AutomaTikZ

UNIVERSITÄT BIELEFELD

Text-Guided Synthesis of Scientific Vector Graphics with TikZ

Jonas Belouadi and **Anne Lauscher** and **Steffen Eger**
Natural Language Learning Group (NLLG)

NLLG

Scientific Figures

Scientists use scientific figures to convey complex ideas or present critical findings, making them central to scientific research. There are various characteristics which we have come to expect:

Source: https://www.britannica.com/biography/Pythagoras

Scientific Figures

Scientists use scientific figures to convey complex ideas or present critical findings, making them central to scientific research. There are various characteristics which we have come to expect:

- high degree of geometric precision

Source: https://www.britannica.com/biography/Pythagoras

Scientific Figures

Scientists use scientific figures to convey complex ideas or present critical findings, making them central to scientific research. There are various characteristics which we have come to expect:

- high degree of geometric precision
- legibility even at small font sizes

Source: https://www.britannica.com/biography/Pythagoras

Scientific Figures

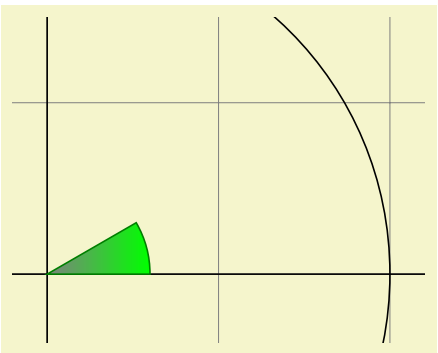
Scientists use scientific figures to convey complex ideas or present critical findings, making them central to scientific research. There are various characteristics which we have come to expect:

- high degree of geometric precision
- legibility even at small font sizes
- searchable text

Source: https://www.britannica.com/biography/Pythagoras

Research Idea

Existing text-to-image approaches fall short of these properties. To address this, we propose the use of TikZ, a well-known graphics language tailored to the creation of scientific figures.



```
\begin{tikzpicture}[scale=3]
  \clip (-0.1,-0.2) rectangle (1.1,0.75);
  \draw[step=.5cm,gray,very thin] (-1.4,-1.4) grid (1.4,1.4);
  \draw (-1.5,0) -- (1.5,0);
  \draw (0,-1.5) -- (0,1.5);
  \draw (0,0) circle (1cm);
  \shadedraw[left color=gray,right color=green, draw=green!50!black]
    (0,0) -- (3mm,0mm) arc (0:30:3mm) -- cycle;
\end{tikzpicture}
```

Source: https://tikz.dev

Basic Approach

We collect a dataset of 120k text-Ti*k*Z pairs. As Ti*k*Z is implemented with TEX macros, a foundational language model capable of understanding it would be desirable.

Source: Dall-E 3

Basic Approach

We collect a dataset of 120k text-TikZ pairs. As TikZ is implemented with TeX macros, a foundational language model capable of understanding it would be desirable.

- surprisingly models that see TeX during pre-training are pretty rare

Source: Dall-E 3

Basic Approach

We collect a dataset of 120k text-TikZ pairs. As TikZ is implemented with TeX macros, a foundational language model capable of understanding it would be desirable.

- surprisingly models that see TeX during pre-training are pretty rare
- however, the recent LLaMA model was trained on arXiv and TeX StackExchange, so it qualifies.

Source: Dall-E 3

Basic Approach

We collect a dataset of 120k text-TikZ pairs. As TikZ is implemented with TeX macros, a foundational language model capable of understanding it would be desirable.

- surprisingly models that see TeX during pre-training are pretty rare
- however, the recent LLaMA model was trained on arXiv and TeX StackExchange, so it qualifies.

Source: Dall-E 3

Idea

Fine-tune LLaMA on our dataset and see how it performs.

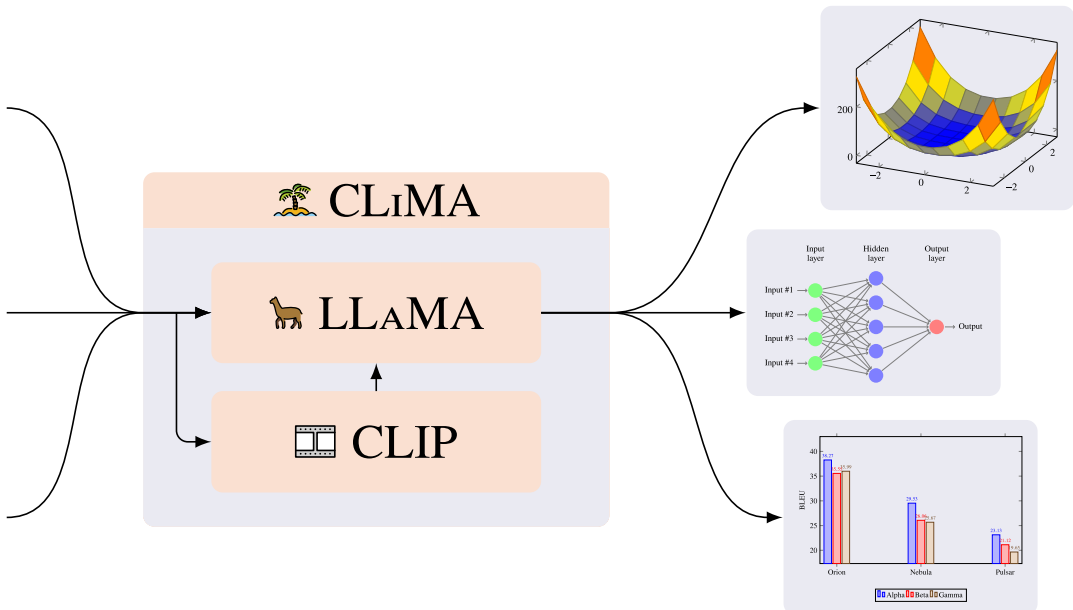
LLaMA learns to see

We hypothesize that incorporating a vision encoder like CLIP could boost our model. We thus incorporate CLIP using its multi-modal projection as a soft prompt, which should allow us to extract visual information from the input text (and optionally images). We call this model CLiMA (CLIP inside LLaMA).

3D contour plot of a loss function, showcasing global and local minima. The color gradient indicates function depth, providing insight into the optimization challenges in machine learning.

Visual representation of a multi-layer perceptron: an interconnected network of nodes, showcasing the structure of input, hidden, and output layers that facilitate complex pattern recognition.

Bar chart comparing BLEU scores of Alpha, Beta, and Gamma models across Orion, Nebula, and Pulsar datasets, with Alpha consistently leading.



Experiments

We evaluate the following range of models on a held-out test set:

Experiments

We evaluate the following range of models on a held-out test set:

- LLAMA variants with 7 billion LLAMA_{7B} and 13 billion LLAMA_{13B} parameters

Experiments

We evaluate the following range of models on a held-out test set:

- **LLaMA** variants with 7 billion LLaMA_{7B} and 13 billion LLaMA_{13B} parameters
- **CLiMA** CLiMA_{7B} and CLiMA_{13B}, as well as CLiMA_{IMG} which has the compiled images as an additional input

Experiments

We evaluate the following range of models on a held-out test set:

- **LLaMA** variants with 7 billion LLaMA_{7B} and 13 billion LLaMA_{13B} parameters
- **CLiMA** CLiMA_{7B} and CLiMA_{13B}, as well as CLiMA_{IMG} which has the compiled images as an additional input

general-purpose models GPT-4 and Claude 2

Experiments

We evaluate the following range of models on a held-out test set:

- **LLaMA** variants with 7 billion LLaMA_{7B} and 13 billion LLaMA_{13B} parameters
- **CLiMA** CLiMA_{7B} and CLiMA_{13B}, as well as CLiMA_{IMG} which has the compiled images as an additional input

general-purpose models GPT-4 and Claude 2

To facilitate that we employ a range of automatic metrics:

Experiments

We evaluate the following range of models on a held-out test set:

- **LLaMA** variants with 7 billion LLaMA_{7B} and 13 billion LLaMA_{13B} parameters
- **CLiMA** CLiMA_{7B} and CLiMA_{13B}, as well as CLiMA_{IMG} which has the compiled images as an additional input

general-purpose models GPT-4 and Claude 2

To facilitate that we employ a range of automatic metrics:

- **code-based** CrystalBLEU, Extended-Edit Distance (EED)

Experiments

We evaluate the following range of models on a held-out test set:

- **LLaMA** variants with 7 billion LLaMA_{7B} and 13 billion LLaMA_{13B} parameters
- **CLiMA** CLiMA_{7B} and CLiMA_{13B}, as well as CLiMA_{IMG} which has the compiled images as an additional input

general-purpose models GPT-4 and Claude 2

To facilitate that we employ a range of automatic metrics:

- **code-based** CrystalBLEU, Extended-Edit Distance (EED)
- **image-based** CLIPScore, CLIPScore_{IMG}, Kernel Inception Distance (KID)

Experiments

We evaluate the following range of models on a held-out test set:

- **LLaMA** variants with 7 billion LLaMA_{7B} and 13 billion LLaMA_{13B} parameters
- **CLiMA** CLiMA_{7B} and CLiMA_{13B}, as well as CLiMA_{IMG} which has the compiled images as an additional input

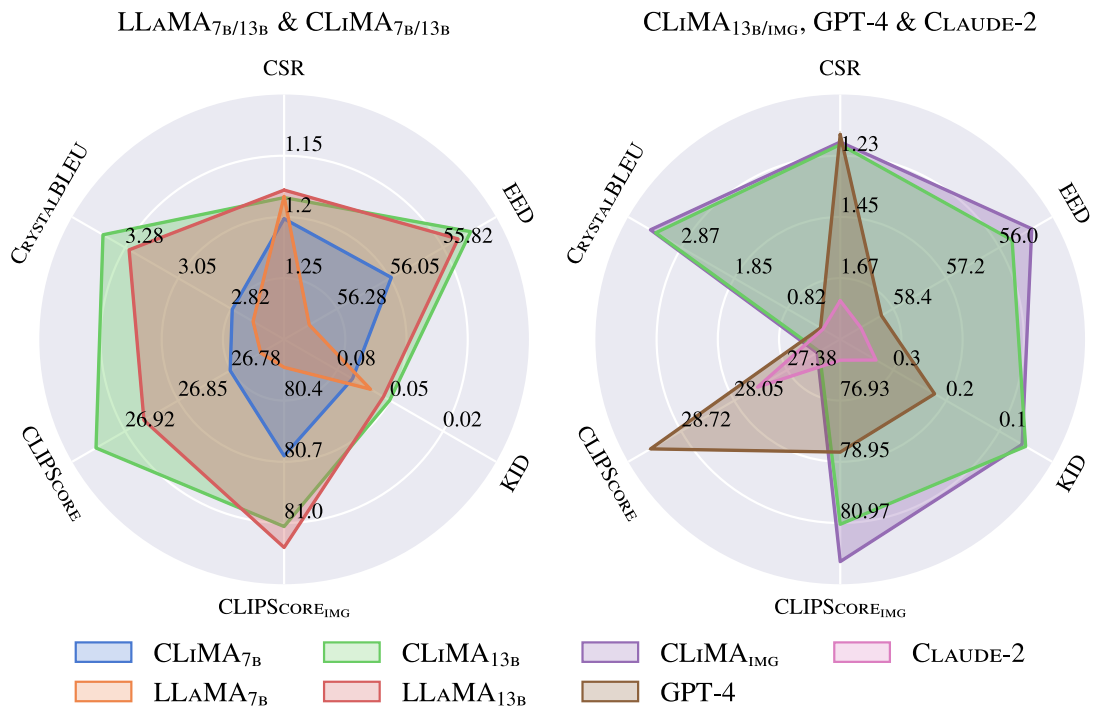
general-purpose models GPT-4 and Claude 2

To facilitate that we employ a range of automatic metrics:

- code-based CrystalBLEU, Extended-Edit Distance (EED)
- image-based CLIPScore, CLIPScore_{IMG}, Kernel Inception Distance (KID)
- sampling-based Compilation Sampling Rate (CSR)

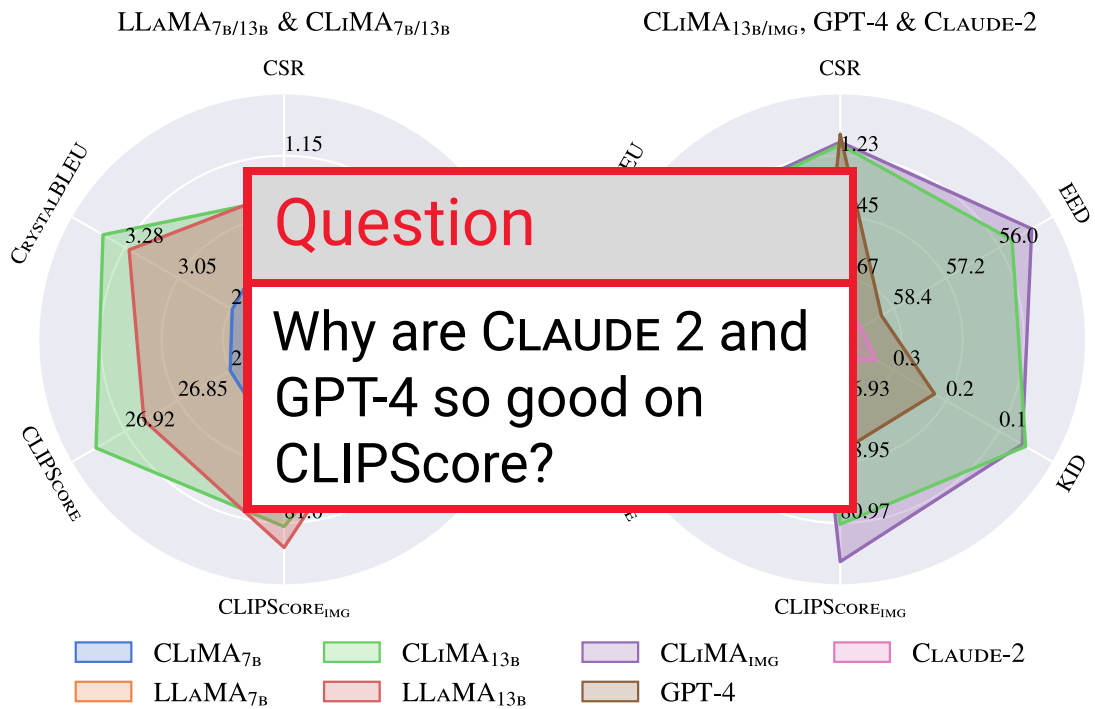
Results

Overall, CLIMA$_{7B}$ and CLIMA$_{13B}$ outperform their respective LLAMA models in five out of seven metrics each, with CLAUDE 2 and GPT-4 substantially underperforming all of them. While CLIMA$_{IMG}$ unsurprisingly improves upon CLIMA$_{13B}$, CLIMA$_{13B}$ is the best model with only textual inputs.



Results

Overall, CLIMA$_{7B}$ and CLIMA$_{13B}$ outperform their respective LLAMA models in five out of seven metrics each, with CLAUDE 2 and GPT-4 substantially underperforming all of them. While CLIMA$_{IMG}$ unsurprisingly improves upon CLIMA$_{13B}$, CLIMA$_{13B}$ is the best model with only textual inputs.

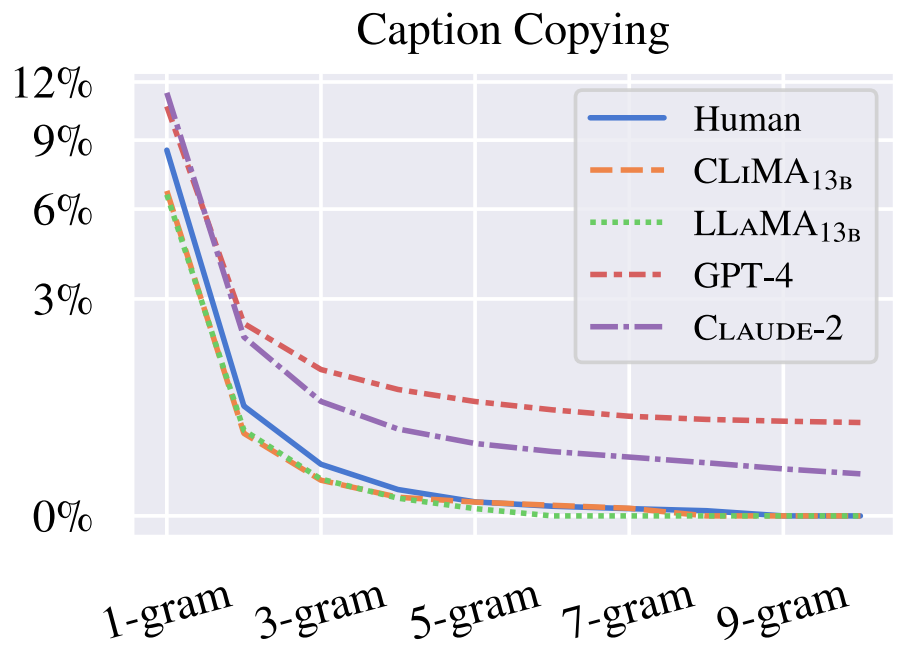


Question

Why are CLAUDE 2 and GPT-4 so good on CLIPScore?

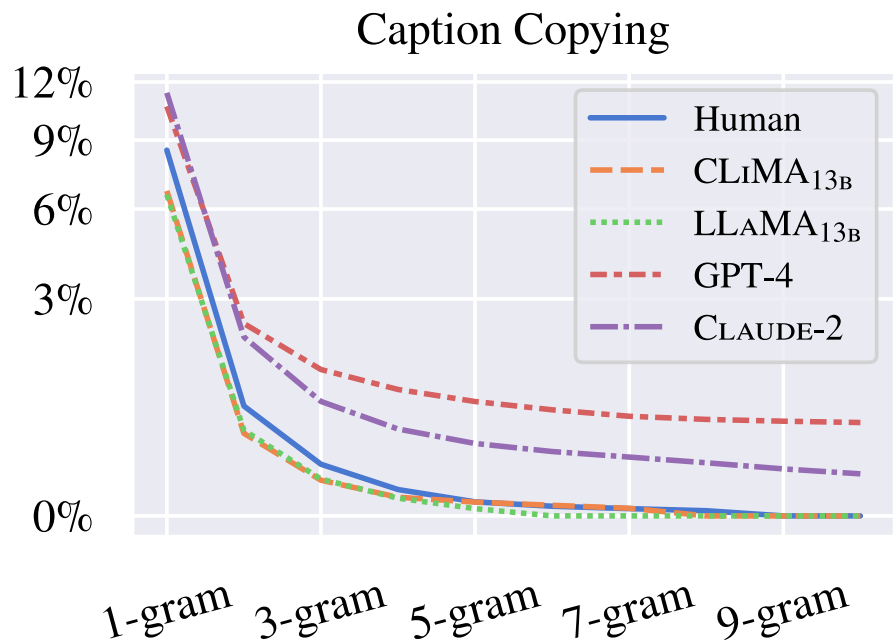
Caption Copying

We calculate the proportion of n-grams from the caption that were copied verbatim into the output code.



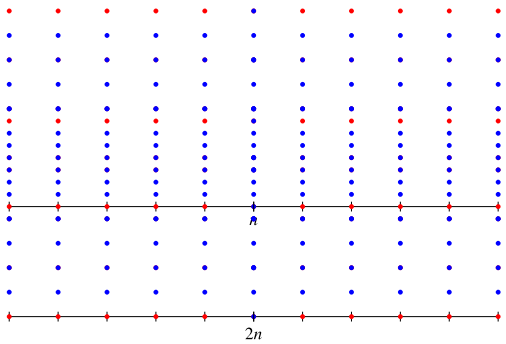
Caption Copying

We calculate the proportion of n-grams from the caption that were copied verbatim into the output code.

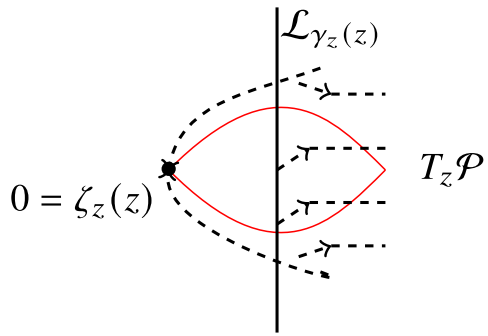


We find that GPT-4 and Claude 2 tend to produce degenerate images, which visibly copy n-grams of the input caption into the output image. This can trick CLIPScore into providing high scores.

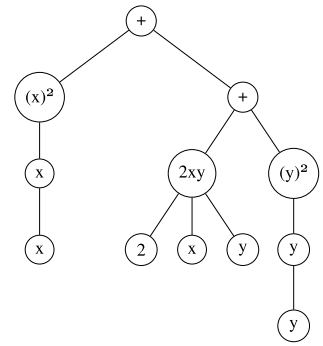
Examples



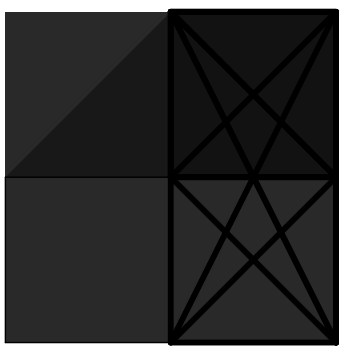
CLIMA$_{13B}$ (good): The ansatz (5.17) and (5.18) for $\alpha = 2$. The red points are evenly spaced, and the blue points scale quadratically with n. *The image is a white background with blue and red dots scattered across it. The dots are placed in various arrangements, creating a visually interesting pattern. Some dots are clustered together, while others are spaced further apart, covering the entire background.*



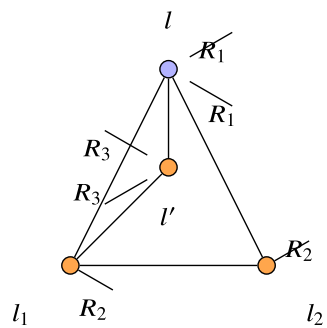
LLAMA$_{13B}$ (good): Local configuration of the path and the foliation (in red) around the point $0 = \Psi_z(z)$. *The image displays a complex mathematical formula with various symbols and notation, including a black dot, waves, and arrows. The formula seems to be related to physics or engineering, and it is written on a white background for easy readability.*



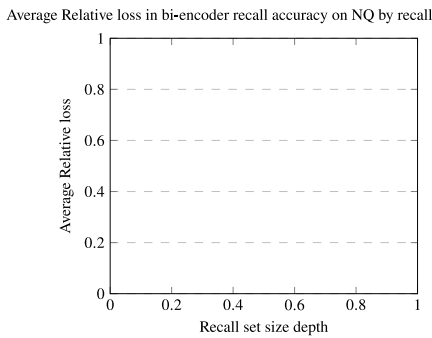
GPT-4 (good): Expression graph for algebraic expression $x^2+2xy+y^2$. *The image shows a tree with a symbol at its root, representing a mathematical concept. The tree has a series of logical connections, and there are variables and mathematical symbols throughout the structure. The image conveys a sense of order and organization in the presentation of the mathematical concept.*



CLIMA$_{13B}$ (bad): For the $P = 2$ scheme, the regular cell footprint is a standard five-point Laplacian, and if any point in the footprint is a cut cell, it is then "irregular." Cut cells are shown with dark shading, irregular cells with light shading, and the remaining white cells are "regular."



LLAMA$_{13B}$ (bad): Illustration of the proof of Theorem 2.7. *The image displays a tree with blue and orange labels on a white background. The tree has a unique structure, with branches that don't follow a typical tree layout. The labels seem to be representing a formula or a set of instructions, and the tree is accompanied by several equations in the surrounding area.*



GPT-4 (bad): Average Relative loss in bi-encoder recall accuracy on NQ by recall set size depth on the baseline, Pretrained Alignment (PT), Data Augmentation (DA), and Contrastive Alignment Post Training (CAPOT) on noisy queries.

Interested? There's more!

In our paper, we conduct a human evaluation and also demonstrate that all models exhibit few memorization problems and generate novel outputs. However, we also show that GPT-4 and CLAUDE 2 tend to generate simpler code than the other models, that compiles to less complex images.

Paper	`https://arxiv.org/abs/2310.00367`
Code	`https://github.com/potamides/AutomaTikZ`
Dataset	`https://huggingface.co/datasets/nllg/datikz`
Demo	`https://huggingface.co/spaces/nllg/AutomaTikZ`
Group	`https://nl2g.github.io`