Scaling and Evaluating Sparse Autoencoders

Authors: Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, Jeffrey Wu

Publisher: OpenAI Superintelligence Interpretability Team

Presenters: Yu Wang and Keshav Worathur

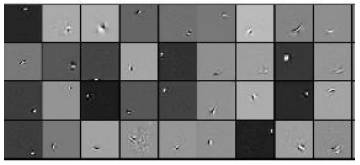
Georgia Tech

Motivation

Motivation

Why study sparse autoencoders?

- Learn interpretable features for **mechanistic interpretability**
  - Important as black box models with increasingly complex architectures proliferate in industry.



(a) $k = 70$

(d) $k = 10$

As we increase sparsity (decrease k), features become more interpretable

Science & technology | Artificial intelligence

Researchers are figuring out how large language models work

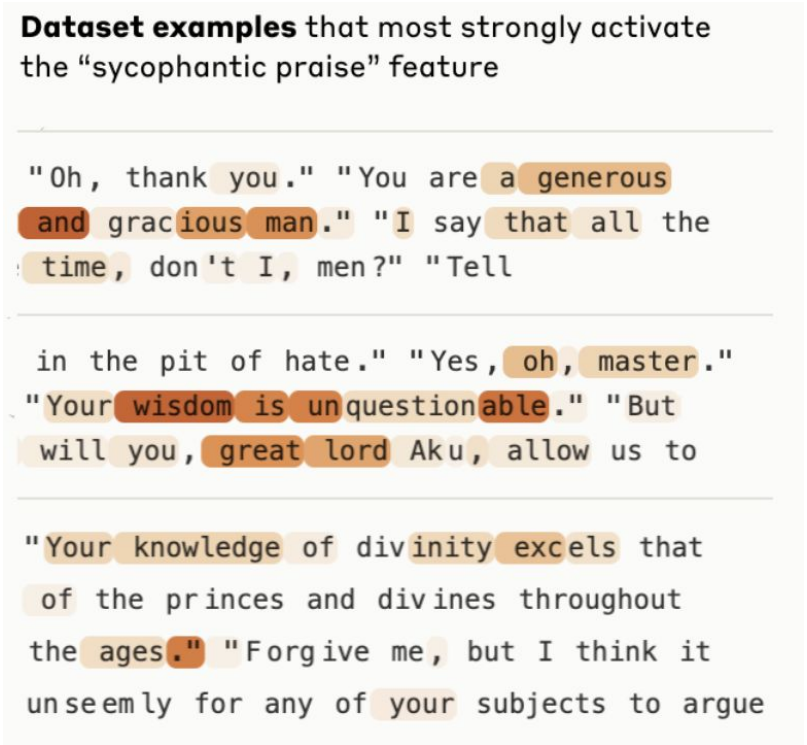
Such insights could help make them safer, more truthful and easier to use

Antropic Researchers Apply Sparse Autoencoders to Claude Sonnet 3 to Learn Interpretable Features (source: The Economist)

Motivation, Continued

How do we apply sparse autoencoders (SAE) to interpret LLMs?

- Intermediate layers of an LLM output **embeddings**
- 💡 “slice open” an LLM at one of these layers and inspect the embeddings.
- Problem: embeddings are too fine grained to be interpretable.
  - In GPT2, each embedding is a 768 dimensional vector
- Solution: autoencoders trained on the embeddings!



Problem: Dead Latents

- Hypothesis: Larger LMs -> high dimensional AE latents
- Training sparse autoencoders on LLM features is difficult:
  - If we make the autoencoder *too wide* (increase embedding dim) or *too sparse* (stronger regularization), we encounter **dead latents**
- Dead latents: units that don't activate for 10 million tokens or more, and they add significant training overhead.

- **Proportion of Dead Features**:
  - **1M SAE**: ~2% dead features.
  - **4M SAE**: ~35% dead features.
  - **34M SAE**: ~65% dead features.

Percentage of Dead Latents Increases as SAE becomes Wider (Templeton et. al, 2024) *Scaling monose-*
manticity: Extracting interpretable features from claude 3 sonnet. Transformer Circuits Thread.

Background

Background: k-sparse autoencoders

Adding a Sparsity Objective to Training

- L0 regularization is the most explicit way of enforcing sparsity
  - problem: this function is not differentiable!
  - previously: use L1 regularization, which makes activations small and adds a new hyperparameter

Baseline Method

$$
\begin{cases}
z = \mathbf{ReLU}(W_{\text{enc}}(x - b_{\text{pre}}) + b_{\text{enc}}) \\
\hat{x} = W_{\text{dec}} z + b_{\text{pre}} \\
\mathcal{L} = ||x - \hat{x}||_2^2 + \lambda ||z||_1
\end{cases}
$$

Background: k-sparse autoencoders, continued

- Simple non-linear activation that acts just like an L0 regularizer by decreasing value of k
- Can be used standalone or in combination with other activation functions

$$\text{TopK} : \mathbb{R}^m \to \mathbb{Z}^k$$

1. **Input**: hidden layer values $\mathbf{z} \in \mathbb{R}^{256}$

$$\mathbf{z} = \begin{bmatrix} 1.53 \\ 0.45 \\ -2.34 \\ \vdots \\ 5.89 \end{bmatrix}$$

Background: k-sparse autoencoders, continued

- Simple non-linear activation that acts just like an L0 regularizer by decreasing value of k
- Can be used standalone or in combination with other activation functions

$$\text{TopK} : \mathbb{R}^m \rightarrow \mathbb{Z}^k$$

2. Sort hidden layer values and keep top k values (example shown for $k = 3$)

$$\begin{bmatrix} 5.89 \\ 1.53 \\ 0.45 \\ \vdots \\ \sout{2.34} \end{bmatrix}$$

Background: k-sparse autoencoders, continued

- Simple non-linear activation that acts just like an L0 regularizer by decreasing value of k
- Can be used standalone or in combination with other activation functions

$$\text{TopK} : \mathbb{R}^m \to \mathbb{Z}^k$$

3. Return $\text{supp}_k(\mathbf{z})$, the indices of the largest values (example shown for $k = 3$)

$$\text{supp}_3(\mathbf{z}) = \begin{bmatrix} 255 \\ 1 \\ 0 \end{bmatrix}$$

Contributions

Contributions, Continued

- Novel training procedure for sparse autoencoders on LLMs
- Scaling laws for autoencoders trained on GPT-2 and GPT-4 residual activations
- New metrics for evaluating feature quality:
  - feature recovery
  - explainability of features
  - sparsity of downstream effects

Contributions, Continued

Training Technique for Mitigating Dead Latents

- Initialize the encoder weights to be the same as the decoder weights.
- Train autoencoder on the residuals of embeddings

$$z = \mathrm{TopK}(W_{\mathrm{enc}}(x - b_{\mathrm{pre}}))$$

- Use embeddings from a layer close to end of the network that is not specialized for next token prediction (based on intuition!)
- Add an auxiliary loss function $\mathcal{L}_{\mathrm{aux}}$

Contributions, Continued

Auxiliary Loss

- Compute forward pass of autoencoder on input x to obtain reconstruction x^hat
- Let $e = x - \hat{x}$ be the reconstruction loss
- Replace ReLU activation of latents to exponential function
- Compute the decoded vector $\hat{e}$ using the top k_{aux} dead latents
- Compute the auxiliary loss as

$$\mathcal{L}_{\text{aux}} = ||e - \hat{e}||_2^2$$

- Scale aux loss by a small constant (e.g. 1/32) and add to reconstruction loss

$$\mathcal{L} + \alpha \mathcal{L}_{\text{aux}}$$

Contributions, Continued

- Novel training procedure for sparse autoencoders on LLMs
- Scaling laws for autoencoders trained on GPT-2 and GPT-4 residual activations
- New metrics for evaluating feature quality:
  - feature recovery
  - explainability of features
  - sparsity of downstream effects

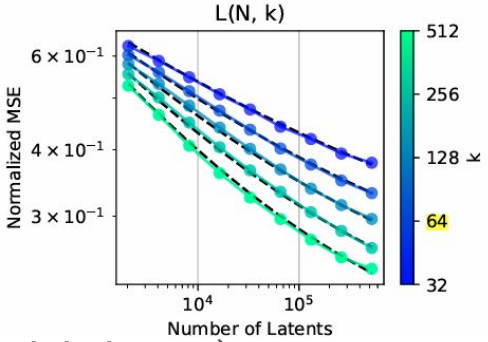
Contributions, Continued

Scaling Law: **number of latents (n), sparsity level (k)**

- Training to compute-MSE frontier (L(C)) - unprincipled
- Training to convergence(L(N)) - GPT-2 small $\Theta(n^{0.6})$ and GPT-4 $\Theta(n^{0.65})$

Jointly fitting sparsity (L(N,K))

$$L(n,k) = \underbrace{\exp(\alpha + \beta_k \log(k) + \beta_n \log(n) + \gamma \log(k)\log(n))}_{\text{Scaling loss}} + \underbrace{\exp(\zeta + \eta \log(k))}_{\text{Irreducible loss}}$$



γ=−0.042 (exponent on n: β_n+γlogk, loss decreases faster with n when k is larger)

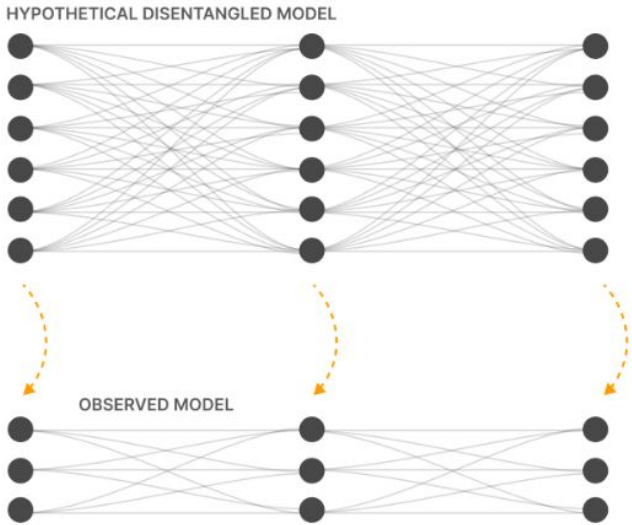
η=−0.085, as k increases, this term decreases

Using Sparse Autoencoders to Interpret LMs

Neurons are **polysemantic**: they respond to mixtures of seemingly unrelated inputs
Superposition: neural networks represent more features than they have neurons

Sparse autoencoders for feature decompositions

- Decompose dense activations into sparse, interpretable component features
- Scale to very large datasets
- Similar in architecture with neural networks just strong enough to recover features



Contributions, Continued

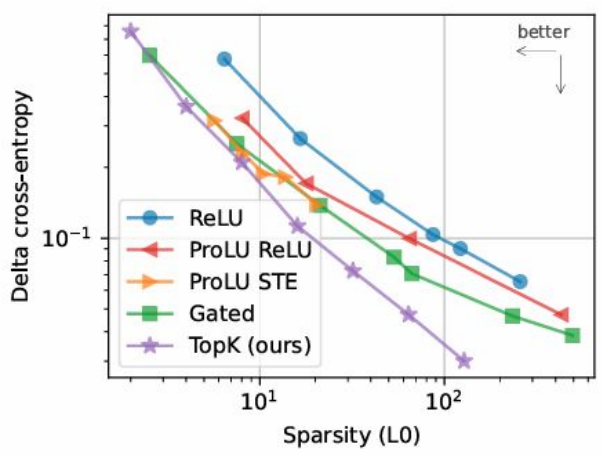
- Novel training procedure for sparse autoencoders on LLMs
- Scaling laws for autoencoders trained on GPT-2 and GPT-4 residual activations
- New metrics for evaluating feature quality:
  - feature recovery
  - explainability of features
  - sparsity of downstream effects

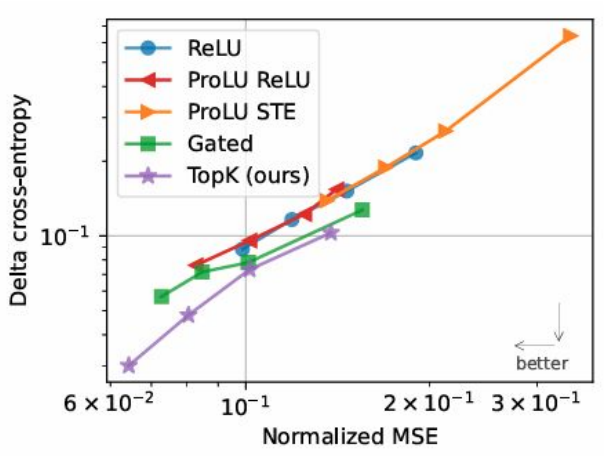
Evaluation

Downstream loss when replacing the residual stream by the reconstructed value

Compute downstream Kullback-Leibler (KL) divergence and cross-entropy loss.

- Relative amount of pretraining compute instead of raw loss for better interpretation.





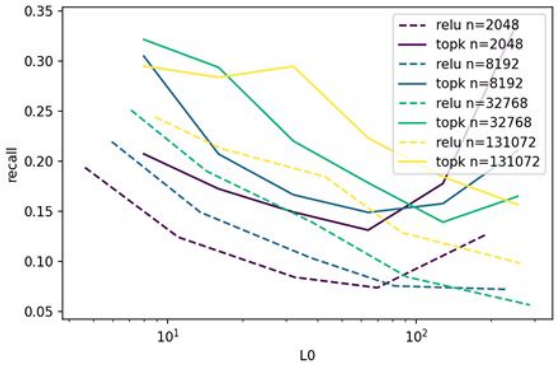
Evaluation

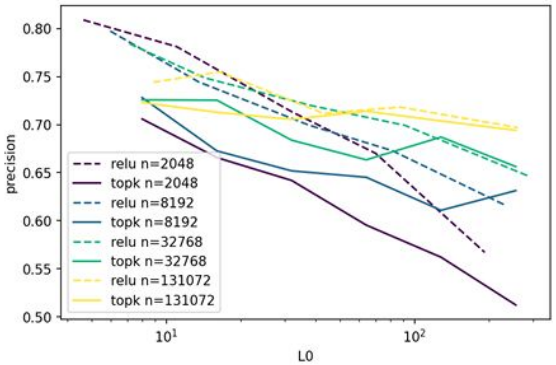
Probe loss checks whether features are present

- 1d logistic probe on each latent using the Newton-Raphson

$$\min_{i,w,b} \mathbb{E}\left[y \log \sigma\left(wz_i + b\right) + (1-y) \log\left(1 - \sigma\left(wz_i + b\right)\right)\right]$$

Pros: computationally cheap **Cons:** known features

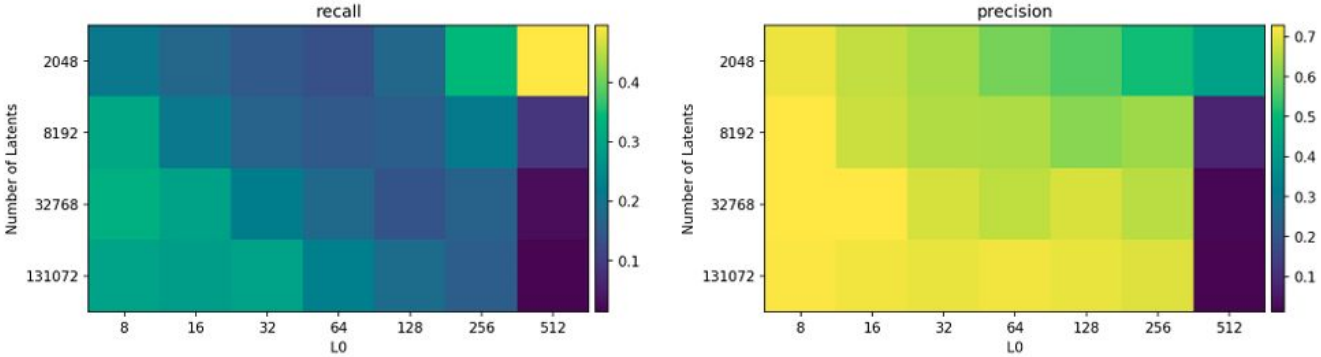




Evaluation

Explainability-Neuron to Graph(N2G) Intepretable graph representation of neurons

- Predicting activations on input text
- Evaluate representation by comparing to the ground truth activations
- Efficiently measure recall, precision, f1 score



(a) Recall of N2G explanations
$P(\text{n2g} > 0 | \text{act} > 0)$

(b) Precision of N2G explanations
$P(\text{act} > 0 | \text{n2g} > 0)$

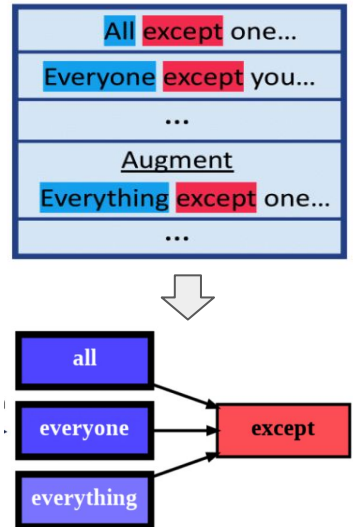


Fig. Example N2G Graph

Conclusion

The paper uses k-sparse autoencoder to directly control sparsity for extracting interpretable features and effectively scaled to GPT-4. Additionally, the paper demonstrates scaling laws for training SAEs and filling the gap in the literature to explore evaluation metrics.

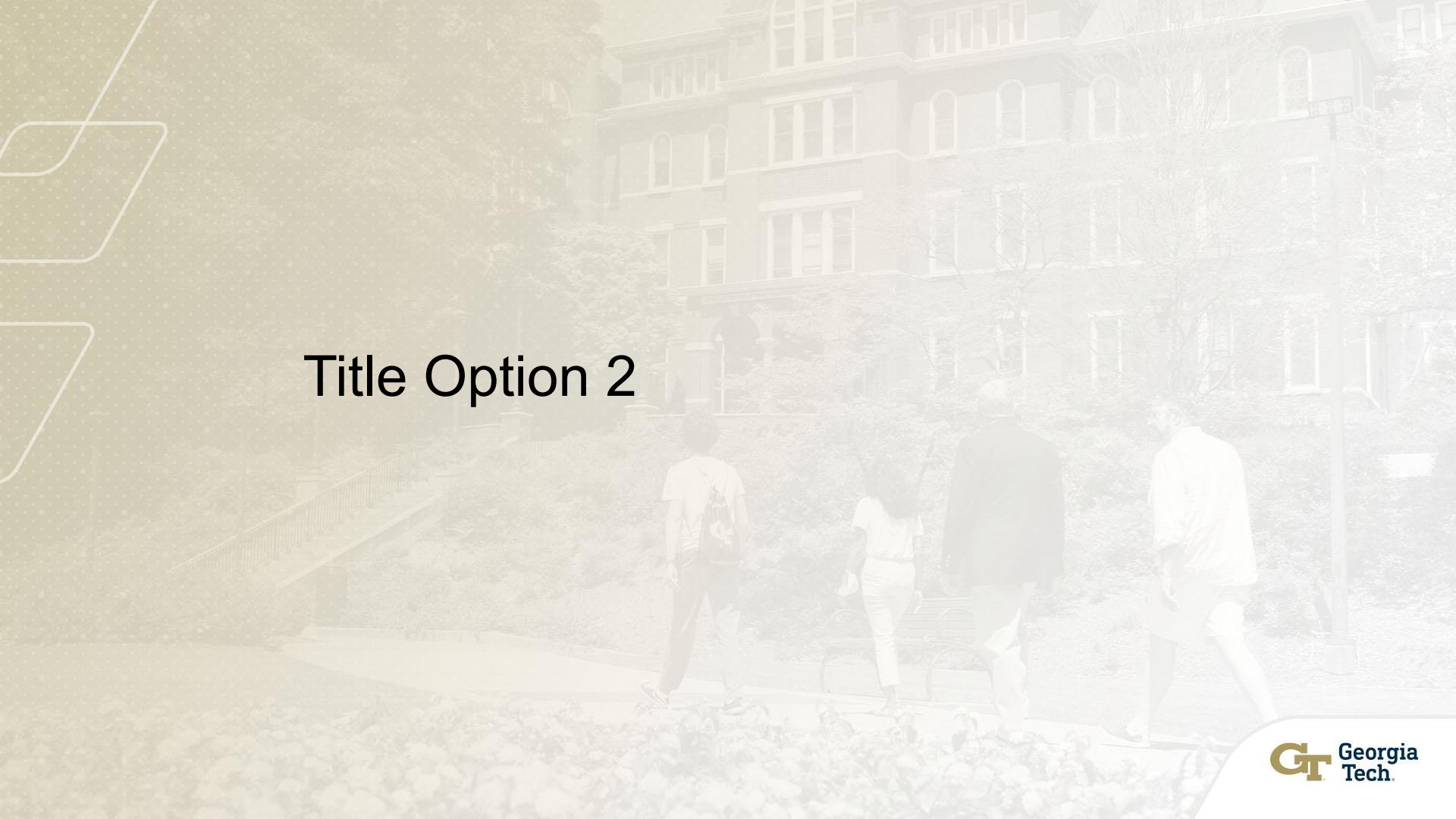
Limitations:

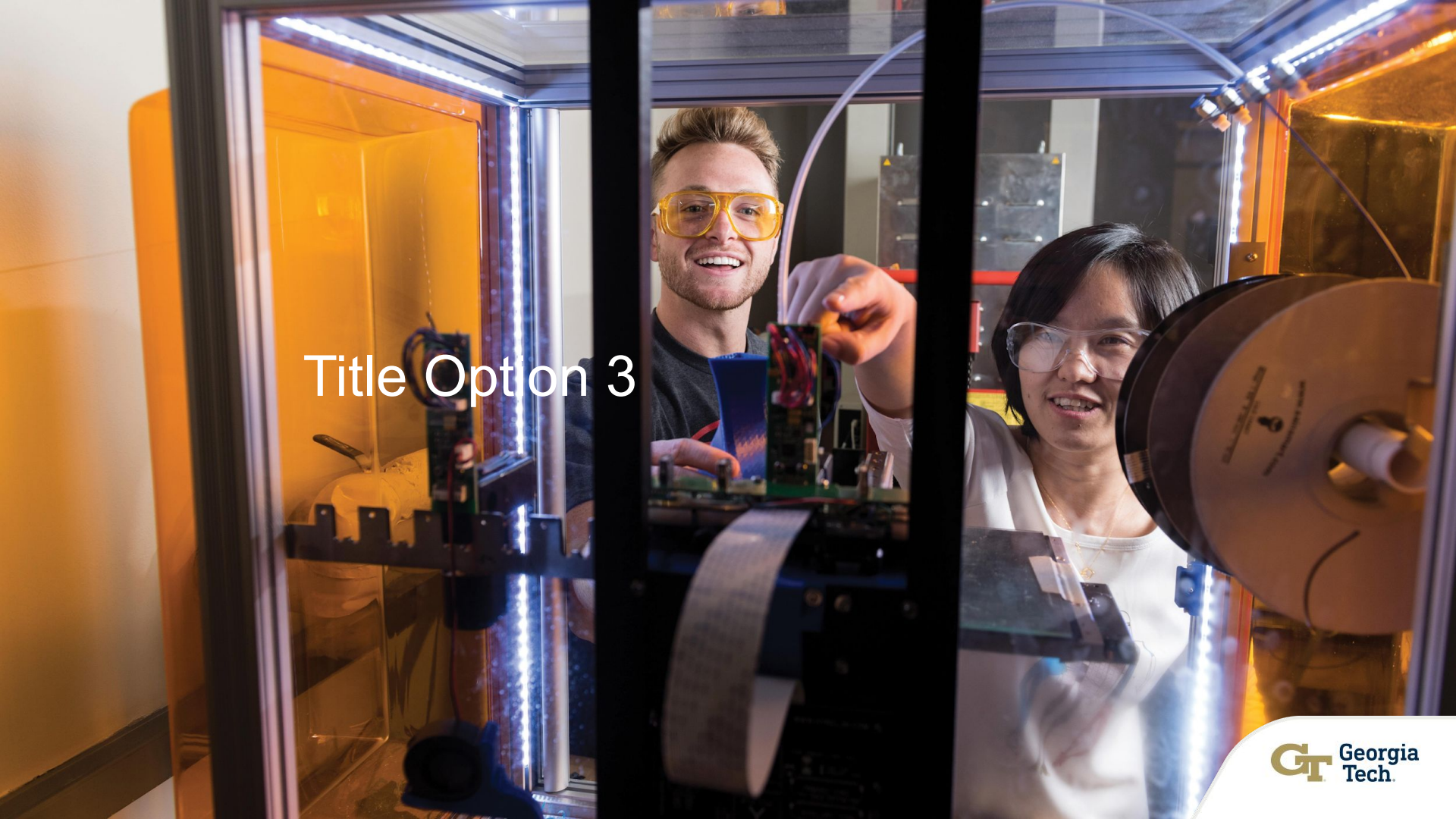
- Only tested up to 64 tokens, while GPT-4 uses much more
- Fixed k for every token
- Evaluations only on GPT-4, closed source

Title Option 1

Georgia Tech

Title Option 2




Title Option 3

Content Blocks

Georgia Tech

Headlines – 36pt Roboto Bold

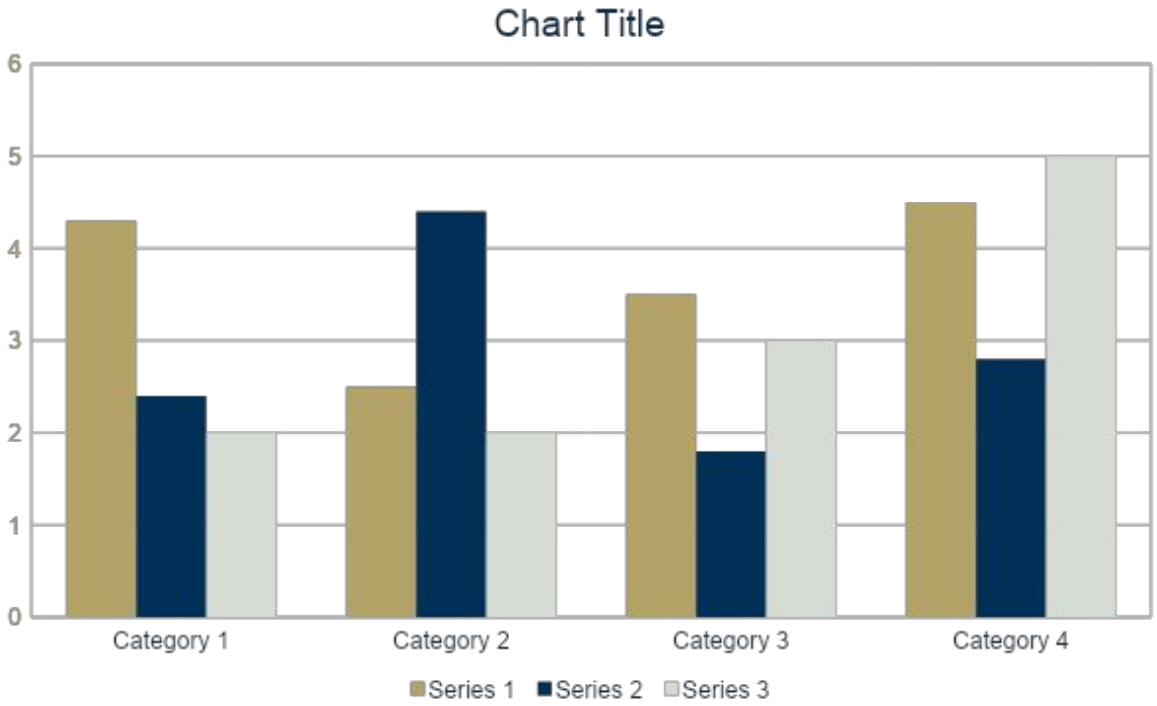
- Main text – Navy Roboto 24px.
  - Example of a sub point. 18px Roboto
  - Subpoint example
    - Sub-sub point example the first
    - And the second

Headlines with Text and Images

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Nullam placerat purus vitae pellentesque semper. Duis urna tortor, aliquam eu urna et, mattis molestie orci. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia curae; Cras tempus neque id ligula commodo condimentum. Etiam dui eros, rutrum et tellus at, maximus finibus dolor. Fusce eget tellus euismod, volutpat orci id, convallis lorem. Proin ultricies pharetra ligula vitae blandit. Aliquam porttitor mi nec finibus lacinia. Curabitur pretium convallis tellus, eu egestas nisi tincidunt ut.

Full Page Photo

Charts



- Captions for charts
- Please use chart example designs on slides 14–17.

Shapes

Hex Shape Options and Text Fill

GOLD

Shape Fill RGB - 179, 163, 105

Shape Stroke RGB - 133, 116, 55

NAVY

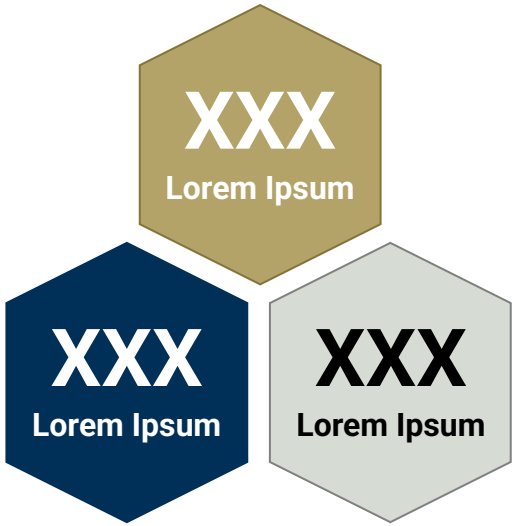
Shape Fill RGB – 0, 48, 87

Shape Stroke RGB – 23, 21, 67

PI MILE

Shape Fill RGB – 214, 219, 212

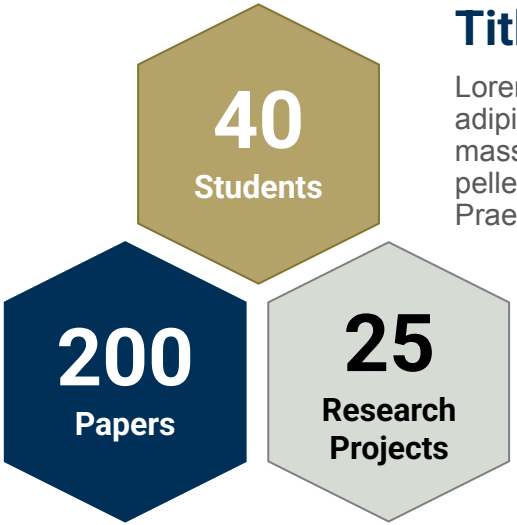
Shape Stroke RGB – 128, 128, 128



- *Hex shape content should be used for displaying minimal information and statistics. Font color to use for shapes is only white and black, with black only being used for Gold and Pi Mile color options. Examples will be displayed on next page for common uses.*

Georgia Tech

Hex Shape Examples

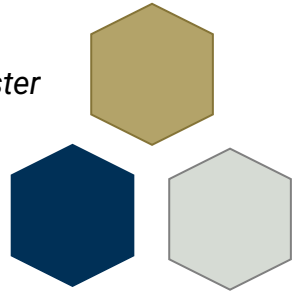


Title Goes Here

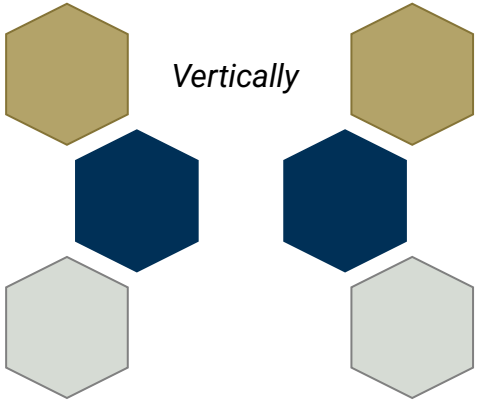
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Fusce dapibus nibh et massa auctor faucibus. Aliquam pellentesque egestas nulla in placerat. Praesent eu varius risus.

Horizontally

Cluster



Vertically



- *Hex shapes can be displayed in various ways. Vertically, Horizontally, or as clusters. Shapes should be evenly spaced from one another with proper spacing. Shape colors options should stay as the ones provided, Gold, Navy, and Pi Mile.*

Hex Shape Options and Text Fill

GOL
D

Shape Fill RGB - 179, 163, 105

Shape Stroke RGB - 133, 116, 55

NAV
Y

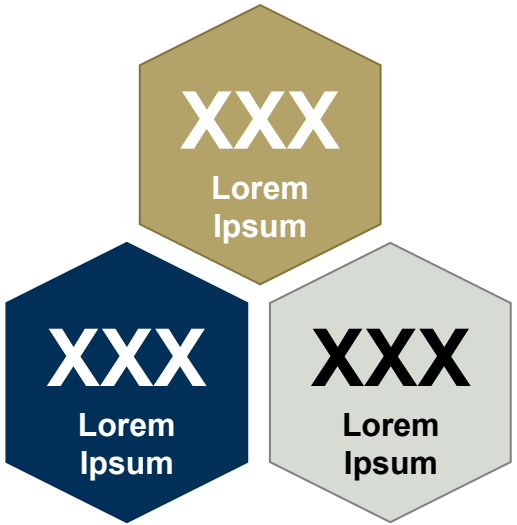
Shape Fill RGB – 0, 48, 87

Shape Stroke RGB – 23, 21, 67

PI
MILE

Shape Fill RGB – 214, 219, 212

Shape Stroke RGB – 128, 128, 128

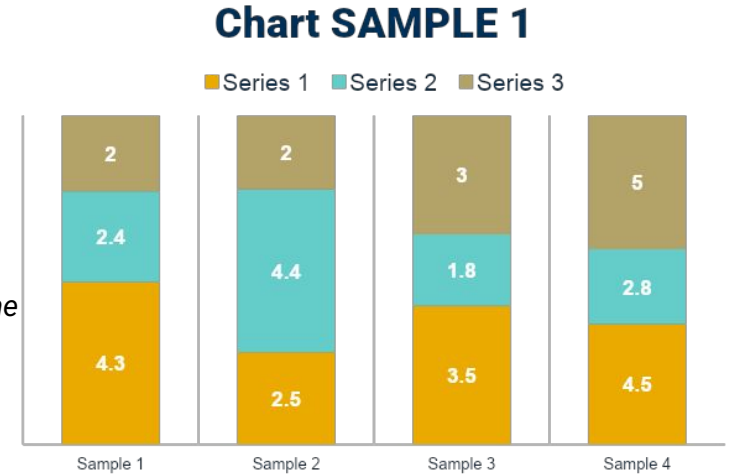


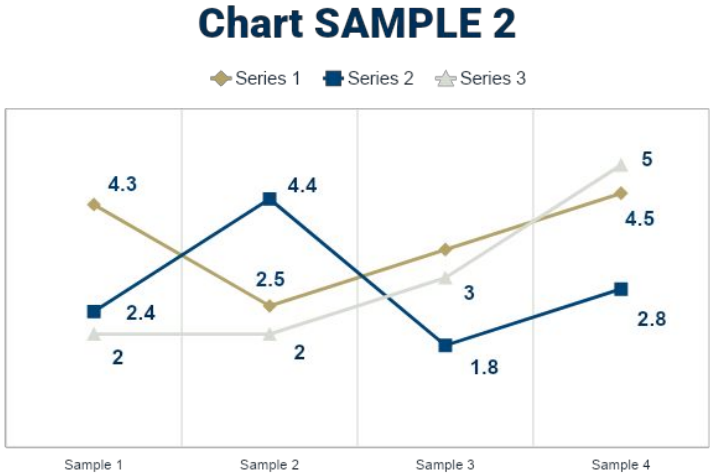
- *Hex shape content should be used for displaying minimal information and statistics. Font color to use for shapes is only white and black, with black only being used for Gold and Pi Mile color options. Examples will be displayed on next page for common uses.*

Charts

Charts

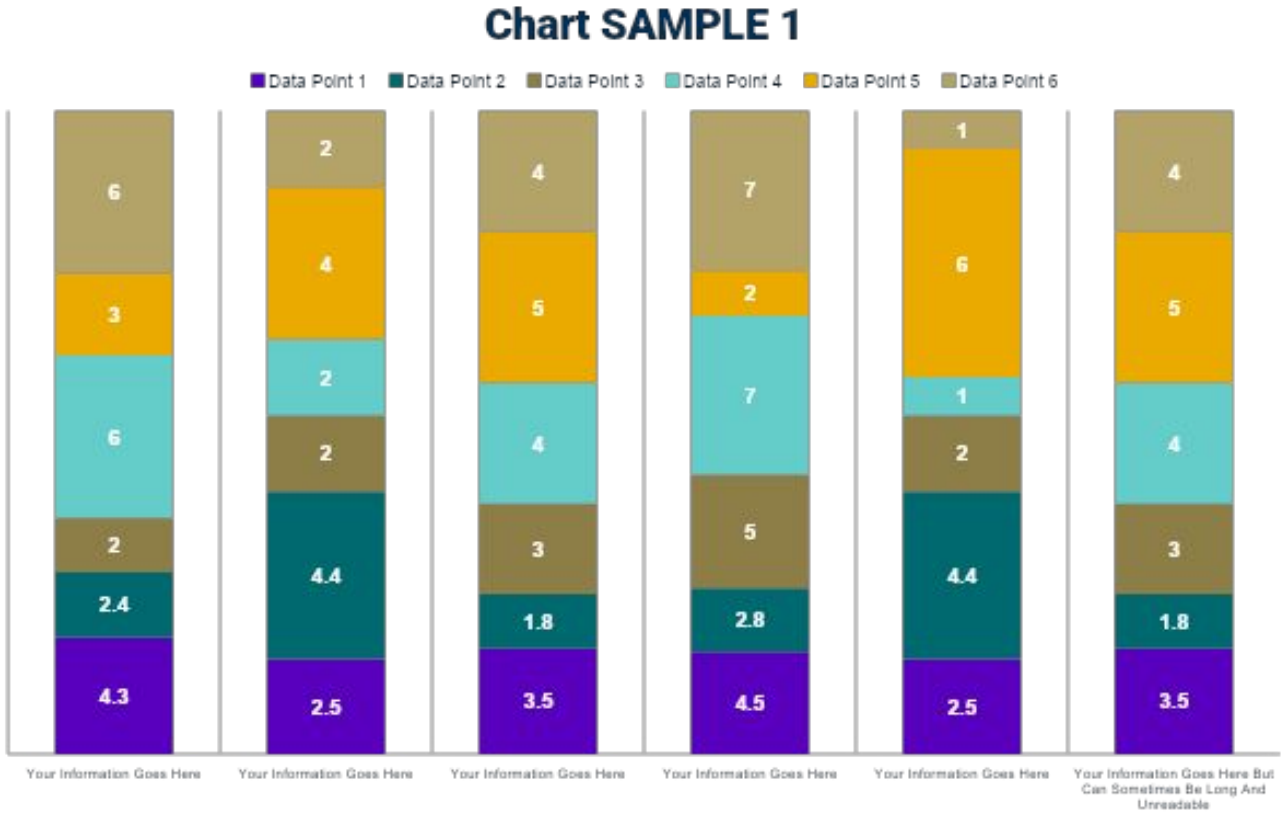
- ***Chart Sample 1*** *utilizes a Column known as 3D Stacked within Powerpoint. Samples and Series numbers can be increased as needed. It incorporates Gold, Navy, and Dark Gold as color options, but can be changed out for other colors within the secondary and tertiary colors located in the Georgia Tech Brand Standards Manual.*





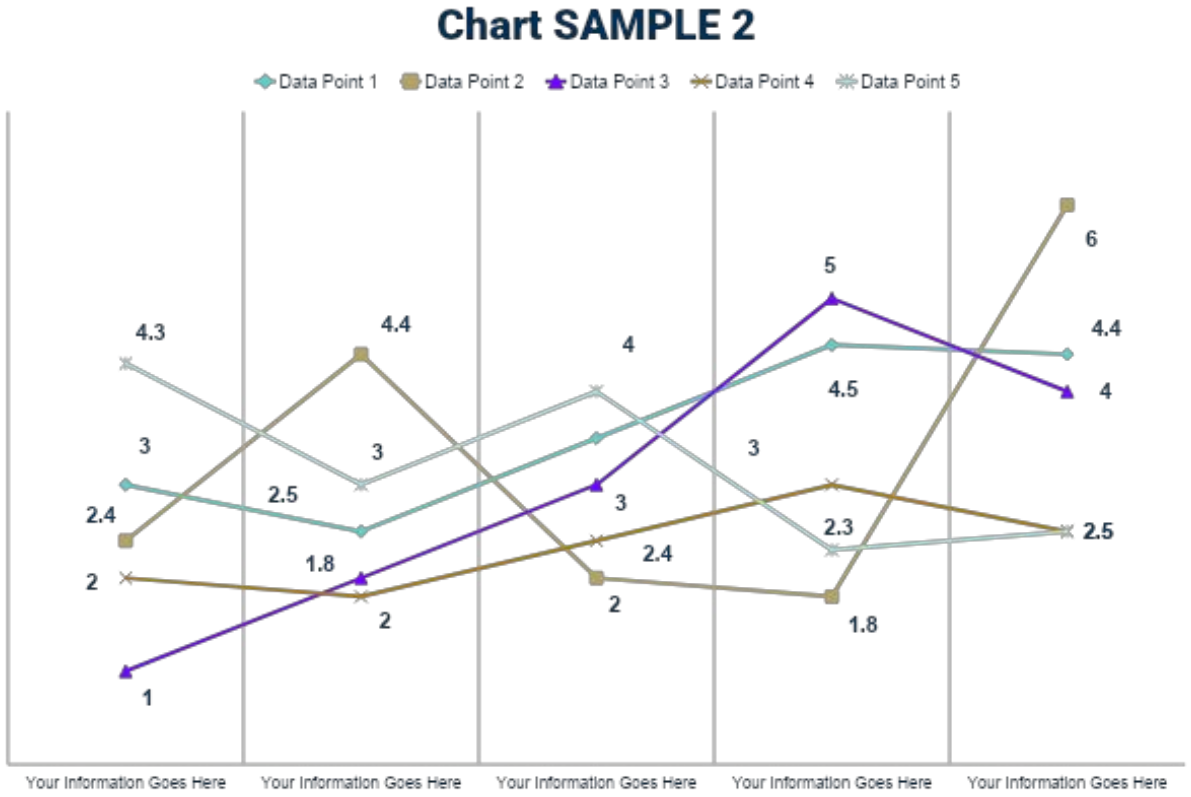
- ***Chart Sample 2*** *utilizes a Line chart known as Line Markers within Powerpoint. Samples and Series numbers can be increased as needed. It incorporates Gold, Navy, and Dark Gold as color options, but can be changed out for other colors within the secondary and tertiary colors located in the Georgia Tech Brand Standards Manual.*

Chart Sample 1



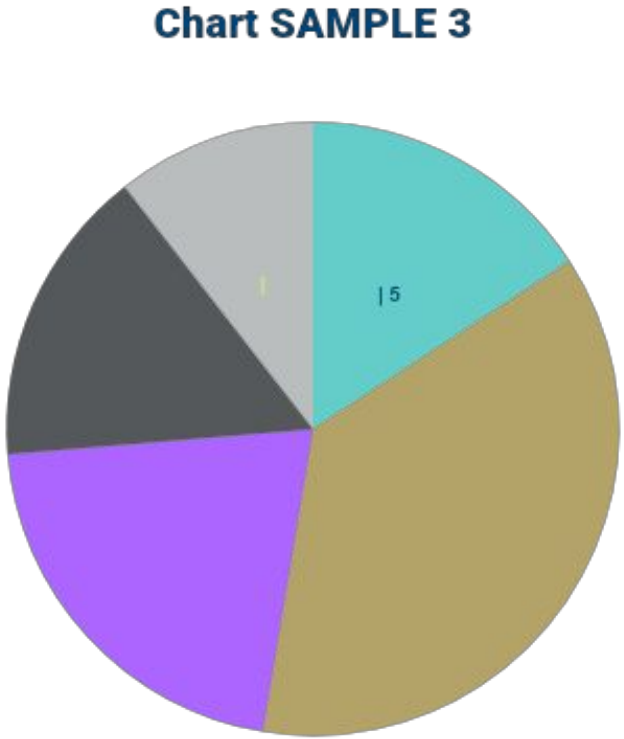
- ***Chart Sample 1*** *using multiple data points and descriptors below. The chart should not exceed more than 6 data points as this will make the information below less readable.*

Chart Sample 2



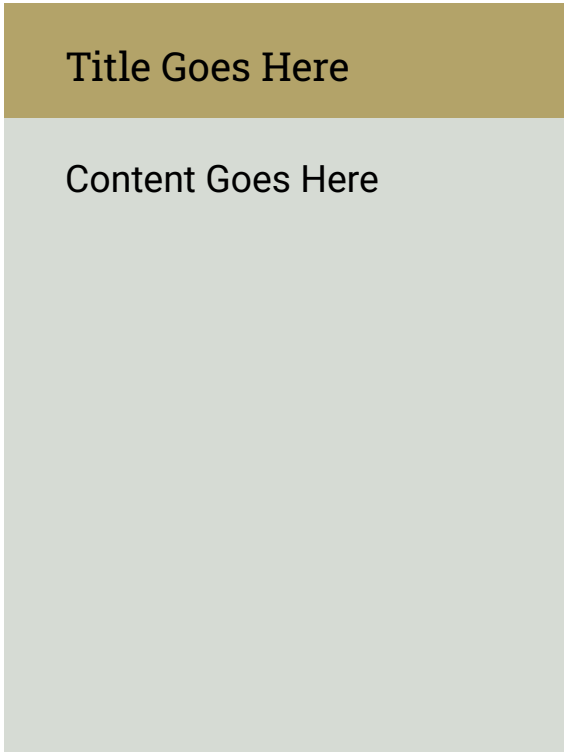
- ***Chart Sample 2*** *using multiple data points and descriptors below. The chart should not exceed more than 5 data point as this will make the information below less readable.*

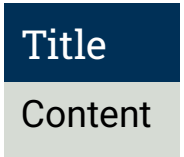
Chart Sample 3

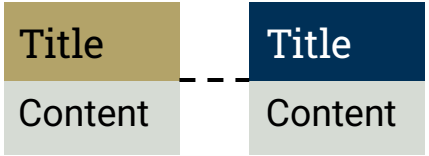


- ***Chart Sample 3*** *using multiple data points and descriptors below. The pie chart is more flexible with data points, but should not exceed more than 40 characters.*

Additional Chart Layouts







- *Linking charts based on context with the use of a dashed line.*

- *Stacked chart layout should only utilize three color options, Gold, Navy, and PiMile. Gold and Navy being used for the top title bar, while Pi Mile as the content block.*
- *Text size for Title should be 2pts larger than the content below it.*

Additional Chart Layouts Examples

Title Goes Here	Title Goes Here	Title Goes Here
Phasellus hendrerit facilisis nisi eget consequat. Donec vitae nunc risus. In eu augue aliquam, imperdiet est et, accumsan ante. Integer et vestibulum quam, nec varius urna. Pellentesque in ornare massa. Proin lobortis leo non lectus aliquam venenatis.	Phasellus hendrerit facilisis nisi eget consequat. Donec vitae nunc risus. In eu augue aliquam, imperdiet est et, accumsan ante. Integer et vestibulum quam, nec varius urna. Pellentesque in ornare massa. Proin lobortis leo non lectus aliquam venenatis.	Phasellus hendrerit facilisis nisi eget consequat. Donec vitae nunc risus. In eu augue aliquam, imperdiet est et, accumsan ante. Integer et vestibulum quam, nec varius urna. Pellentesque in ornare massa. Proin lobortis leo non lectus aliquam venenatis.

Icons

Georgia Tech

Icons

*Icons should only be used as a visual representation of an action or idea. Icons can be located under **Insert** in PPT. There you will find a variety of icons to choose from. Color options for icons can vary based on use but staying within Georgia Tech's brand should be key.*

Icon Use Examples

Title Goes Here

Phasellus hendrerit facilisis nisi eget consequat. Donec vitae nunc risus. In eu augue aliquam, imperdiet est et, accumsan ante.

Phasellus hendrerit facilisis nisi eget consequat. Donec vitae nunc risus. In eu augue aliquam, imperdiet est et, accumsan ante.

Phasellus hendrerit facilisis nisi eget consequat. Donec vitae nunc risus. In eu augue aliquam, imperdiet est et, accumsan ante.

Phasellus hendrerit facilisis nisi eget consequat. Donec vitae nunc risus. In eu augue aliquam, imperdiet est et, accumsan ante.