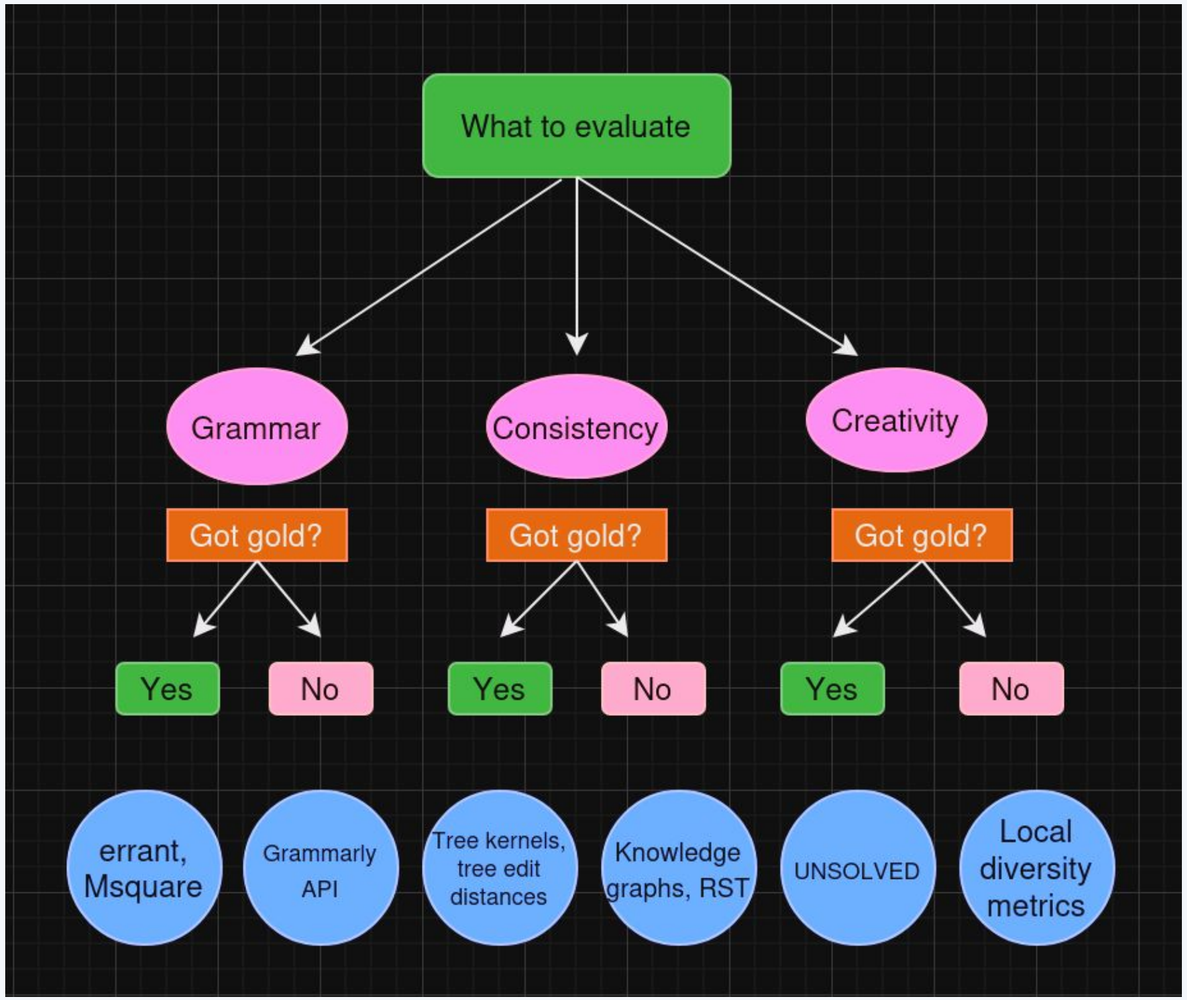


epoch count v. data



across models

keeping score

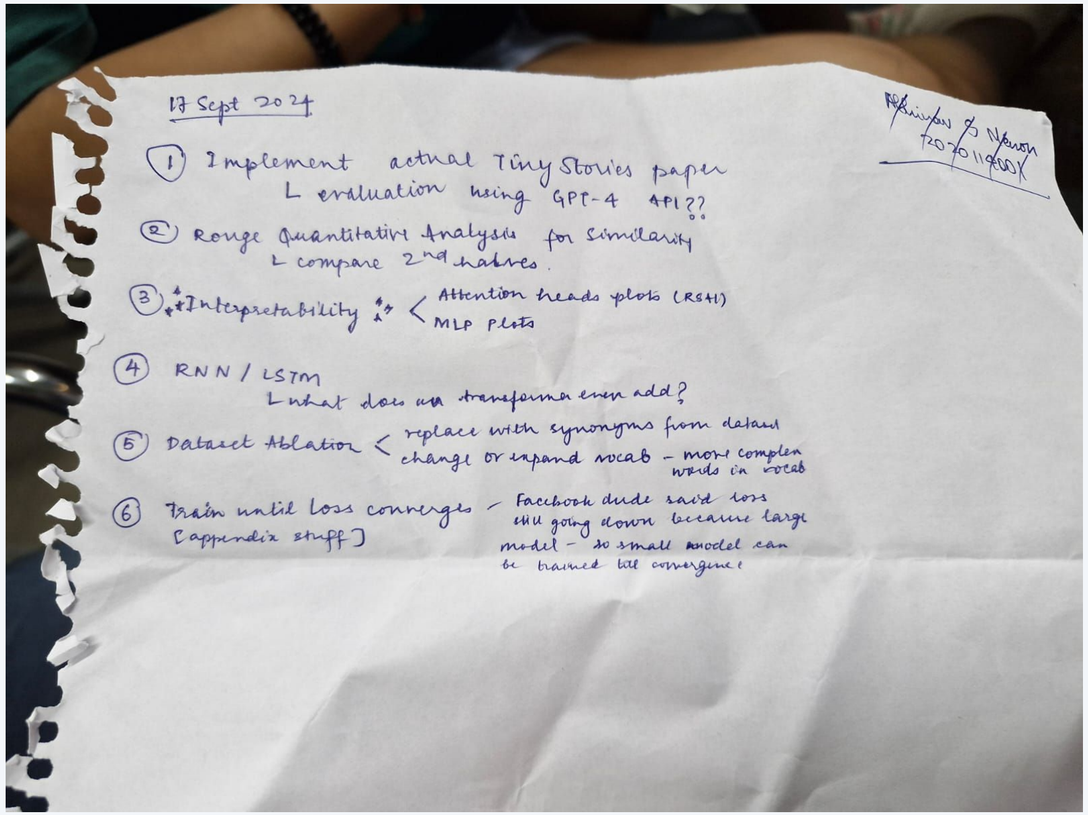


04.

limitations & future scope

limitations & future scope

- Look at + discover future metrics
- Interp
- using other models (including smaller linear models)
- does our models trained on tinyStories generalize for harder data?



a Nanda panic sheet

thank you

also, thanks to Sreeram Vennam for showing us that one paper