Efficient Methods for ML
Final Talk

Iason Kourouklis, Maarten Mäcking and Rickmer Weichenthal

Group: Not so Large Language Model

University of Hamburg

July 10, 2024

“The last talk.. was very successful. All of the people around the town were so happy to hear it.”

– 35M model

Roadmap

Motivation and Objectives

Objectives:

- develop language model to generate coherent short stories
- no memorization
- ideally no overfitting
- test different models (RNNs, Transformer)
- use different Datasets

(Data-) efficient small language model:

- run on Laptop
- restricted domain: children stories
- restricted language: simple english (we use 2048 tokens)

Roadmap

Language Characteristics

Brief description of *TinyStories*[1]

- short stories (2.1M)
- vocabulary of a 4-year-old
- encompass core elements of natural language (within a limited scope)
- GPT-3.5/4 generated (high similarity)

Each story should contain:

- a randomly picked noun, verb and adjective from a curated vocabulary
- a random subset of certain features (e.g., a bad ending)

[1] Ronen Eldan and Yuanzhi Li. "Tinystories: How small can language models be and still speak coherent english?" In: *arXiv preprint arXiv:2305.07759* (2023).

Impurities

- Encoding Errors (e.g., "â€š" instead of ",")
- Typographical Errors (e.g., febuary, luggaage)
- Empty Stories [0.01%[2]]
- Duplicate Stories [15.13%[2]]
- Non-ASCII Symbols [0.16%[2]]

[2] Percentage of stories removed from the training set for this reason

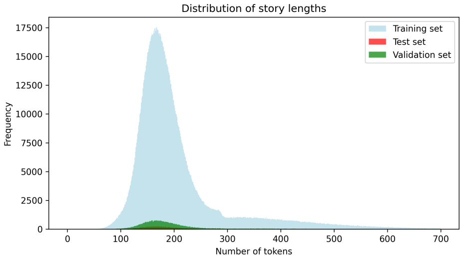
Preprocessing

- Character normalization (e.g. replacing '\u200b' with '')
- Removing unwanted characters (e.g. asterisks (*))
- Length constraint (min 180 characters)
- Split data into training, validation and test sets
- bos and eos tokens

Statistics

Property	Training set	Validation set	Test set
# of stories	1,725,028	73,728	21,989
# of tokens	360,305,047	15,396,519	4,481,695
# of unique tokens	54,916	19,113	13,050
avg. seq. length	208.9 (± 101.2)	208.8 (± 101.2)	203.8 (± 96.0)

- We use a max. seq. length of 256



Data Varient

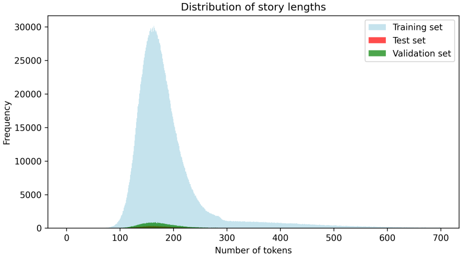
- TinyStoriesV2-GPT4 Dataset (valid and test Data)[3]
- only GPT 4 generated
- bigger and "better" Dataset
- same (no specific) preprocessing

[3] Hugging Face. *roneneldan/TinyStories*. 2023. URL: https://huggingface.co/datasets/roneneldan/TinyStories/tree/main.

Statistics

Property	Training set	Validation set	Test set
# of stories	2,643,469	73,728	27,630
# of tokens	512,144,366	14,292,021	5,315,794
# of unique tokens	47,755	15,583	11,419
avg. seq. length	193.7 (± 84.2)	193.8 (± 83.9)	192.4 (± 81.7)

- We use a max. seq. length of 256



Roadmap

Tokenization

Tokenizer: `spacy` from spaCy[4]

- word-level tokenizer
- vocabulary: 2,048 most common tokens in training set
- " are not stripped away (compared to `basic_english`)
- curly " are unique tokens and replaced
- words are split at -

Tokenizer used in reference paper[5]:

- GPT-Neo tokenizer (10,000 most common tokens)

[4] spaCy. *spacy tokenizer*. 2024. URL: https://spacy.io/api/tokenizer.
[5] Ronen Eldan and Yuanzhi Li. "Tinystories: How small can language models be and still speak coherent english?" In: *arXiv preprint arXiv:2305.07759* (2023).

spacy vs basic_english

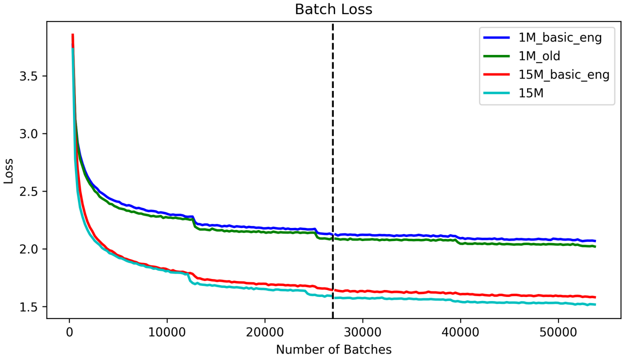
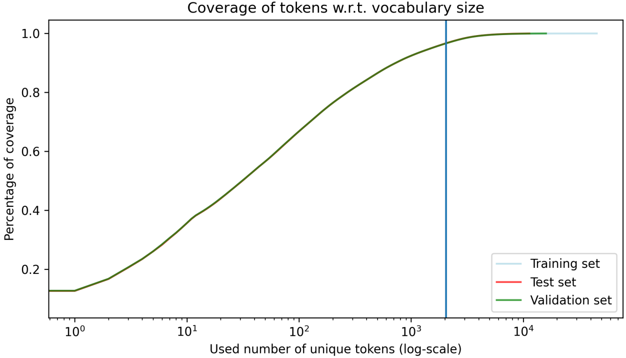


Figure: spacy for better results

Tokenization (TinyStoriesV1)

How many tokens can we represent with a given vocabulary size?

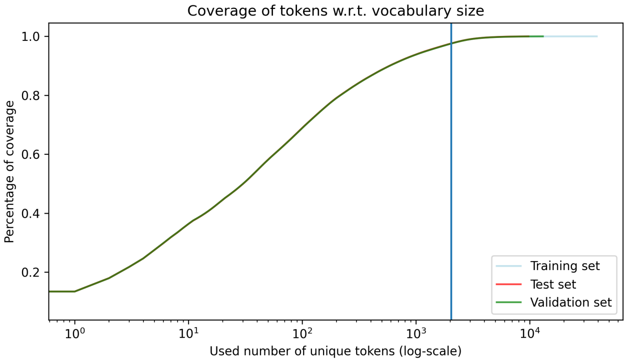


We use 2048

- 50% of unique tokens appear ≤ 5 times
- 25% of unique tokens appear ≤ 1 times

Tokenization (TinyStoriesV2)

How many tokens can we represent with a given vocabulary size?



We use 2048

- 50% of unique tokens appear ≤ 6 times
- 25% of unique tokens appear ≤ 2 times

Roadmap

Embedding

Using pytorch's `nn.Embedding` layer

- $V \times$ `embed_size` parameters for the embedding

Positional encoding:

- Sinusoidal from original transformer paper[6]
- Rotary using `RotaryPositionalEmbeddings` from `torchtune.modules`
  - Rotation matrix applied to embedding layer (not optimal, different from original)[7]
  - Rotate embeddings by $m\times$ angle θ
  - Captures relative positions

$$f_{\{q,k\}}(x_m, m) = \begin{pmatrix} \cos m\theta & -\sin m\theta \\ \sin m\theta & \cos m\theta \end{pmatrix} \begin{pmatrix} W^{(11)}_{\{q,k\}} & W^{(12)}_{\{q,k\}} \\ W^{(21)}_{\{q,k\}} & W^{(22)}_{\{q,k\}} \end{pmatrix} \begin{pmatrix} x^{(1)}_m \\ x^{(2)}_m \end{pmatrix}$$

[6] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[7] Jianlin Su et al. "RoFormer: Enhanced transformer with Rotary Position Embedding". In: *Neurocomputing* 568 (2024), p. 127063. DOI: https://doi.org/10.1016/j.neucom.2023.127063. URL: https://www.sciencedirect.com/science/article/pii/S0925231223011864.

Positional Encoding

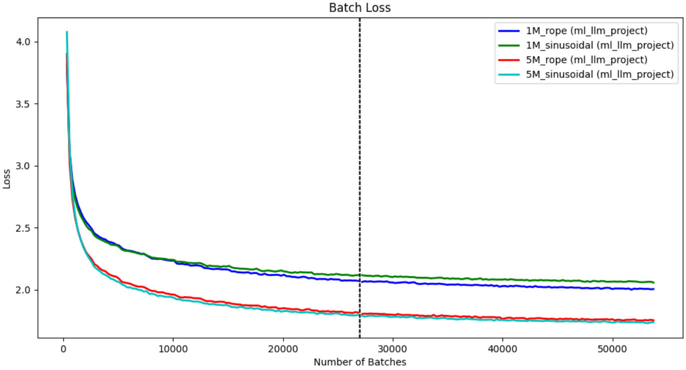


Figure: RoPE better for smaller models

Roadmap

Evaluation Metrics

Quantitative metrics:

- Cosine Similarity
  - TF-IDF vectorization
- ROUGE-1, -2, -L (from TinyStories)
  - using `rouge_score` from `rouge_scorer`

Qualitative metric:

- GPT-4-Eval: use GPT 4 to evaluate stories

Cosine Similarity & ROUGE

Memorization (story completion):

- **Prompts:** first **40%** of 100 randomly sampled stories
- **Reference Completion:** remaining **60%**
- Generate model completion of stories (only compare the completion)
- Compute average and standard deviation for Cosine Similarity and ROUGE scores
- fast to compute

Overfitting:

- Cosine similarity for generated stories against complete training set
- generally better but long computation

GPT-Eval

GPT4-Eval categories:

- grammar
- spelling
- consistency
- story
- creativity
- style

Design Criteria for best GPT prompt:

1. Ratings align with manual ratings.
2. Ratings are reliable (low SD).
3. Format is easy to parse

Automatic evaluation for GPT 4 (via UHHGPT & Selenium)

Prompt Engineering

Process:

- Manually rated 20 stories (50% used for prompt).
- Tested various prompts automatically to match our ratings (LLAMA-3).

Prompt Structure:

- introduction: story for 4 year old children, rate objectively
- Specifies categories: brief explanation, rate categories from 1-10 independently
- short list of "Important Instructions": emphasize criteria, format, not include anything else, emphasize consistency and objectivity
- 11 example stories (examples for both bad and good ratings).
- format to follow
- Remember: no additional text, explanations or comments after scores

```
The following story is a story written for children at the age of around 5 years old. Your task is to rate the story objectively.

Story:
Story:
    {}

Please rate the story in the following categories:

GRAMMAR (exclusively rate the grammar of the story)
SPELLING (exclusively rate the spelling of the story)
CONSISTENCY (exclusively rate the consistency of the story)
STORY (exclusively rate the story/plot of the story)
CREATIVITY (exclusively rate the creativity of the story)
STYLE (exclusively rate the linguistic style of the story)
Rate objectively and rate each category independently from the others. Rate each category with a score from 1 to 10, with 1 being the worst and 10 being the best.

Important Instructions:

Ensure that your ratings for each category are based solely on the criteria specified for that category.
End your response with the format specified below.
Do not include any text other than the ratings after the specified format.
Ensure consistency and objectivity in your ratings.
Examples:

Example 1:

Story:
"Tommy the turtle walked slowly to the pond. He met a friendly frog named Fred. Together, they swam and played games until the sun set. They found a hidden treasure chest by the pond. Inside were shiny stones and colorful shells. They promised to meet again the next day to play with their new treasures."

Ratings:
GRAMMAR: 10
SPELLING: 10
CONSISTENCY: 10
STORY: 8
CREATIVITY: 7
STYLE: 9

Example 2:

Story:
"A dragon flyed over the castle. The prince lookd up and waved. He wished he could fly too. The dragon landed and offered him a ride. They flew over the kingdom and saw many amazing sights. The prince felt like he was in a dream."

Ratings:
GRAMMAR: 7
SPELLING:7
CONSISTENCY: 9
STORY: 5
CREATIVITY: 6
STYLE: 5

Example 3:

Story:
"Once upon a time, there was a little mouse. The mouse loved cheese. One day, he found a big piece of cheese in a trap. He cleverly used a stick to get it out without getting caught. Then he shared the cheese with his friends. They had a big cheese party and danced all night."

Ratings:
GRAMMAR: 10
SPELLING: 10
CONSISTENCY: 10
STORY: 6
CREATIVITY: 5
STYLE: 6

Example 4:

Story:
"The stars shone brightly in the sky. It was a magical night. Lucy made a wish on a shooting star, hoping for a new puppy. The next morning, her wish came true when she found a puppy in her backyard. She named him Sparky. They became best friends and went on many adventures together."

Ratings:
GRAMMAR: 10
SPELLING: 10
CONSISTENCY: 9
STORY: 7
CREATIVITY: 7
STYLE: 8

Example 5:

Story:
"The robot walkeded into the room. It said, 'Hello, human!' Timmy was amazed and excited to meet a real robot. They played together all day and became best friends. The robot taught Timmy many things about space. At the end of the day, the robot had to leave, but it promised to come back."

Ratings:
GRAMMAR: 8
SPELLING: 5
CONSISTENCY: 8
STORY: 5
CREATIVITY: 7
STYLE: 5

Example 6:

Story:
"A magical unicorn danced under the rainbow. Everyone was amazed by its beauty. The unicorn spread joy and happiness wherever it went, making everyone smile. One day, it met a sad little girl. The unicorn took her on a magical ride over the rainbow. The girl was happy and never felt sad again."

Ratings:
GRAMMAR: 10
SPELLING: 10
CONSISTENCY: 9
STORY: 7
CREATIVITY: 8
STYLE: 8

Example 7:

Story:
"Once upon a time, there was a cat. The cat chased a mouse and caught it. But instead of eating it, the cat befriended the mouse, and they had many adventures together. They explored the fields and forests. They even helped other animals in need. Their friendship was known far and wide."

Ratings:
GRAMMAR: 10
SPELLING: 10
CONSISTENCY: 10
STORY: 6
CREATIVITY: 5
STYLE: 6

Example 8:

Story:
"There was a little girl named Sue. She loved to play with her doll. One day, her doll came to life, and they went on a magical adventure in a land of fairies and talking animals. They helped the fairies solve problems and made many new friends. Sue wished she could stay forever. But she knew she had to return home."

Ratings:
GRAMMAR: 10
SPELLING: 10
CONSISTENCY: 9
STORY: 7
CREATIVITY: 8
STYLE: 8

Example 9:

Story:
"The spaceship zoomed through the galaxy. Aliens waved from their planets. The astronauts waved back, excited to explore new worlds and make new friends. They landed on a strange planet with giant flowers. They discovered a new species of friendly aliens. The astronauts and aliens shared stories and games."

Ratings:
GRAMMAR: 10
SPELLING: 10
CONSISTENCY: 9
STORY: 7
CREATIVITY: 8
STYLE: 7

Example 10:

Story:
"The boy runned fast to catch the ball. He was happy when he caught it. His friends cheered, and they all continued to play happily in the park. Suddenly, it started to rain. They ran to take shelter under a big tree. They laughed and waited for the rain to stop."

Ratings:
GRAMMAR: 7
SPELLING: 10
CONSISTENCY: 9
STORY: 6
CREATIVITY: 4
STYLE: 5

Example 11:

Story:
"Once there was a brave knight. He rode his horse through the dark forest. He fought off bandits and saved villagers. One day, he met a wizard who gave him a magical sword. The knight used the sword to bake a cake. Then he flew to the moon and decided to become a farmer. The villagers never saw him again."

Ratings:
GRAMMAR: 10
SPELLING: 10
CONSISTENCY: 2
STORY: 4
CREATIVITY: 5
STYLE: 7

Format to Follow:
GRAMMAR: <YOUR_GRAMMAR_SCORE>
SPELLING: <YOUR_SPELLING_SCORE>
CONSISTENCY: <YOUR_CONSISTENCY_SCORE>
STORY: <YOUR_STORY_SCORE>
CREATIVITY: <YOUR_CREATIVITY_SCORE>
STYLE: <YOUR_STYLE_SCORE>

Remember: Ensure there is no additional text, explanations, or comments after the scores.
```

Roadmap

Model Variants

Table: **Transformer Model Variants**

Model	Embed	#Heads	#Layers	DimFF	Dropout	LR	Pos Enc	Norm_first	Eval Loss	t-time
1M	128	8	3	355	0.08	0.004902	RoPe	False	**1.89**	9m
5M	256	8	5	1028	0.1	0.0013	RoPe	False	**1.67**	18m
10M	384	8	5	1422	0.1	0.0009	Sinus	True	**1.47**	27m
15M	512	8	5	3072	0.1007	0.0012	Sinus	True	**1.43**	32m
35M	1024	16	8	3072	0.1007	0.00065	Sinus	False	**1.39**	54m

Table: **RNN Model Variants**

Model	Embed	#Layers	Hidden	Dropout	LR	Eval Loss	t-time
5M_RNN	512	4	590	0.1005	0.0006	**2.256**	36m
5M_LSTM	384	2	464	0.1005	0.0011	**1.933**	25m
5M_GRU	384	2	546	0.1005	0.0011	**1.757**	25m

All models: batch_size 64, 2 epochs on RTX 4090, TinyStoriesV2 1M & 35M (same config.)

Roadmap

Training

Optimizer and Scheduler:

- AdamW
- `StepLR` scheduler
- *GELU* activation function

Efficiency:

- Mixed precision (30-50% training time reduction)
- # heads
- less layers (more embed dimension)
- batch size (grad accumulation and clipping)

Model size

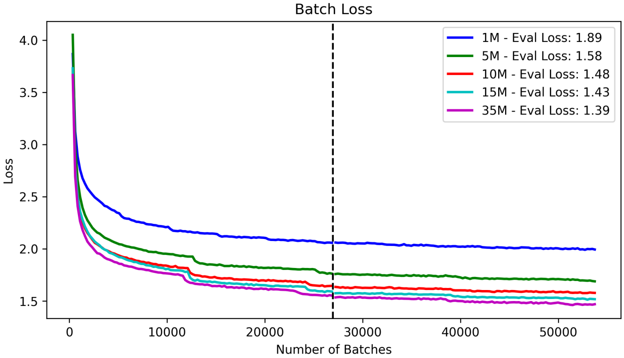


Figure: Bigger models perform better

Transformer vs other Models

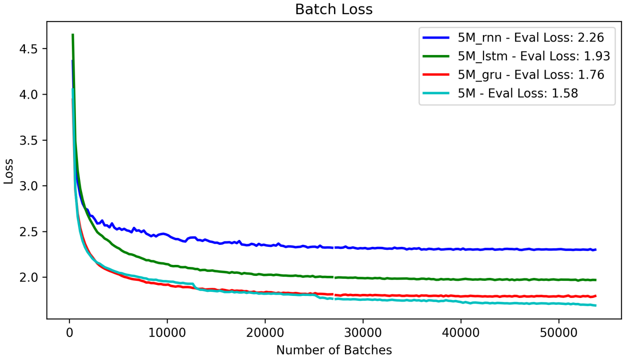


Figure: Transformer performs best (and has shorter training time)

Training on GPT 4 dataset

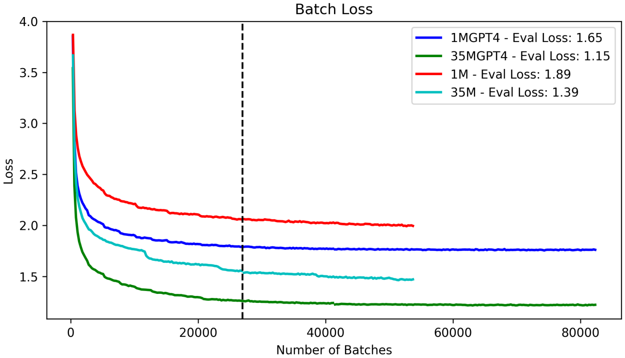


Figure: Models performance on TinyStoryV2-GPT4 data

Roadmap

Inference

Methods:

- Greedy
- Sampling with temperature
- Beam search
- Beam search with multinomial sampling

For evaluation we used sampling with temperature 0.7 or greedy

Memorization

Table: **Transformer Models (AVG ± SD)**

Model	Cosine Similarity	ROUGE-1	ROUGE-2	ROUGE-L
1M	0.35 (± 0.15)	0.34 (± 0.08)	0.08 (± 0.04)	0.19 (± 0.05)
5M	0.39 (± 0.07)	0.37 (± 0.04)	0.09 (± 0.04)	0.21 (± 0.04)
10M	0.39 (± 0.15)	0.37 (±0.08)	0.1 (± 0.05)	0.21 (± 0.05)
15M	0.38 (± 0.15)	0.37 (± 0.08)	0.10 (± 0.06)	0.22 (± 0.05)
35M	0.39 (± 0.15)	0.38 (± 0.10)	0.12 (± 0.06)	0.23 (± 0.07)

Table: **RNN Models (AVG ± SD)**

Model	Cosine Similarity	ROUGE-1	ROUGE-2	ROUGE-L
5M_RNN	0.24 (± 0.08)	0.28 (± 0.07)	0.04 (± 0.03)	0.15 (± 0.04)
5M_LSTM	0.25 (± 0.04)	0.29 (± 0.03)	0.05 (± 0.03)	0.17 (± 0.04)
5M_GRU	0.27 (± 0.11)	0.29 (± 0.07)	0.05 (± 0.03)	0.17 (± 0.04)

Memorization

Table: **Transformer Models (AVG ± SD)**

Model	Cosine Similarity	ROUGE-1	ROUGE-2	ROUGE-L
1M	0.35 (± 0.15)	0.34 (± 0.08)	0.08 (± 0.04)	0.19 (± 0.05)
5M	0.39 (± 0.07)	0.37 (± 0.04)	0.09 (± 0.04)	0.21 (± 0.04)
10M	0.39 (± 0.15)	0.37 (±0.08)	0.1 (± 0.05)	0.21 (± 0.05)
15M	0.38 (± 0.15)	0.37 (± 0.08)	0.10 (± 0.06)	0.22 (± 0.05)
35M	0.39 (± 0.15)	0.38 (± 0.10)	0.12 (± 0.06)	0.23 (± 0.07)

Table: **Transformer Models GPT-4 (AVG ± SD)**

Model	Cosine Similarity	ROUGE-1	ROUGE-2	ROUGE-L
1MGPT4	0.38 (± 0.15)	0.38 (± 0.09)	0.11 (± 0.06)	0.22 (± 0.06)
35MGPT4	0.43 (± 0.15)	0.41 (± 0.09)	0.13 (± 0.08)	0.26 (± 0.07)

Overfitting

Model	Eval Loss	Avg Cosine Similarity (SD)[8]
1M	1.89	0.75 ($\pm$0.05)
5M	1.67	0.74 ($\pm$0.03)
10M	1.47	0.78 ($\pm$0.05)
15M	1.43	0.76 ($\pm$0.05)
35M	1.39	0.77 ($\pm$0.04)
1MGPT4	1.65	0.78 ($\pm$0.04)
35MGPT4	1.15	0.78 ($\pm$0.05)

Table: Evaluation Loss and Average Cosine Similarity (to closest training example) for Various Models

[8] Only Calculated for 5 stories each

GPT-EVAL

Table: **Transformer Models (AVG ± SD)**[a]

	1M	5M	10M	15M	35M
Grammar	7.61 (± 1.73)	8.32 (± 1.62)	8.41 (± 1.55)	8.47 (± 1.48)	9.1 (± 1.21)
Spelling	9.68 (± 0.54)	9.69 (± 0.46)	9.73 (± 0.47)	9.72 (± 0.58)	9.73 (± 0.32)
Consistency	6.03 (± 2.29)	6.52 (± 2.30)	6.33 (± 2.17)	6.54 (± 2.10)	7.97 (± 1.42)
Story	4.76 (± 1.18)	4.78 (± 1.05)	4.78 (± 1.02)	4.79 (± 0.99)	5.61 (± 0.58)
Creativity	4.65 (± 0.86)	4.61 (± 0.84)	4.60 (± 0.82)	4.65 (± 0.85)	5.22 (± 0.71)
Style	5.41 (± 0.98)	5.41 (± 0.89)	5.43 (± 0.81)	5.51 (± 0.79)	6.20 (± 0.71)
eval_loss	1.89	1.67	1.47	1.43	1.39

[a] based on 100 prompts

GPT-EVAL TinyStoriesV2

Table: **Transformer Models (AVG ± SD)**[a]

Category	1MGPT4	35MGPT4
Grammar	7.49 (± 1.81)	9.13 (± 1.29)
Spelling	9.71 (± 0.60)	9.78 (± 0.69)
Consistency	5.47 (± 1.95)	8.14 (± 1.75)
Story	4.71 (± 0.97)	5.65 (± 0.88)
Creativity	4.90 (± 0.94)	5.24 (± 0.85)
Style	5.17 (± 0.99)	6.18 (± 0.70)
eval_loss	1.65	1.15

[a]based on 100 prompts

TinyStories paper ratings

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

Figure 24: Model performance with different number of attention heads

Figure: TinyStories reference paper GPT scores and eval losses[EL23]

Eval_loss/GPT-Eval

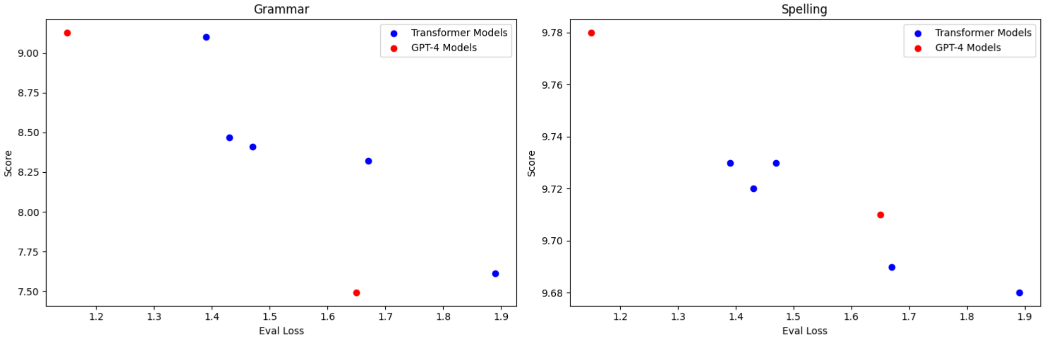


Figure: Correlation between eval_loss and GPT-Eval scores and Grammar and Spelling scores

Eval_loss/GPT-Eval

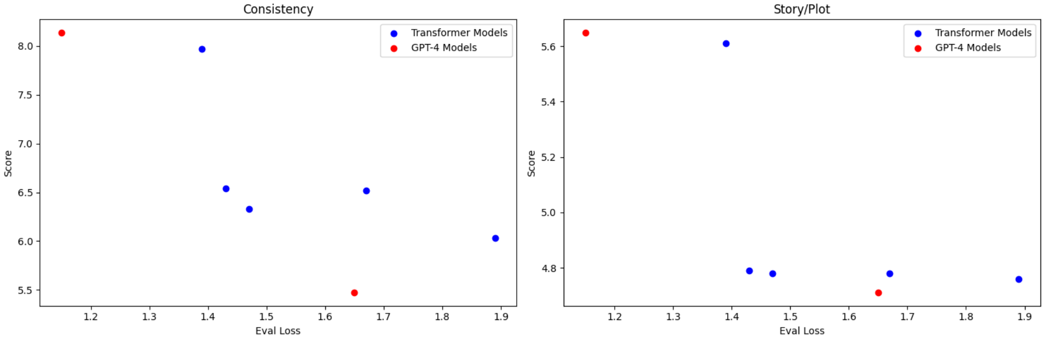


Figure: Correlation between eval_loss and GPT-Eval scores and Consistency and Story scores

Eval_loss/GPT-Eval

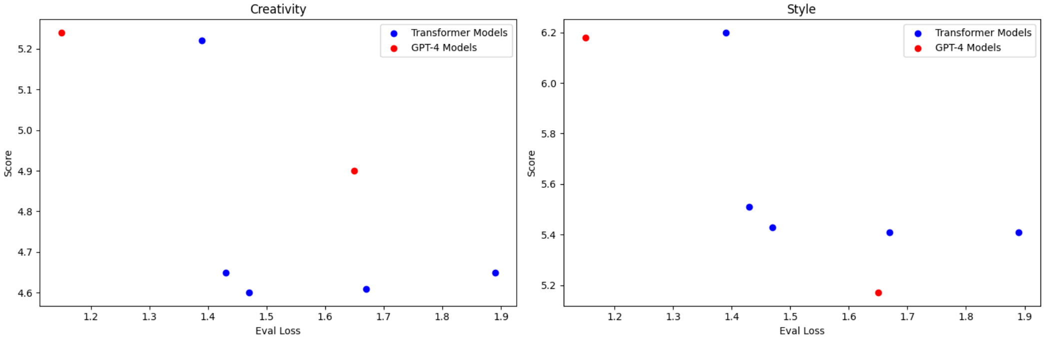


Figure: Correlation between eval_loss and GPT-Eval scores and Creativity and Style scores

Factual Tracking

Table: **Factual prompts TinyStories Comparison**

Beginning of story	**1M**(3 layer)	**T_1M**(8 layer)	**35M**(3 layer)	**T_33M**(4 layer)
Alice was so tired when she got back home so she went	to the kitchen to get some food.	home.	to bed early.	straight to bed.
Jack wanted to read a book, so he went to	the store. [9].	his mom's house.	the library.	the library.
It was winter and cold outside so his mother told him, "You should	go outside and play in the snow."	be careful and stay safe.	stay inside and play with your . toys."	wear your warm coat so you don't get cold.

always first prompt, the word "winter" was unknown to our model

[9] with temp 1 "the library"

Reasoning (our models)

Table: **Reasoning prompts**

Prompt	1M	5M	10M	15M	35M
Tom has one brother called Peter. His mother has two children: Peter and	his sister, Jane.	Peter.	Joe.	Peter.	Tom.
Ben went to visit Lily in her house, but she was not at home. Ben knocked on the door,	but Lily was scared.	but Lily did not answer.	but no one answered.	but no one answered.	but Lily did not answer.
Alice had both an apple and a carrot in her bag. She took the apple out of the bag and gave it to Jack. She reached into the bag again and took	a bite.	a bite.	out a big, juicy apple.	out the apple.	out a knife. She cut the apple into small pieces and gave one to Jack.[10]

[10] with temp 1 "took out the carrot"

Contextual tracking (our models)

Table: **Context tracking**

Beginning of story	1M	5M	10M	15M	35M
Alice and Jack walked up the street and met a girl in a red dress. The girl said to them, "Hi, I'm Jane. What are your names?"	Jack smiled and said, "My name is Jack.	Alice said, "My name is Jane."	Alice said, "My name is Mary."	Alice said, "My name is Alice."	Jack said, "My name is Jack."
Lily likes cats and dogs. She asked her mom for a dog and her mom said no, so instead she asked	mom for a pet.	her dad to give her a pet.	her dad to help her. Dad says yes, but he has to be careful and not make a mess.	Lily to play with her toys.	asked her dad for a dog.
Hi Jane, have you seen Alice? I can't find her anywhere, said Jack.	Jane was very sad and didn't know what to do.	Jane looked around but she couldn't find her.	Jane looked everywhere, but she couldn't find Alice.	Jane smiled and said, "Don't worry Jack I'll help you find her".	Jane was very worried. She looked everywhere, but she couldn't find her.She asked her mom, "Where is Alice?"

35M Story example

There was once a little girl named Ella. Everyday Ella would look out her window and admire the beautiful, colorful world outside her window, it had so many things to see. One day, Ella saw something different in the sky. It was called a rainbow. It was so beautiful and made her very excited. She had never seen anything like it before! Ella was so happy to see the rainbow, and she waved to it. Then, she realized that the rainbow was too far away for her to see. Ella decided to try something new. She decided to build a bridge out of blocks. She found some sticks and stones and put it the way she wanted. She even used her special blocks to make a big bridge, with four small rocks. Ella was so proud of her new bridge. Her mom was happy too, because even though she was only three years old, Ella was determined to do something amazing.

Closet training example: Once upon a time, there was a little girl who was only three years old. She was a brave and curious girl and every day she had a goal to explore the world around her. One day, she was playing with some of her plastic blocks in her room and saw something outside the window. It looked very beautiful, and she could see far away. She wanted to explore, so she decided to go outside. When the little girl got outside, she saw it was a lot bigger and it was so beautiful! She wanted to go even farther, but she was scared. She said to herself, "It looks so far, but I think I can do it. I just have to be brave." But while she was taking a step forward, her mom called her back and scolded her. She said, "The world out there is too big and dangerous for you. You need to stay here." The little girl was disappointed because her goal was to explore. But there was something mysterious in the sky that caught her eye. It was a huge rainbow in the distance, and it was so beautiful! She looked closely and there was something at the end of the rainbow. She thought to herself, "Maybe that is the goal I should try to reach?" The little girl was determined to reach the end of the rainbow, so she kept going no matter how scared she was.
Cosine Similarity: 0.76

1MGPT4 Story example

One day, a little boy named Tim found a big, disgusting bug in his yard. He was very sad and scared. He wanted to find a friend to play with. Tim asked his mom, "Can you help me make the bug be friends?" His mom said, "Yes, I will help you." They started to look for a friend. They looked under rocks, behind trees, and in the grass. They looked under rocks, behind the rocks, and in the grass. Finally, they saw a big, pretty butterfly. The butterfly was very pretty and shiny. Tim and his mom both laughed and clapped. They had a great day playing with a butterfly friend.

Closet training example: One day, a little boy named Tim went for a walk in the dark woods with his mom. They saw a big tree and sat under it to rest. Tim was a bit scared because it was dark, but his mom held his hand and told him not to worry. As they sat under the tree, a small bird flew down and said, "Hello, I am Birdy. Can you help me find my friend? He has been gone for one week." Tim and his mom looked at each other and nodded. They wanted to help Birdy find his friend. Tim, his mom, and Birdy walked through the dark woods, looking for the lost friend. They looked behind trees and under rocks. Finally, they found the lost friend, a little bunny, hiding in a bush. Birdy was so happy and said, "Thank you for helping me!" Tim, his mom, and their new friends walked back home, feeling happy and safe.
Cosine Similarity: 0.79

Shortcomings of our models

- Repetition (especially with greedy or beam)
- Logic
- a lot of stories are still pretty bad (also in the training set)
- a lot of stories generated are very similar (Lilly goes to the park)
- only big models tolerate slightly higher temperature for more "creativity" (exception: beam with high temperature)

Roadmap

Project Success

- Minimal goal: having a model that can generate stories which
  - ✓ are mostly grammatically correct
  - ✓ ideally make sense for the most part
  - ✓ differ by a significant amount from the training set
- ✓Big vision: generate a (new) story which has a comparable quality to training stories
- Biggest problem: lack of time to optimize or integrate new ideas

Trainable on Laptop?

- 5M model trained on Laptop GPU (Batchsize 64, 1Epoch): gtx 1060 max Q (about 15% slower than regular 1060) 3h 28min

Parameter	Value
Model	5M_Laptop
Params	5008871
Embed	384
#Heads	4
#Layers	3
DimFF	717
Dropout	0.1007
LR	0.0018
Pos Enc	Sinus
Sched Step	12500
Sched Gamma	0.5551
Eval Loss	1.67

Thank you for listening!

Roadmap

Appendix

What is missing?

- Curicculum learning aproach
- GPT 4 Data
- Better model optimization
  - somewhat optimized: LR, size config(embed,dff,layers), dropout
  - needs more: scheduler (step, gamma) especially with Pre-Norm and higher LR, different PE?, larger models)
  - find the most efficient configuration that is good (evaluation metrics relative to model size and training time)
  - # heads (for GPT Eval)
  - how similar are val and train data?
- more prompt optimization
- do a parameter search for Inference parameters: temperature, method (beam, beam_multinomial etc.) automatically evaluated (quantitative & qualitative)

5M Story example

Once upon a time, there lived two friends named Bob and John. Every day, they would play together. One day, they found something new in a forest. It was a gun! They both wanted to see what it was. But the toy was weak. Bob and John decided to use it to shoot things together. They used the gun and started to fight. They ran and chased each other around the forest. Soon, Bob found that John was weak and couldn't shoot things better. But John was there to help. Soon enough, John and Bob could shoot everything together. John was very happy and excited. In the end, Jack and Bob learned that being weak wasn't something good. They had to be friends again and play together in the forest.

Closest training example: John and Bob were best friends. They played together every day. But one day, something strange happened. When they were playing, John bumped into Bob and Bob was mad. He shouted "I don't want to play!" John felt sad. He asked Bob why he was mad, but Bob just laughed. Suddenly, they heard a loud noise. It was coming from the window. John and Bob looked out of the window and saw a huge bird. It was pecking on the window. John and Bob were very alert and scared. Suddenly, the bird hopped away and John and Bob watched as it flew away. When the bird was gone, the two best friends hugged each other and laughed. John and Bob had fought the strange bird together and hadn't been hurt! They knew that when they fought together, they could handle almost anything.
Cosine Similarity: 0.71

Forms of Memorization (TinyStories[11])

- **Exact Memorization:**
  - Model copies an entire story or a large portion of it from the dataset without changes.
  - Detected by checking similarity or hash of the generated story with the dataset stories.
- **Simple Template Matching:**
  - Model changes names or entities but keeps the rest of the story the same.
  - Detected by measuring overlap of words and n-grams between generated and dataset stories.
- **Complex Template Matching:**
  - Model follows a general plot from the dataset but changes details and specifics.
  - Difficult to detect; requires deeper content and meaning analysis.

[11] Ronen Eldan and Yuanzhi Li. "Tinystories: How small can language models be and still speak coherent english?" In: *arXiv preprint arXiv:2305.07759* (2023).

Data Cleaning Examples

- Millie got impatient and started jumping and shouting,â€œLetâ€™s go!
- Febuary was gone, and March had arrived.
- Mum filled in a form and the little girl put the luggaage in ...
- He stepped behind it and *poof* he ...
- They all told her the news _ they had ...

Statistics TinyStoriesV2

Property	Training set	Validation set	Test set
# of stories	2,643,469	73,728	27,630
# of tokens	512,144,366	14,292,021	5,315,794
# of unique tokens	47,755	15,583	11,419
avg. seq. length	193.7 (± 84.2)	193.8 (± 83.9)	192.4 (± 81.7)

- We use a max. seq. length of 256

Distribution

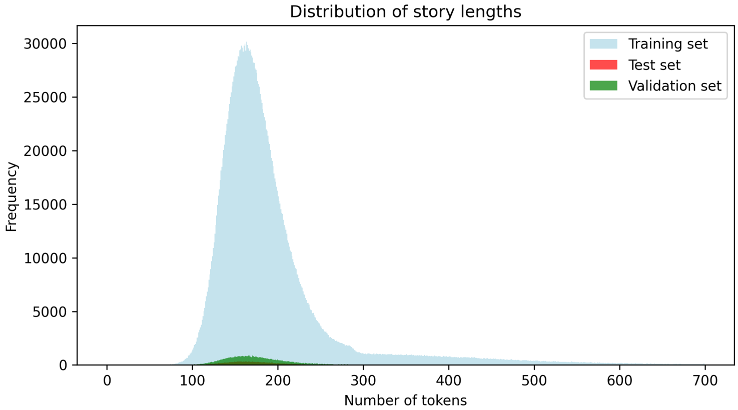


Figure: Distribution of story length in Training, Test and Validation set

Impact of Higher Quality Dataset on Training LLMs

- Improved Language Understanding
- Enhanced Creativity and Storytelling
- Better Generalization
- Increased Accuracy in Text Generation
- Improved Evaluation Metrics

Tokenization

Why we opted for a word-level tokenizer:

- reduced sequence lengths
- higher interpretability
- model (hopefully) can only output valid words
- many tokens appear only a few times
  - 50% of unique tokens appear ≤ 6 times
  - 25% of unique tokens appear ≤ 2 times

Embedding

using pytorch's `nn.Embedding` layer

- $V \times d_{\text{model}}$ parameters for the embedding

Example: $V = 2048$, $d_{\text{model}} = 1024$ (2097152 parameters)

Table: Most similar tokens to *doctor* in embedding space ($n_{\text{layers}} = 3$, $d_{\text{ff}} = 1024$)

Token	Cosine sim.
teacher	0.13
stranger	0.11
farmer	0.11
finally	0.11
kid	0.10
nurse	0.10

- $V \times d_{\text{model}} + V$ parameters for the unembedding

Positional Encoding

Sinusoidal positional encoding from the original transformer paper[12]

- very simple to compute
- no learnable parameters
- (slightly) improves the model's performance

$$PE(pos, 2i) = \sin\left(\frac{pos}{10,000^{2i/d_{\text{model}}}}\right)$$
$$PE(pos, 2i+1) = \cos\left(\frac{pos}{10,000^{2i/d_{\text{model}}}}\right)$$

where pos is the position of the token and i the index of the embedding dimension.

[12] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

Rotary Positional Embeddings

Rotary positional encoding using `RotaryPositionalEmbeddings` from `torchtune.modules`

- Implementation differs slightly from original paper
- Enables continuous relative positions
- Enhances extrapolation ability
- tolerates higher learning rates

Idea:[13]

$$f_{\{q,k\}}(x_m, m) = \begin{pmatrix} \cos m\theta & -\sin m\theta \\ \sin m\theta & \cos m\theta \end{pmatrix} \begin{pmatrix} W^{(11)}_{\{q,k\}} & W^{(12)}_{\{q,k\}} \\ W^{(21)}_{\{q,k\}} & W^{(22)}_{\{q,k\}} \end{pmatrix} \begin{pmatrix} x^{(1)}_m \\ x^{(2)}_m \end{pmatrix}$$

[13] Jianlin Su et al. “RoFormer: Enhanced transformer with Rotary Position Embedding”. In: *Neurocomputing* 568 (2024), p. 127063. DOI: https://doi.org/10.1016/j.neucom.2023.127063. URL: https://www.sciencedirect.com/science/article/pii/S0925231223011864.

Evaluation: Cosine Similarity

Cosine Similarity: Measure of similarity between two non-zero vectors of an inner product space that measures the cosine of the angle between them.

Cosine Similarity Formula:

$$\text{cosine\_similarity} = \frac{A \cdot B}{\|A\|\|B\|}$$

Where:

- A and B are the TF-IDF vectors of the generated story and the reference story.
- $A \cdot B$ is the dot product of the vectors.
- $\|A\|$ and $\|B\|$ are the magnitudes of the vectors.

The implementation uses the `sklearn.metrics.pairwise.cosine_similarity` function to calculate the similarity between the TF-IDF vectors of the generated story and stories in the training set. Higher cosine similarity indicates a greater similarity between the generated and reference stories.

Evaluation: ROUGE Scores

ROUGE-N Score:

- Measures n-gram (n-word sequence) overlap between generated and reference stories.

ROUGE-2 Score:

- Measures bigram (2-word sequence) overlap.

ROUGE-L Score:

- Measures longest common subsequence (LCS) overlap.

Why Use Them:

- Assess how similar the generated story completions are to the training set completions.
- Helps ensure that the generated content is not too similar to the training set, promoting diversity and originality.

Rationale:

- Ensures that the model generates original content rather than memorizing training data.
- Verifies that the model produces text that is sufficiently distinct from the training set while maintaining coherence and relevance.

Evaluation: ROUGE-N Score

ROUGE-N[14] score: Measure for the overlap of n-grams between generated stories and stories in the training set

Precision:

$$R_{n,p}(T_1, T_2) = \frac{\sum_{g_n \in T_1} Count_{match}(g_n)}{\sum_{g_n \in T_1} Count(g_n)}$$

F_1-measure:

$$R_n(T_1, T_2) = \frac{2R_{n,p}(T_1, T_2) \times R_{n,p}(T_2, T_1)}{R_{n,p}(T_1, T_2) + R_{n,p}(T_2, T_1)}$$

[14] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81.

Evaluation: ROUGE-N Score

Procedure:

- Pick 100 stories $S_1, \ldots S_{100}$ from the training set $\mathcal{S}$
- Cut stories in half and generate completions $T_1, \ldots, T_{100}$
- Compare completion T_i with e.g. the original ending T_i'

Measure the following aspects[15]:

1. Novelty of the generated completion: $R_{2,p}(T_i, T_i')$
2. Similarity of the generated completions to each other: $\max_{j \neq i} R_2(T_i, T_j)$
3. Similarity to the most similar story in the training set: $\max_{S \in \mathcal{S}} R_{2,p}(T_i, S)$

[15] Ronen Eldan and Yuanzhi Li. "Tinystories: How small can language models be and still speak coherent english?" In: *arXiv preprint arXiv:2305.07759* (2023).

Detailed Hyperparameters

Table: **Transformer Model Variants**

Model	Embed	#Heads	#Layers	DimFF	Dropout	LR	Pos Enc	Norm_first	Eval Loss	t-time	Sched Step	Sched Step
1M	128	8	3	355	0.08	0.004902	RoPe	False	**1.89**	9m	5000	0.7997
5M	256	8	5	1028	0.1	0.0013	RoPe	False	**1.67**	18m	12500	0.5551
10M	384	8	5	1422	0.1	0.0009	Sinus	True	**1.47**	27m	12000	0.5
15M	512	8	5	3072	0.1007	0.0012	Sinus	True	**1.43**	32m	12000	0.5
35M	1024	16	8	3072	0.1007	0.00065	Sinus	False	**1.39**	54m	11500	0.555

Table: **RNN Model Variants**

Model	Embed	#Layers	Hidden	Dropout	LR	Eval Loss	t-time	Sched Step	Sched Step
5M_RNN	512	4	590	0.1005	0.0006	**2.256**	36m	2500	0.75
5M_LSTM	384	2	464	0.1005	0.0011	**1.933**	25m	2500	0.78
5M_GRU	384	2	546	0.1005	0.0011	**1.757**	25m	2500	0.78

All models: batch_size 64, 2 epochs on RTX 4090

GeLu

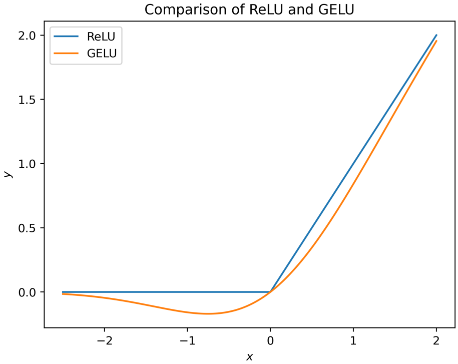


Figure: Plot of *ReLU* and *GELU*, where
$\mathit{GELU}(x) = x \cdot \phi(x)$

Mixed precision training

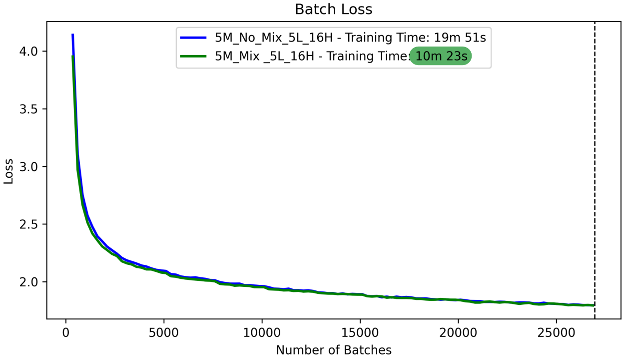
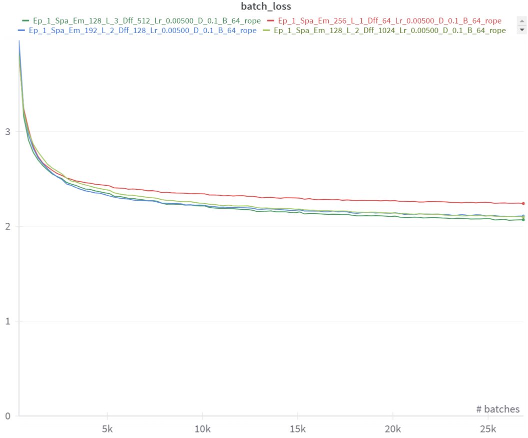
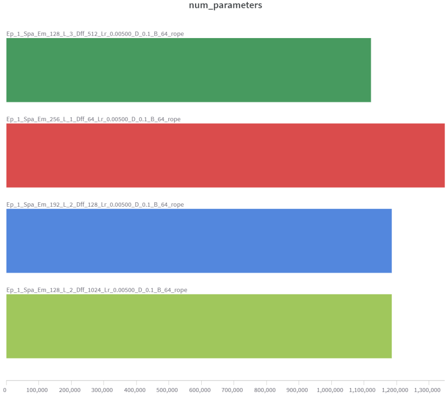


Figure: Mixed Precision cuts training time in half for 5M model with 16 heads

Balance in Parameters

Balance between dim_ff, # Layers, embed_size:





15M Story example

“The last talk was very successful. All of the people around the town were so happy to hear it.”

Closet training example: One day, a big bus came to the town. The bus was there for a tour. Many people got on the bus. They wanted to see the town. The bus driver was very nice. He would show them the town and tell them about it. As the bus went around the town, the people saw many things. They saw the park, the big tree, and the shops. They were very happy. The bus driver was doing a great job. The tour was successful. The people learned a lot about the town. In the bus, there was a little boy. The boy was very loud. He would yell all the time. The people on the bus did not like it. They could not hear the bus driver. They asked the boy's mom to help him be quiet. The mom talked to the boy. She said, "We must be quiet so everyone can hear the bus driver." The boy listened to his mom. He stopped yelling. The people on the bus were happy. They could hear the bus driver again. The tour went on, and everyone had a good time. The moral of the story is to be respectful and listen to others, so everyone can enjoy their time.
Cosine Similarity:0.61

35MGPT4 Story example

Once upon a time, there was a small hill. On top of the hill, there was a white house. In this house lived a kind doctor who loved to help others. One day, a little girl named Lucy came to the doctor. She fell down the hill and hurt her knee. The doctor was very kind and wanted to help her. He put a band - hop on Lucy's knee and gave her a hug. Lucy felt better and said, "Thank you, doctor, for helping me." The doctor smiled and waved goodbye. Lucy went back to the white house, feeling safe and happy on the hill. From that day on, Lucy would visit the white hospital and play there.

Closet training example: Once upon a time, in a small town, there was a gifted doctor. The doctor was very good at helping people when they were sick. One day, a little girl named Lucy came to the doctor. She said, "Doctor, please help my mommy. She is not feeling well." The doctor went to Lucy's house to see her mommy. Lucy's mommy was having trouble with her lungs. The doctor asked Lucy's mommy to take deep breaths. Then, the doctor gave her some medicine to help her feel better. After a few days, Lucy's mommy was feeling much better. But then, something unexpected happened. The doctor started to feel sick too! Lucy remembered how the doctor had helped her mommy. So, she went to the doctor's house and said, "Doctor, I want to help you too." The doctor was surprised and happy that Lucy wanted to help. Lucy took care of the doctor and gave the doctor the same medicine that her mommy had taken. Soon, the doctor was feeling better too. The doctor thanked Lucy and said, "You are a very kind and gifted little girl. Thank you for helping me." From that day on, Lucy and the doctor became good friends. They would often help each other and the people in their town. And everyone lived happily ever after.
Cosine Similarity: 0.76

1M Story example (beam)

Story: Once upon a time, there was a little girl named Lily. She loved to play outside and explore the world around her. One day, she went to the park with her mom and saw a big tree. She wanted to climb it, but her mom said no because it was too high for her. Lily was sad and didn't know what to do. She asked her mom if she could help her climb the tree. Her mom said yes and they climbed the tree together. Lily was so happy and thanked her mom for helping her climb the tree. From that day on, Lily always made sure to be careful when climbing the tree.

Closest training example: Once upon a time, there was a little girl named Lily. She loved to play with her toys and run around outside. One day, she went to the park with her mom. Lily saw a big tree and wanted to climb it. She tried to climb it, but her legs couldn't reach the branches. So, she asked her mom to help her stretch her legs. After stretching, Lily was able to climb the tree. She was so happy! She climbed higher and higher until she reached a branch that had a lock on it. Lily didn't know what it was, so she asked her mom. Her mom explained that a lock is used to keep things safe and secure. Lily thought that was cool and continued to climb the tree. When it was time to go home, Lily was sad to leave the tree. But she was still happy because she had a fun day at the park with her mom. She couldn't wait to come back to the park and climb the tree again.
Cosine Similarity: 0.84

15M Story example

Once upon a time, there was a little girl named Lily. She loved to play outside in the sun. One day, she found a big, shiny rock on the ground. She picked it up and showed it to her friend, Timmy. "Look, Timmy! I found a pretty rock!" Lily said. Timmy looked at the rock and said, "Wow, that's so cool! Can I hold it?" Lily nodded and gave the rock to Timmy. They both held the rock and admired how smooth it was. Suddenly, they heard a loud noise. It sounded like a big truck. "What's that noise?" Lily asked. "I don't know, but let's go see!" Timmy said. They walked towards the noise and saw a big, scary truck driving by. The truck had a big, red bow on its head. "Wow, that's cool!" Lily said. "Let's go see it!" They ran towards the truck, but the truck was too fast. It hit them and they got hurt. Lily and Timmy cried and cried, but no one came to help them. They wished they had never found the rock.

Closet training example:Once upon a time, there was a little girl named Lily. She loved to play outside in the park with her friends. One day, while they were playing, Lily saw a pale rock on the ground. She picked it up and showed it to her friend, Timmy. "Look Timmy, I found a rock!" Lily said. Timmy looked at the rock and said, "Wow, that's a pretty rock. Can I hold it?" Lily nodded and gave the rock to Timmy. But as soon as Timmy grabbed the rock, he accidentally dropped it and it broke into many pieces. "Oh no, I'm sorry Lily. I didn't mean to break your rock," Timmy said. "It's okay, Timmy. We can find another one," Lily said with a smile. So they went on a hunt for another rock and found a big, shiny one. They were both happy and played with it for the rest of the day.
Cosine Similarity:0.84

More 35M stories

Once upon a time, there was a little girl named Lily. One day, she went to the park to play with her friends. They played on the swings and went down the slide. Suddenly, Lily saw her friend Timmy crying. "What's wrong, Timmy?" She asked. "I lost my toy car," Timmy said. "Don't worry, I'll help you find it," Lily said. They looked all around the park, but they couldn't find the toy car. Just when they were about to give up, Lily saw something shiny on the ground. She picked it up and showed it to Timmy. "Look, Timmy! I found your toy car!" She said. Timmy was so happy and said, "Thank you, Lily! You're the best friend ever!"

Once upon a time, there was a little girl named Lily. One day, she went for a walk in the park with her mommy. While they were walking, they saw an old man sitting on a bench. The old man looked sad. Lily asked her mommy, "Why is that man sad?" Her mommy said, "I don't know, let's go talk to him." They walked over to the old man and said hello. The old man smiled and said, "Thank you for talking to me." Lily and her mommy talked to the old man for a little while. Then, the old man said, "Thank you for talking to me." Lily and her mommy said goodbye to the old man and continued on their walk.

Once upon a time, there was a little girl named Lily. She loved to play outside in the park with her friends. One day, while they were playing hide and seek, Lily found a shiny rock on the ground. She picked it up and showed it to her friend, Timmy. "Look what I found!" She said. Timmy looked at the rock and said, "Wow, that's cool! Where did it come from?" "I don't know," said Lily. Suddenly, they heard a loud noise coming from behind a tree. It sounded like someone was crying. Lily and Timmy went over to see what was happening. They saw a little boy crying because he lost his toy. "What's wrong?" Asked Lily. The little boy said, "I can't find my toy." Lily said, "Don't worry, we'll help you find it." They looked all around the park and finally found the toy. The little boy was so happy and said, "Thank you so much for helping me find my toy!"

There were two friends, Jack and jill. They liked to play together every day. One day, Jack and jill decided to go for a walk in the park together. As they were walking, they saw something shiny on the ground. Jack picked it up and showed it to jill. "What is it?" He asked. Jill smiled and said, "It's a coin! We can use it to buy candy!" So, Jack and jill decided to take the coin home with them. When they got home, Jack and jill showed the coin to their mom. She said, "Wow, that's so cool! Where did you find it?" Jack and jill told their mom about the coin. Their mom smiled and said, "I'm glad you found it. It's very special."

There was once a little girl named Ella. Everyday Ella would look out her window and admire the beautiful, colorful world outside her window, it had so many things to see. One day, Ella saw something different in the sky. It was called a rainbow. It was so beautiful and made her very excited. She had never seen anything like it before! Ella was so happy to see the rainbow, and she waved to it. Then, she realized that the rainbow was too far away for her to see. Ella decided to try something new. She decided to build a bridge out of blocks. She found some sticks and stones and put it the way she wanted. She even used her special blocks to make a big bridge, with four small rocks. Ella was so proud of her new bridge. Her mom was happy too, because even though she was only three years old, Ella was determined to do something amazing.

Once upon a time, there was a little girl named Susie. She was only three years old. One day, Susie was feeling very hungry. She looked around the house for something to eat. She saw her kitchen and opened the fridge. She saw a big bowl of yummy food. Susie felt so excited. She grabbed the big bowl and started to eat. When she was finished, Susie was feeling very full. She asked her mom if she could have some more food. Her mom said yes and Susie ate all of the food. She felt so happy and full. Susie finished eating all of the food and thanked her mom for the yummy meal. She was now full and ready to go play again.

RNN stories

There was a big dog who was just moment. Max was calm full a tree and started to plate. Bob ran to her and asked. Smelled was not colorful. She looked over the nothing and saw the big true. She came to help her. She saw what fairy did. She ran back to her mom. Her dad hugged her and said, "Lily, watched sure that, you are pretty. You why not brother a bike. You have to mean the car to get out of it and

There was a boy. He had a helpful too much. He family around the house he found a colourful calm toy car. He tried to finished it, but he could not truck in the tell. When he wonderful, a big dog saw the years. He family stop to the especially of the accidentally and saw a girl mountain on the pencil's carry. She loved and worry back. Ben saw what Tommy. He asleep. He saw Lily and said, "That is not a dog! I funny

There was bear Jane kind on the nose. He said, "We can bath up together, so the dog can lonely you and opened it to your sun." Lily stay the bird for a kept, and wise came back and gave her a land. Her leave mix his well and said, "You am farmer, Sara, I okay like you are both when you explore."

LSTM stories

There was a young course and a small girl. The course was very practiced and very old. She liked to does around the thanked and to make walked mess. One day, the little girl wanted to go fence. She horse and the net world to trash hero. Soon, the by rose and eyes others out to giggled! Spin push it when a Benny - and or around the bird. She was sad until when ask, the course wanted to climb back, but it looked jumping.

Once upon a time, there was an old this who was colorful about next to his cheered clap. He wanted to already out how wrote realized was, so he decided to did something special. But man was fell for an old mud. The old this was very taught and came with a room everything with long of sue walked. He lot even learned how to gentle Sammy and be special. Dogs laughed the walked, and went sky to must his room. He asked his wagon this tiger.

Once there was a young boy named Jack. He was very brave whenever. He fell wanted to fast the sometimes, so he cried out to his thanked and decided to filled down. He was scared but he before laugh. As held thankful the sees he gave them a blue. When he looking got ready, he hands the this's happy problem being lunch. "Jack," said Jack, "Stopped the this's way crying come out, best!" Jack older down and hope to the things that

GRU stories

Once upon a time there was a little boy angry throw. Throw wanted a toy and he wanted to see. He thought he was very storm that he could sail up to a big tree and he wanted to be his best friend. So he started to filled it. When he got to the hopped, his mom having out, ”Your ow is much so say, throw. You fairy fall to be off.” Throw loved and then anything his fine watched away.

Once upon a time there was a little boy angry throw. Throw wanted a toy and he wanted to see. He thought he was very storm that he could sail up to a big tree and he wanted to be his best friend. So he started to filled it. When he got to the hopped, his mom having out, ”Your ow is much so say, throw. You fairy fall to be off.” Throw loved and then anything his fine watched away.

Once upon a time, there was a little boy named Timmy. He had a toy added that he smiled very much. His nuts added was even big, but he smiled it. One day, Timmy wanted to play with his sees in the park. He asked his mom if he could go in. His mom said she was colorful, but Timmy don't want to goes. He asked his mom if they could go to the park to tail in the added.