Controllable Language Generation at Every Scale

Balancing Precision and Resources

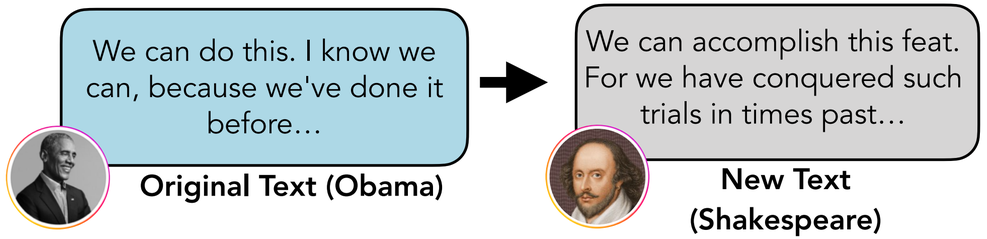
Presented by Jillian Fisher
General Exam 12/4/2024

Controllable Generation

Method which directs a model's output to meet specific criteria.

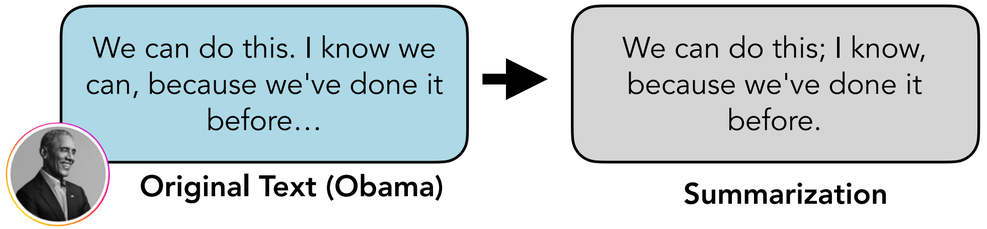
Style Transfer

Criteria: *Target Style*



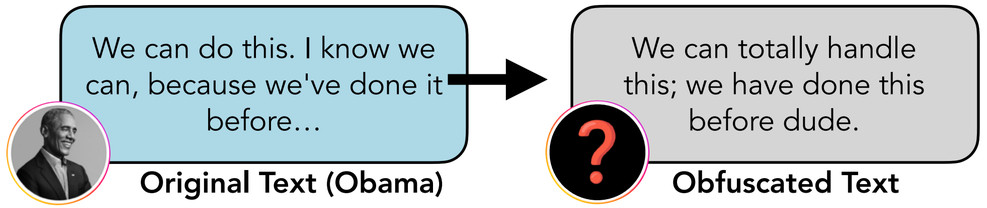
Summarization

Criteria: *Shorter*



Authorship Obfuscation

Criteria: *Not Original Author Style*



Controllable Generation

Authorship Obfuscation

We can do this. I know we can, because we've done it before…

Original Text (Obama)

We can totally handle this; we have done this before dude.

Obfuscated Text

What?

Rewriting text to obscure the original author's identity

*Should maintain the *content* and *sentiment**

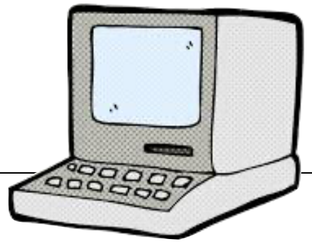
Why?

Blind Review for Scientific Papers

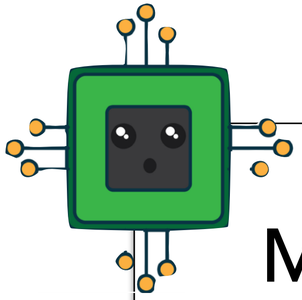
Interaction on Mental Health Forums

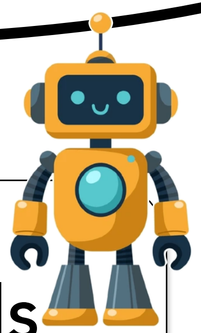
Anonymous Online Review

Controllable Generation



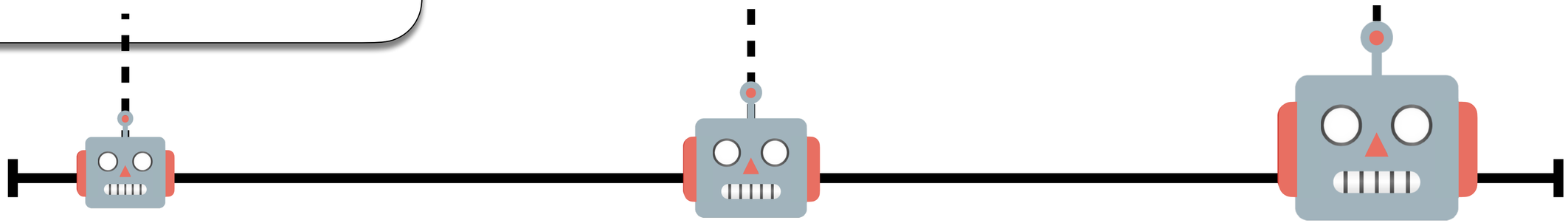
Small Models
(<2B)





Medium Models
(2B - 10B)

Large Models
(>10B)



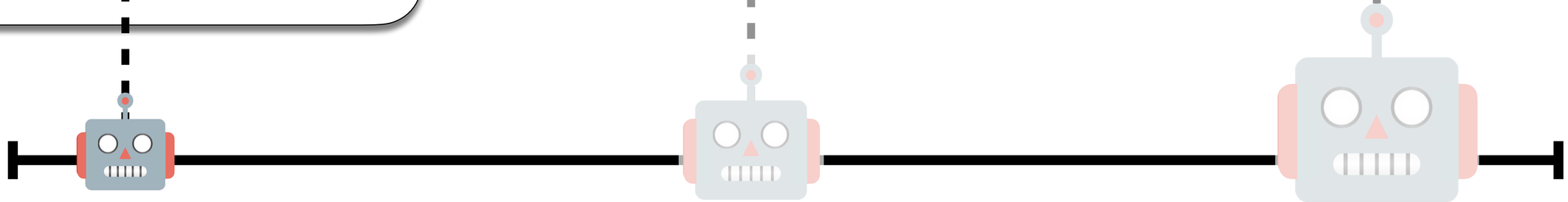
Model Size

Controllable Generation

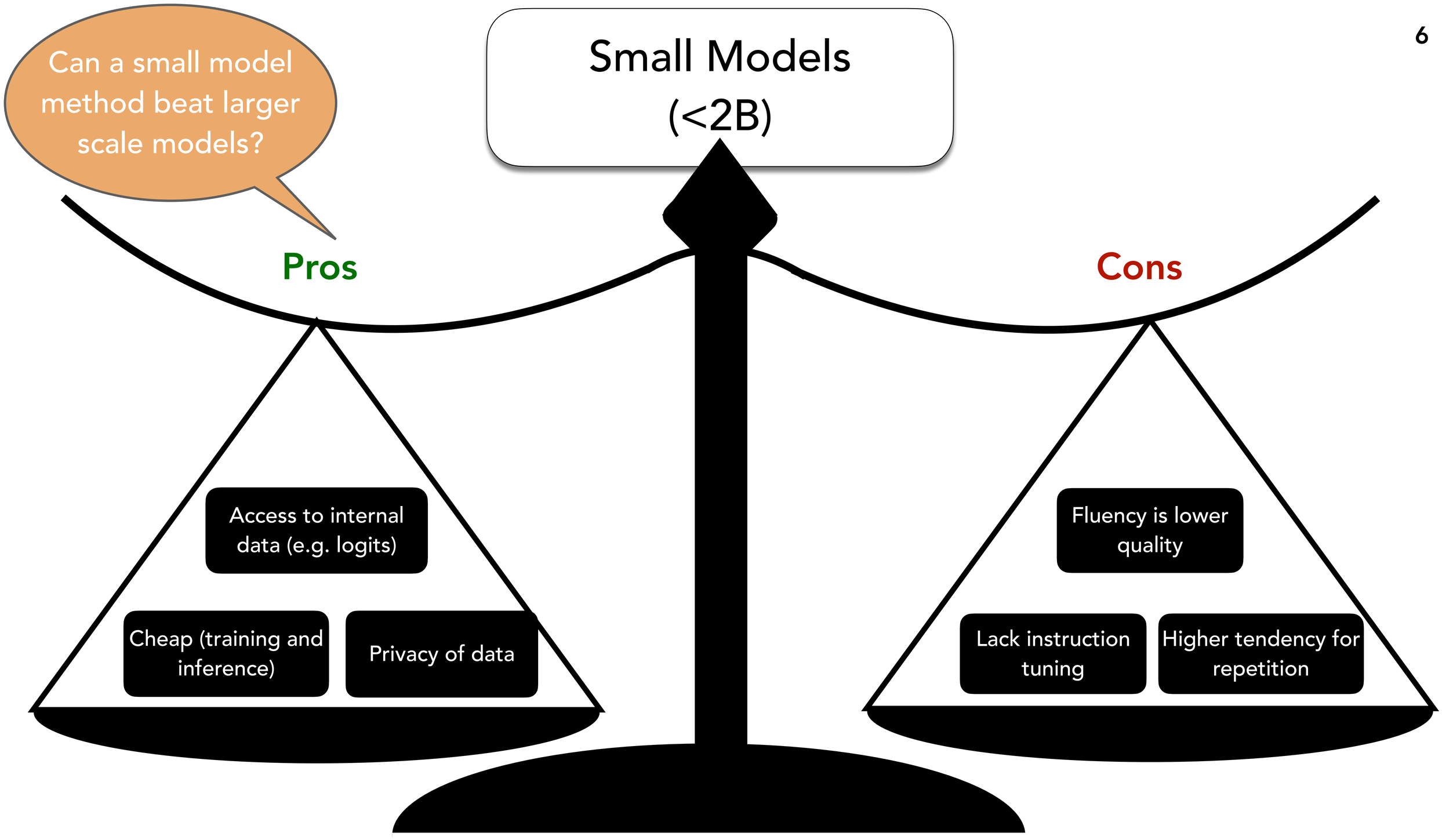
Small Models
(<2B)

Medium Models
(2B - 10B)

Large Models
(>10B)



Model Size


Can a small model method beat larger scale models?
Small Models (<2B)
Pros
Cons
Access to internal data (e.g. logits)
Cheap (training and inference)
Privacy of data
Fluency is lower quality
Lack instruction tuning
Higher tendency for repetition

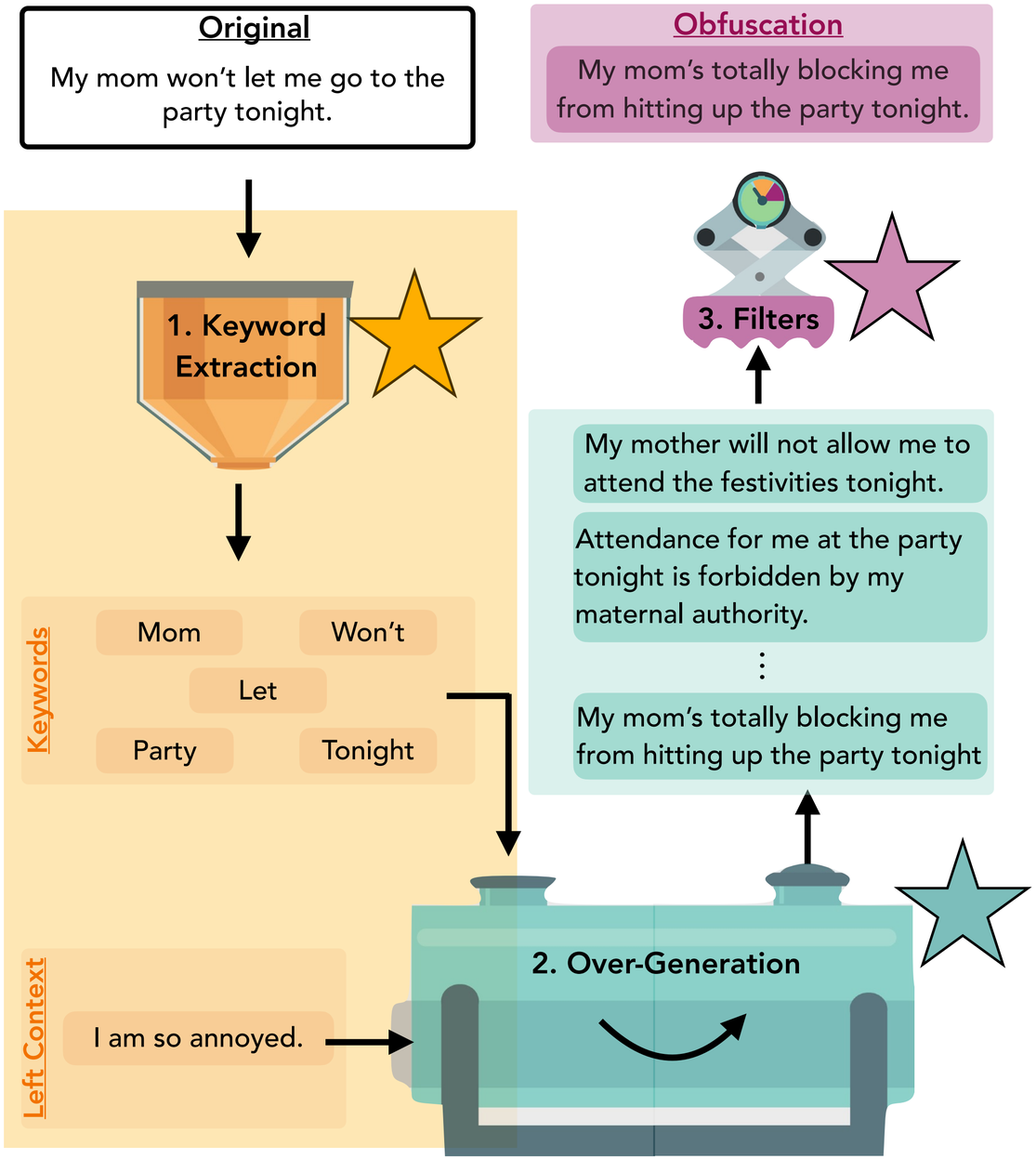
JAMDEC Decoding

Contributions:

- New authorship obfuscation method aimed for small models (JAMDEC Decoding)
- Highlight a new keyword extraction method
- Explore new combined beam-search method

- User-controlled, inference-time algorithm for authorship obfuscation that can be applied to any text and authorship without a separate authorship corpus

- **3 Stage Approach:**
  1. *Keyword Extraction*: Maintain original content
  2. *Over-generation*: Generate diverse outputs that include the keywords
  3. *Filters*: Maintain fluency and content preservation, + user-specified control



Fisher et. al., JAMDEC: Unsupervised authorship obfuscation using con- strained decoding over small language models. (2024)

Innovations: Keyword Extraction

- Current methods rely on cosine similarity of a word-embeddings to a document-embedding

New Likelihood-based Method

- Keywords = top-k tokens with the lowest conditional probabilities, as measured by a specific language model

$$P(t_i \mid \text{context}_i)$$

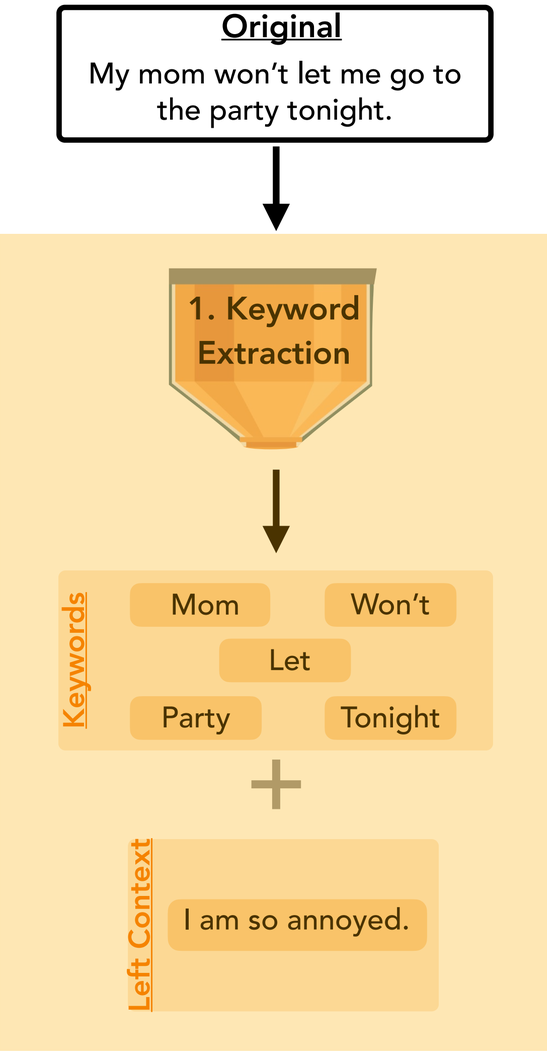
where t_i and context_i is the token and given context at time i.

Auto-Regressive (GPT2)

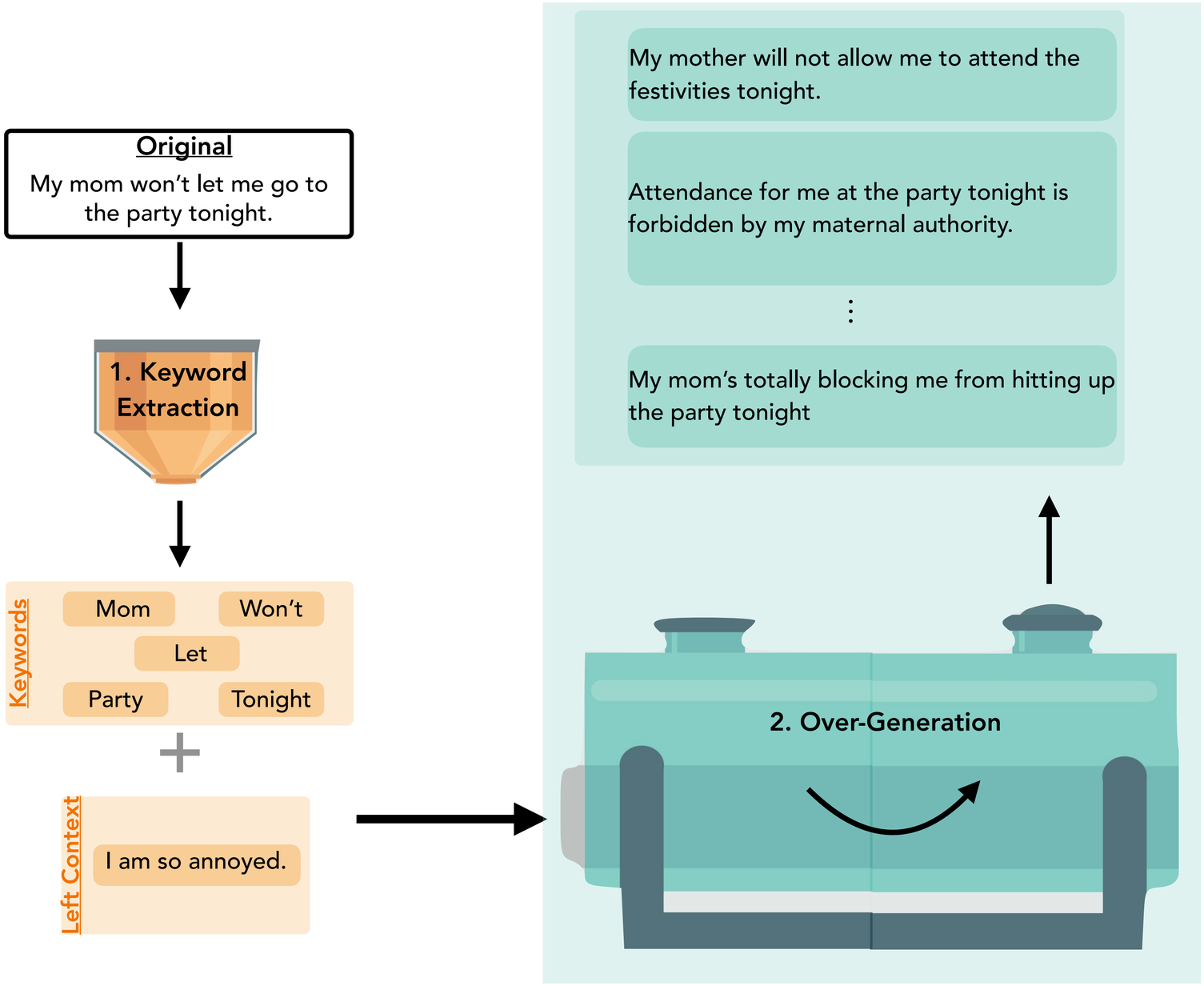
$$P(t_i \mid t_1, t_2, \ldots, t_{i-1})$$

Text-to-Text (T5)

$$P(t_i \mid t_1, \ldots, t_{i-1}, [MASK], t_{i+1}, \ldots, t_n)$$

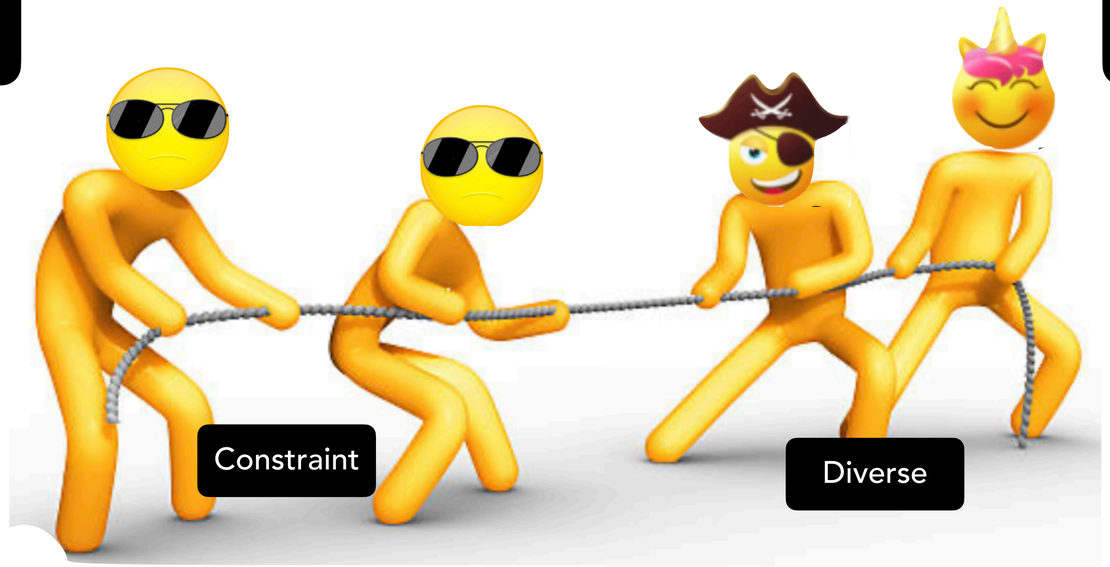


Innovations

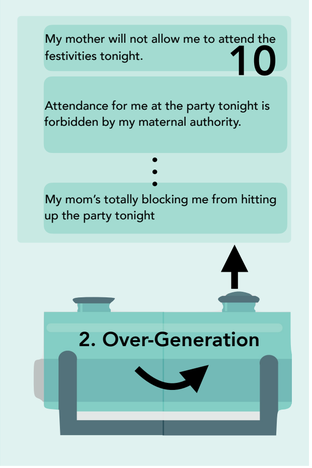


Innovations: Over-Generation

Constrained to original content



Create diverse authorship styles



Constrained + Diverse Beam Search (CoDi-BS)

Constrained + Diverse Beam Search (CoDi-BS)

$$\arg\max_{y \in Y} P_\theta(y \mid x) + \lambda C(y)$$

where x is the sequence of previous tokens, $y \in Y$ is the output sequence, $\theta \in \Omega$ is the parameter vector, λ is a hyperparameter, and $C(y)$ is the constraint penalization.

Add Diversity

$$P^*(y \mid x) = P_\theta(y \mid x) - \lambda F$$

$F \in \mathbb{R}^v$ is a vector of frequencies of tokens chosen in the previous beams, and λ is a hyperparameter

Innovations: Constrained + Diverse Beam Search

$$P^*(y \mid x) = P_\theta(y \mid x) - \lambda F$$

$F \in \mathbb{R}^v$ is a vector of frequencies of tokens chosen in the previous beams, and λ is a hyperparameter

Beam	V1	V2	V3	V4	V5
1	**0.9**	0.8	0.3	0.2	0.7

2	**0.8**	0.5	0.4	**0.8**	0.6

λ − $0.2 \times$

1	0	0	0	0

F

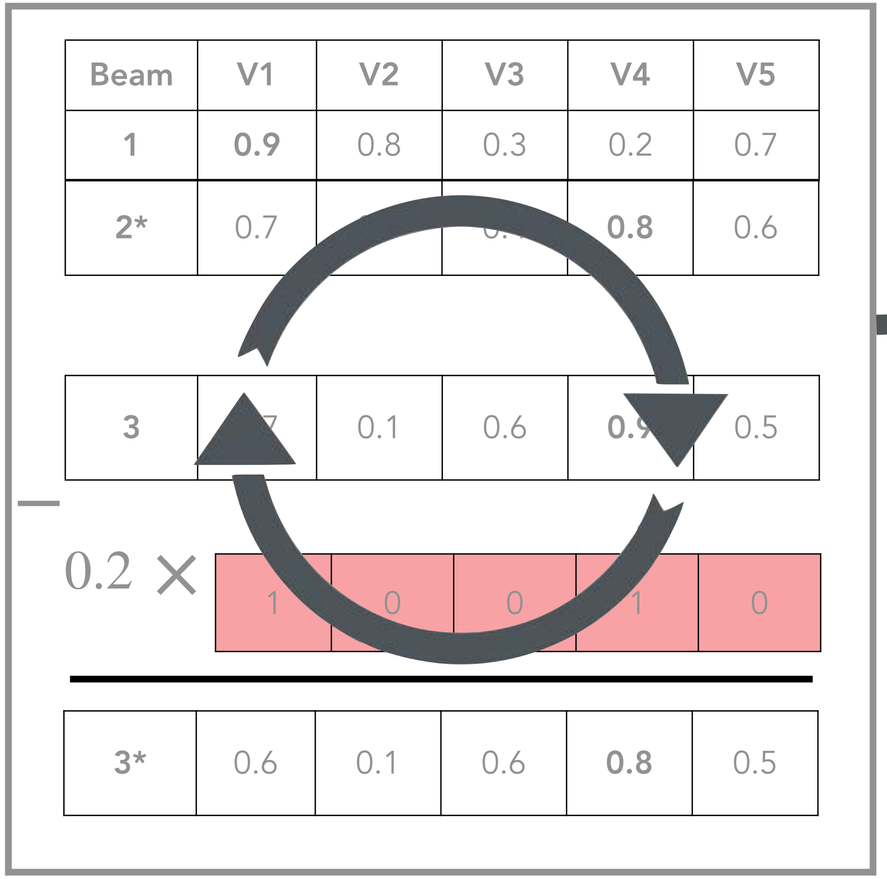
2*	0.7	0.5	0.4	**0.8**	0.6

Replace & Next Beam

Innovations: Constrained + Diverse Beam Search

$$P^*(y \mid x) = P_\theta(y \mid x) - \lambda F$$

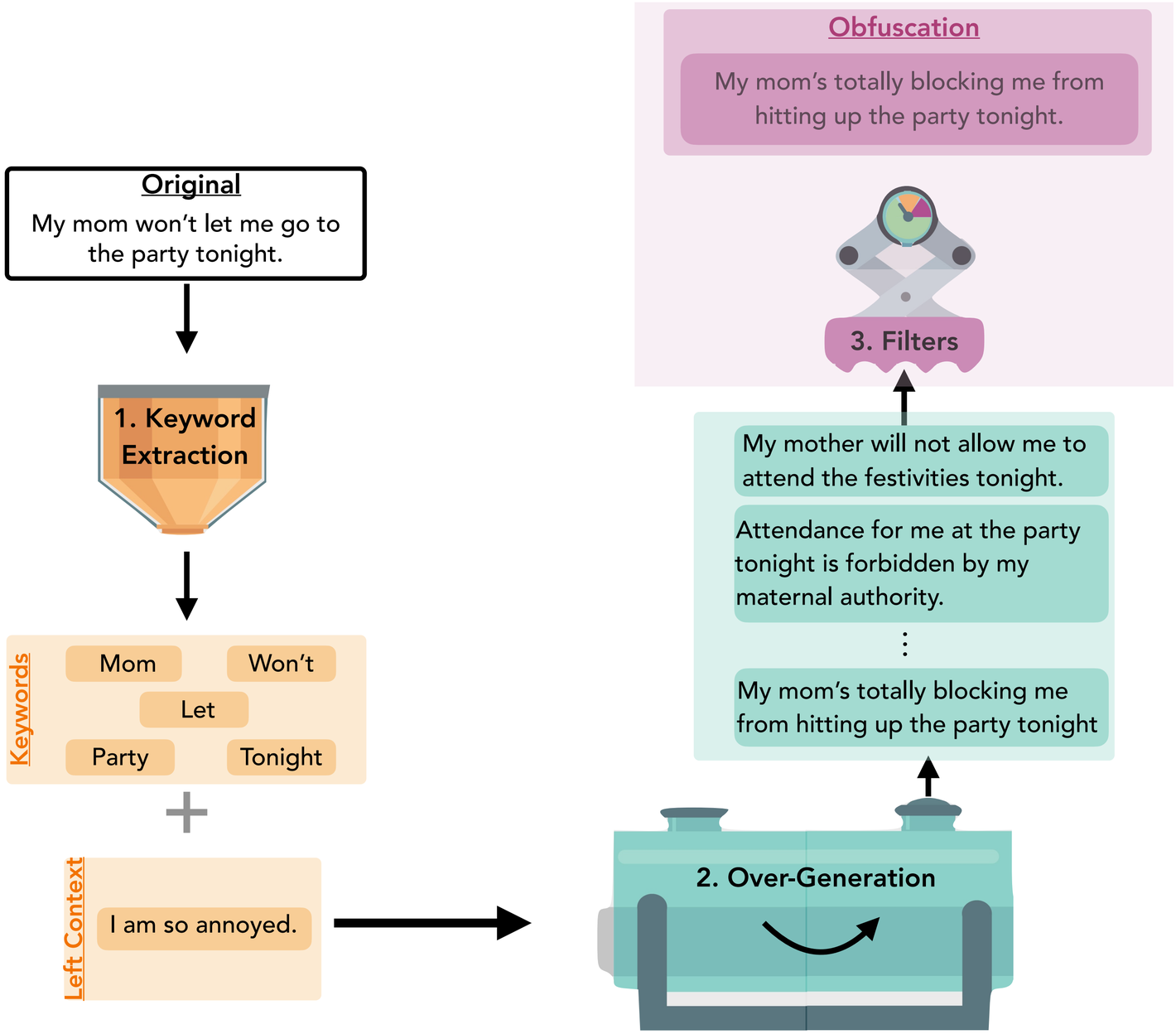
$F \in \mathbb{R}^v$ is a vector of frequencies of tokens chosen in the previous beams, and λ is a hyperparameter



Diversity Processed Logits: $P^*(y \mid x)$

Beam	V1	V2	V3	V4	V5
1	0.9	0.8	0.3	0.2	0.7
2*	0.7	0.5	0.4	0.8	0.6
3*	0.6	0.1	0.6	0.8	0.5
4*	0.3	0.5	0.7	0.4	0.7

Innovations

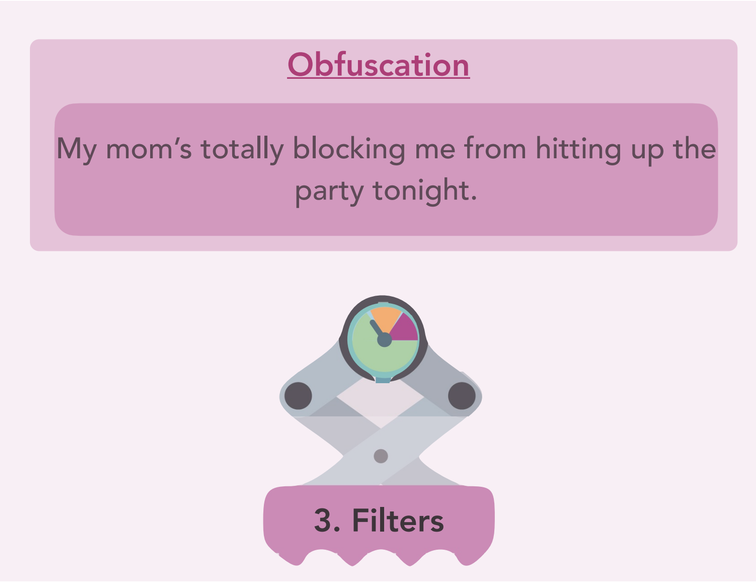


Innovations: Filtering

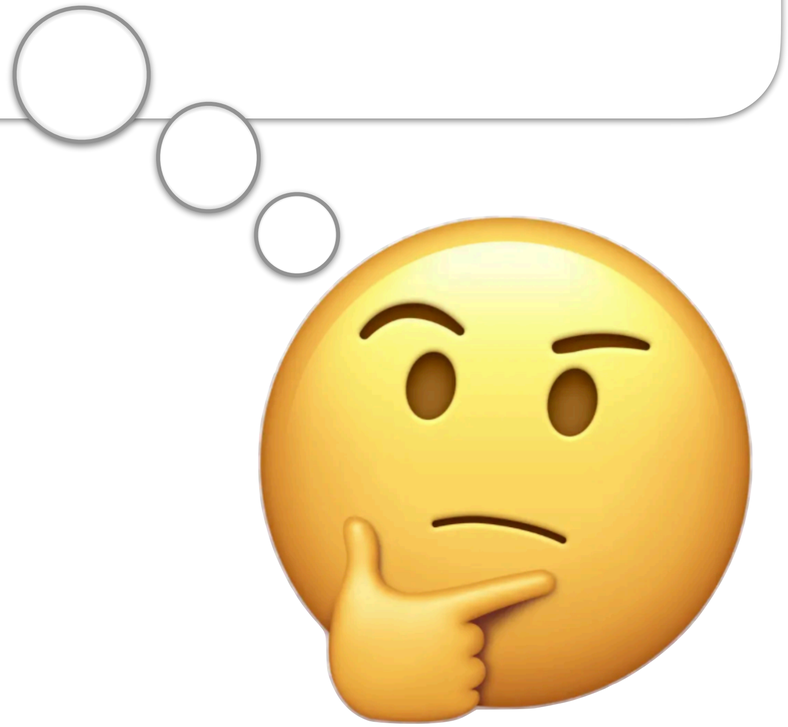
Filtering

- Reduce pool and allow user personalization
- We used the following:
  - Grammar: Corpus of Linguistics Acceptability (CoLA)
  - Content Preservation: Natural Language Inference (NLI)

- Customizable!
  - Length
  - Formality
  - Grade level
  - And more



How does JAMDEC perform compared to other methods?



JAMDEC: Experimental Setup

- **Two Datasets**
  1. Extended-Brennan-Greenstadt: collection of formal scholarly passages
  2. Blog Authorship Corpus: diary-style entries from blog.com
  - Number of Authors: 3,5, or 10

- **Baselines**
  - *Stylometric*: rule-based changes such as synonyms, number of words, punctuation, etc.
  - *Round Trip Machine Translation*: English —> German —> French —> English
  - Paraphrase
  - *Mutant-X*: Iteratively re-writes and combines randomly, uses internal classifier

- **Base Model:** GPT2-XL (1.5B)

JAMDEC: Evaluation Metrics

- Authorship obfuscation is traditionally evaluated (automatically) on:

1. Obfuscation

How well does the rewritten text obfuscate the author style?

Metric: *Drop-Rate* using automatic authorship classifier (ENS and BertAA)

2. Fluency

How understandable is the text?

Metric: *Probability of acceptable grammar* using CoLA model

3. Content Preservation

How similar in meaning is the generation to the original text?

Metric: *Probability of two-way entailment* using NLI model

- Overall Task Score: **average** of the three metrics

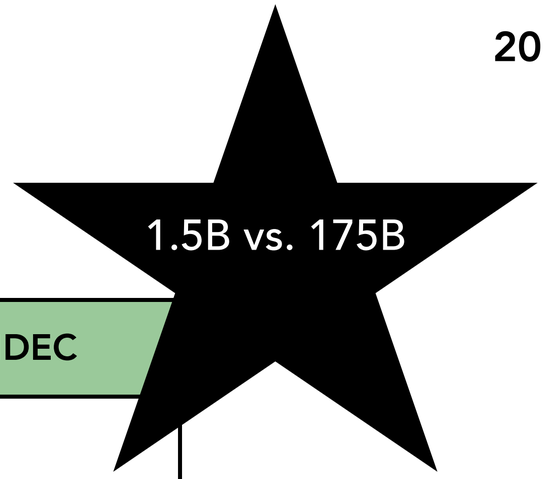
$$\text{Task Score} = \frac{\text{Drop Rate} + \text{NLI} + \text{CoLA}}{3}$$

JAMDEC: Automatic Evaluation

Dataset	Metric	Mutant-X	Paraphrase	Machine	Stylometric	JAMDEC
Scholar - 3	Drop Rate (ENS) ↑	-0.04	0.04	0.04	-0.03	**0.11**
	Drop Rate (BertAA) ↑	0.04	0.04	0.08	**0.12**	0.04
	NLI ↑	0.61	0.62	0.75	0.50	**0.81**
	CoLA ↑	0.51	0.78	0.69	0.46	**0.79**
	Task Score (ENS) ↑	0.36	0.48	0.49	0.31	**0.57**
	Task Score (BertAA) ↑	0.39	0.48	0.51	0.36	**0.55**
Scholar - 5	Drop Rate (ENS)	0.08	0.2	0.2	**0.23**	0.13
	Drop Rate (BertAA)	0	-0.06	0.07	0.04	**0.14**
	NLI	0.57	0.62	0.74	0.48	**0.82**
	CoLA	0.55	0.77	0.69	0.46	**0.79**
	Task Score (ENS)	0.4	0.53	0.54	0.39	**0.58**
	Task Score (BertAA)	0.37	0.44	0.50	0.33	**0.58**
Blog - 10	Drop Rate (ENS)	0.13	**0.35**	0.3	0.21	0.32
	Drop Rate (BertAA)	0.06	0.4	0.11	0.08	**0.32**
	NLI	0.61	0.46	0.62	**0.75**	0.67
	CoLA	0.45	0.62	0.54	0.41	**0.74**
	Task Score (ENS)	0.4	0.48	0.49	0.46	**0.58**
	Task Score (BertAA)	0.37	0.49	0.42	0.41	**0.58**

JAMDEC had the highest overall Task Score on every dataset!

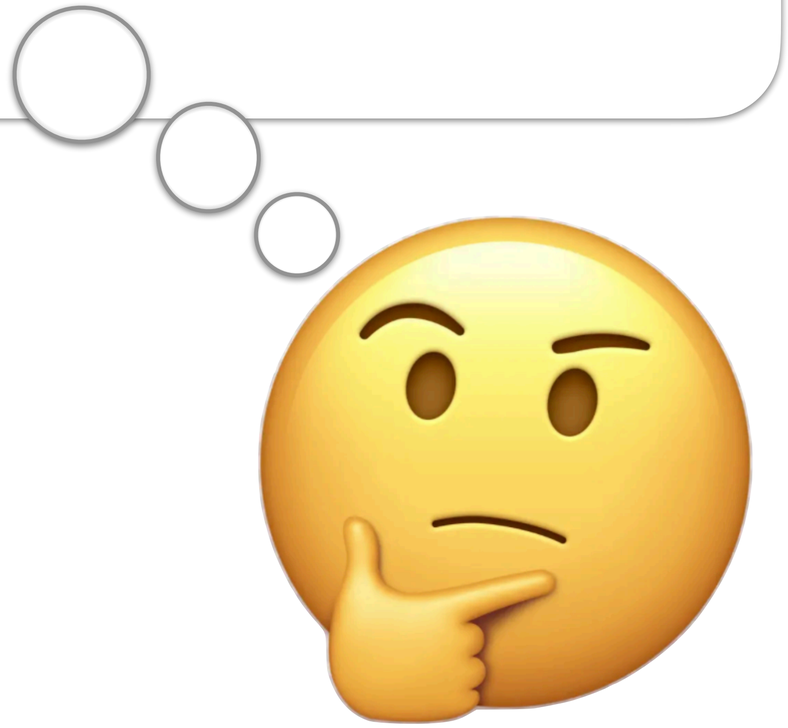
JAMDEC: Automatic Results



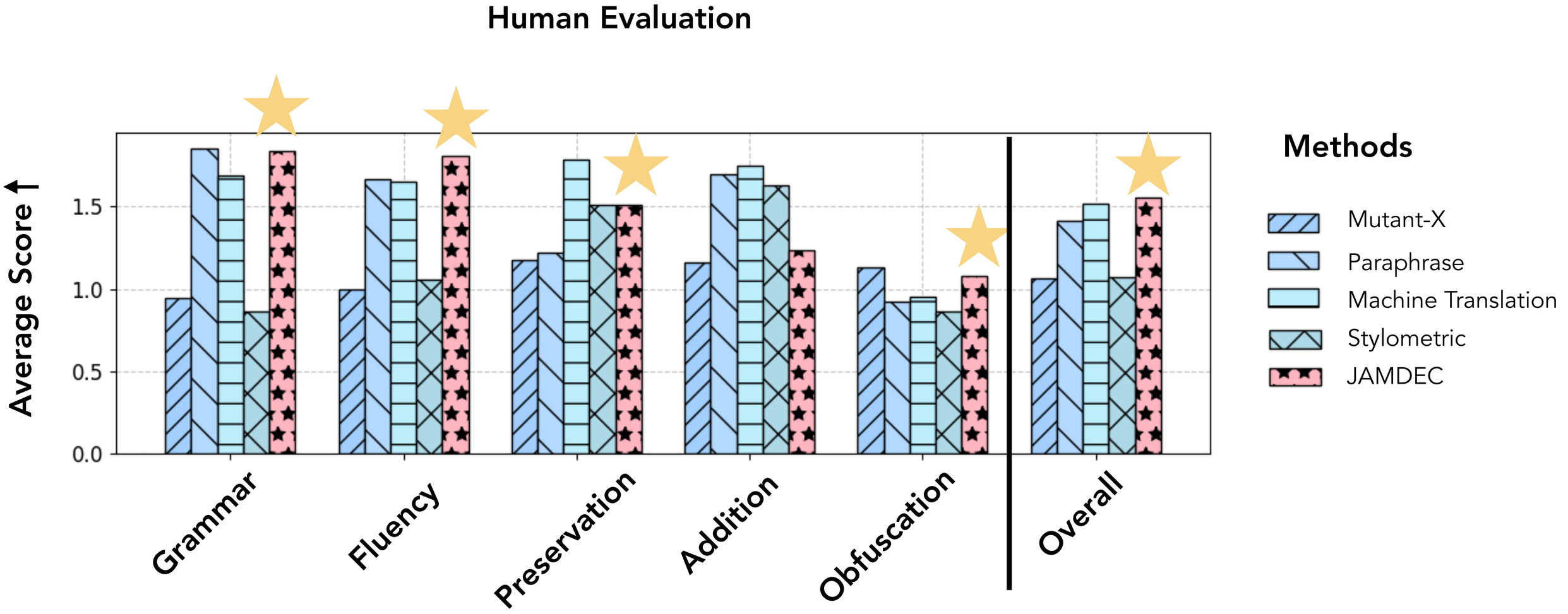
		GPT3-Turbo		JAMDEC
Dataset	**Metric**	**Sentence**	**Paragraph**	
Scholar - 3	Drop Rate (ENS) ↑	**0.23**	**0.23**	0.11
	Drop Rate (BertAA) ↑	**0.13**	0.09	0.04
	NLI ↑	0.77	0.73	**0.81**
	CoLA ↑	0.76	**0.8**	0.79
	Task Score (ENS) ↑	**0.59**	**0.59**	0.57
	Task Score (BertAA) ↑	**0.55**	0.54	**0.55**

Performs similarly to much larger models!

Do humans agree that JAMDEC outperforms other methods?



JAMDEC: Human Evaluation

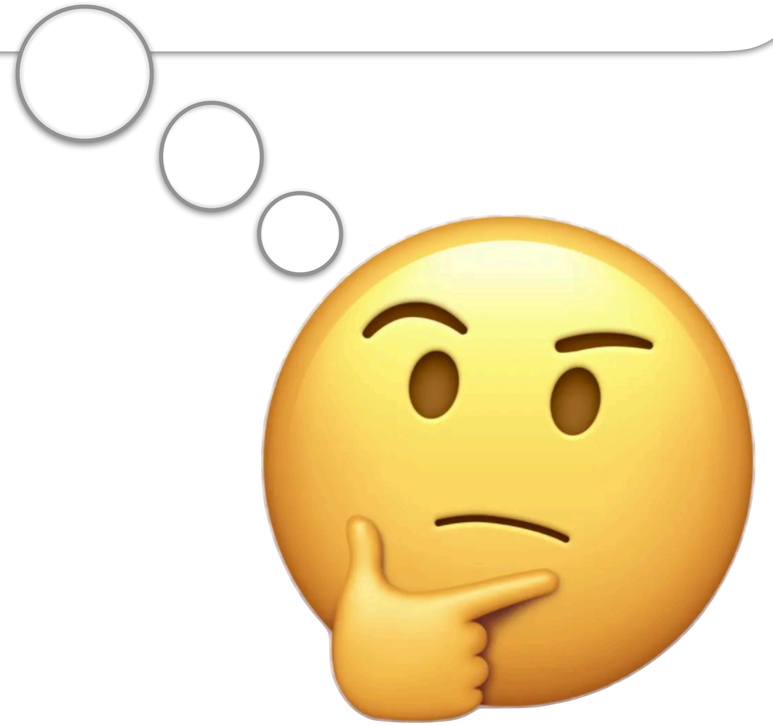


*3-point Likert Scale by 3 raters each

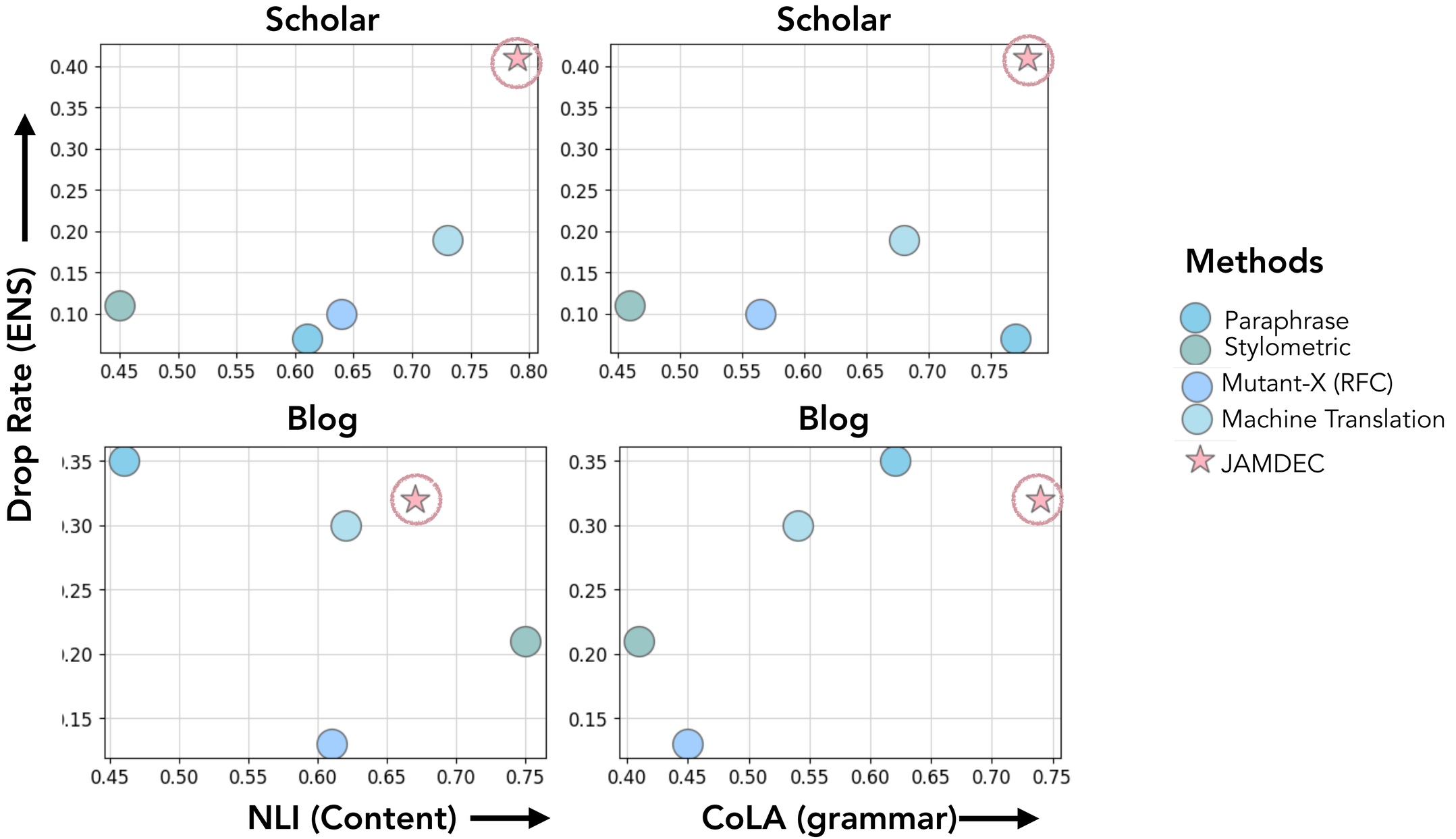
JAMDEC: Qualitative Results

Method	Generation	
Original	The Ex. An ex holding a grudge can do a lot of damage in a short amount of time. He knows enough to open accounts in your name, and he has the motive to hurt you.	
Mutant-X	The Ex. An ex holding a **bitterness able ought** a lot of damage in a **length quantity** of time. He knows enough to **ascend** accounts in **Your prefix**, and he has the **justifiable to impair You.**	Not Grammatical
Paraphrase	**A lot of damage can be done In a short period of time.** He knows **how to** open accounts In your name and he **wants** to hurt you.	Incorrect Content
Machine Translation	**The former.** An **old man who holds a knife** can make a lot of damage in a short time. He knows enough to open accounts in your name, and he has the **reason** to hurt you.	Incorrect Content
Stylometric	An ex **holding, a** grudge can do a lot **inside damage** in a **brief** amount in time, **yet** he knows enough to open accounts in your name, and he has the motive to hurt you.	Missing Meaning
JAMDEC	The Ex. **When the ex is holding his grudge against the person who caused him lot of damage to his life, he is short sighted and will do anything in his power to get back at that person, no matter how much it will hurt the person he is trying to get revenge against.** He knows enough to open accounts in your name, and he has the motive to hurt you.	

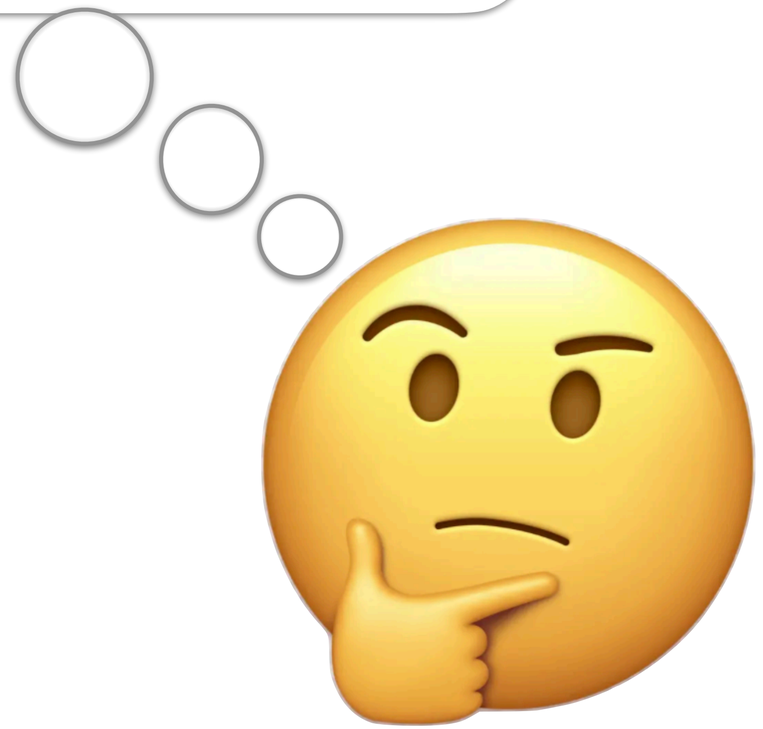
It seems like there might be a tradeoff between obfuscation, content preservation, and fluency...



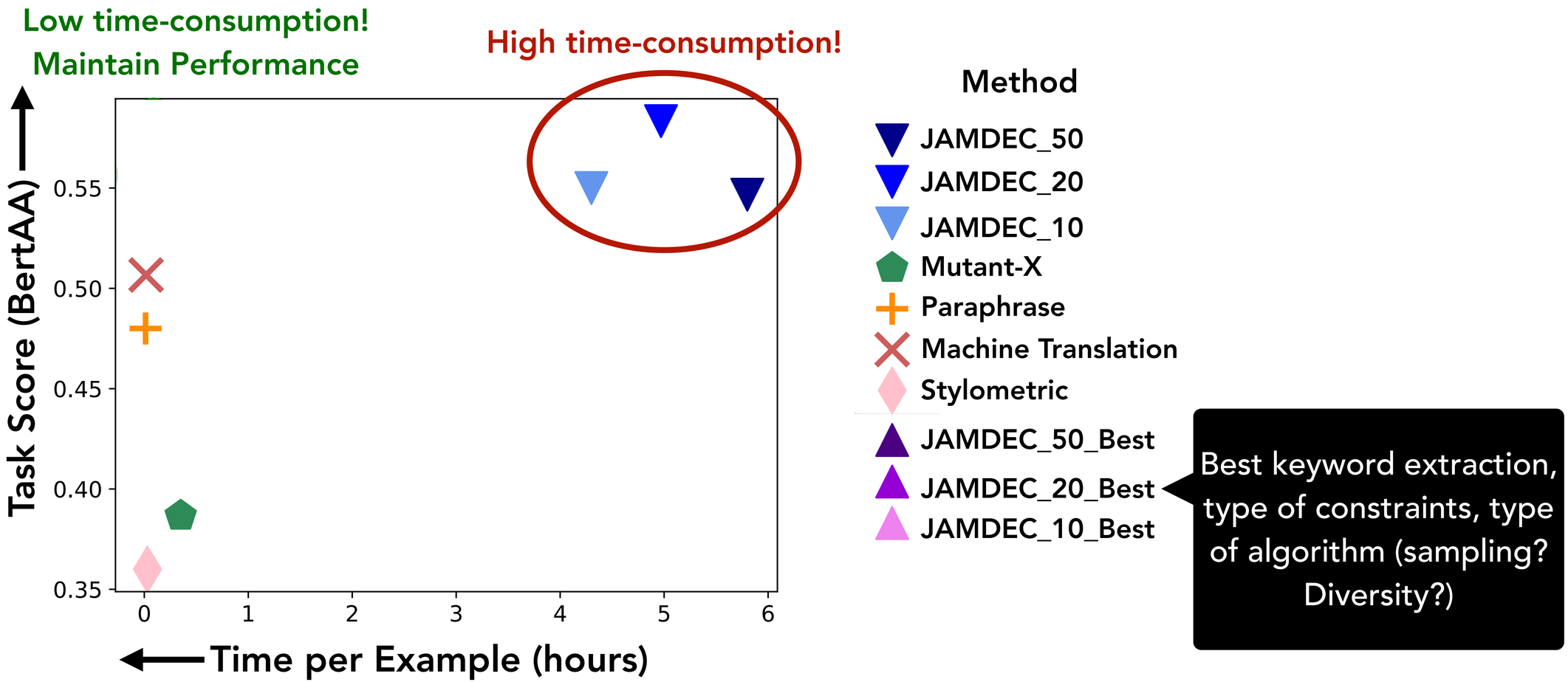
JAMDEC: Inherent Tradeoff



Over-generation seems like it would take more time than other methods…

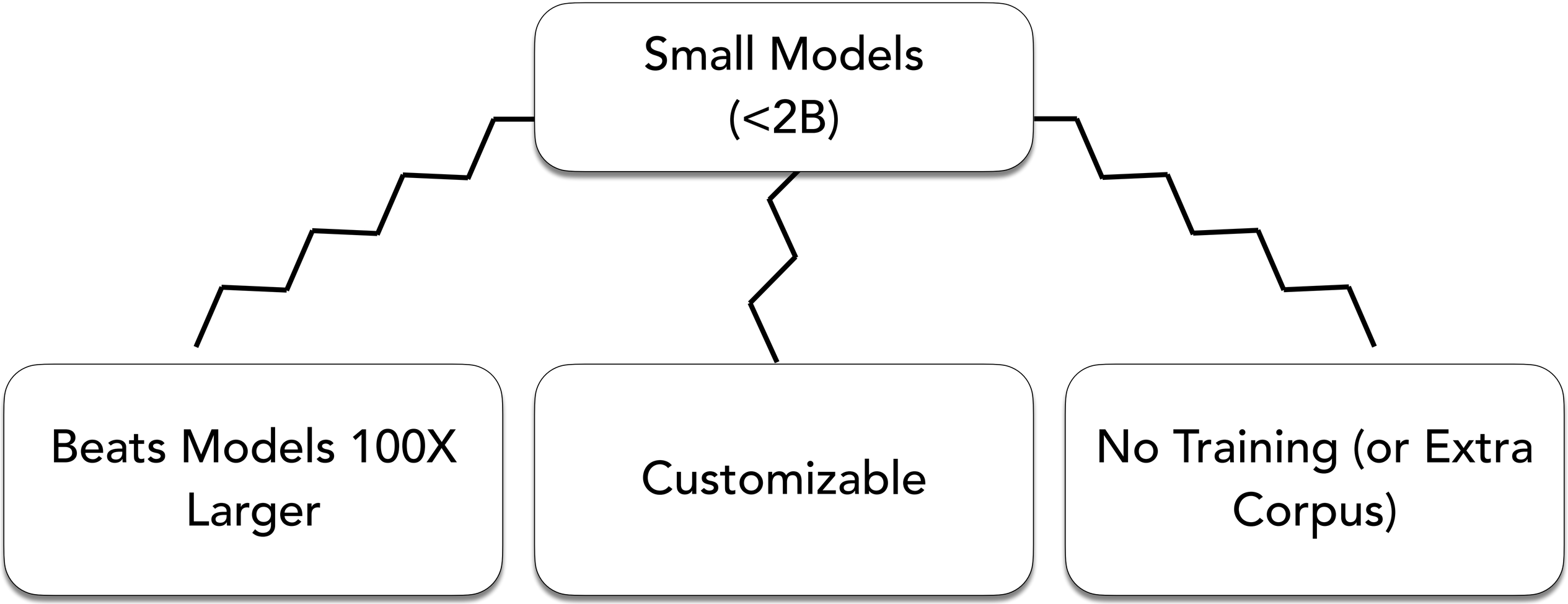


JAMDEC: Computational Time



*Keybert, only the raw constraint (no medium constraints), sampling, ordered, and with diversity

JAMDEC

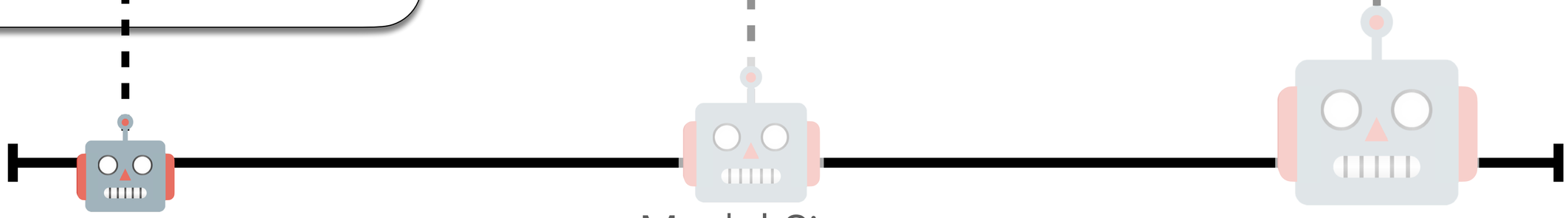

Small Models
(<2B)
Beats Models 100X
Larger
Customizable
No Training (or Extra
Corpus)

Controllable Generation

Small Models
(<2B)

Medium Models
(2B - 10B)

Large Models
(>10B)



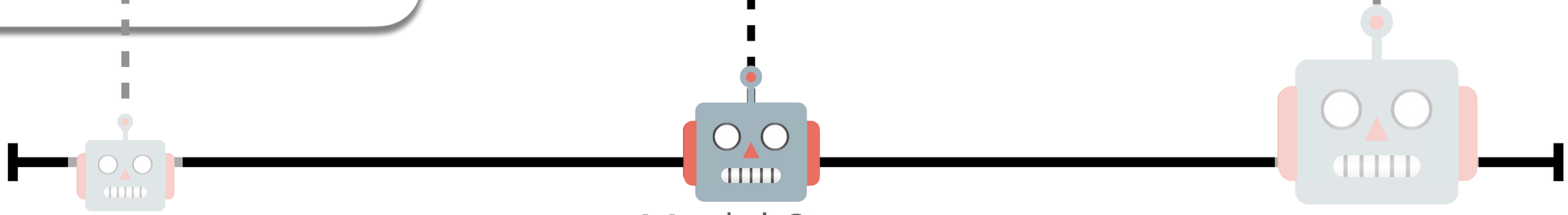
Model Size

Controllable Generation

Open Small Models
(<2B)

Medium Models
(2B - 10B)

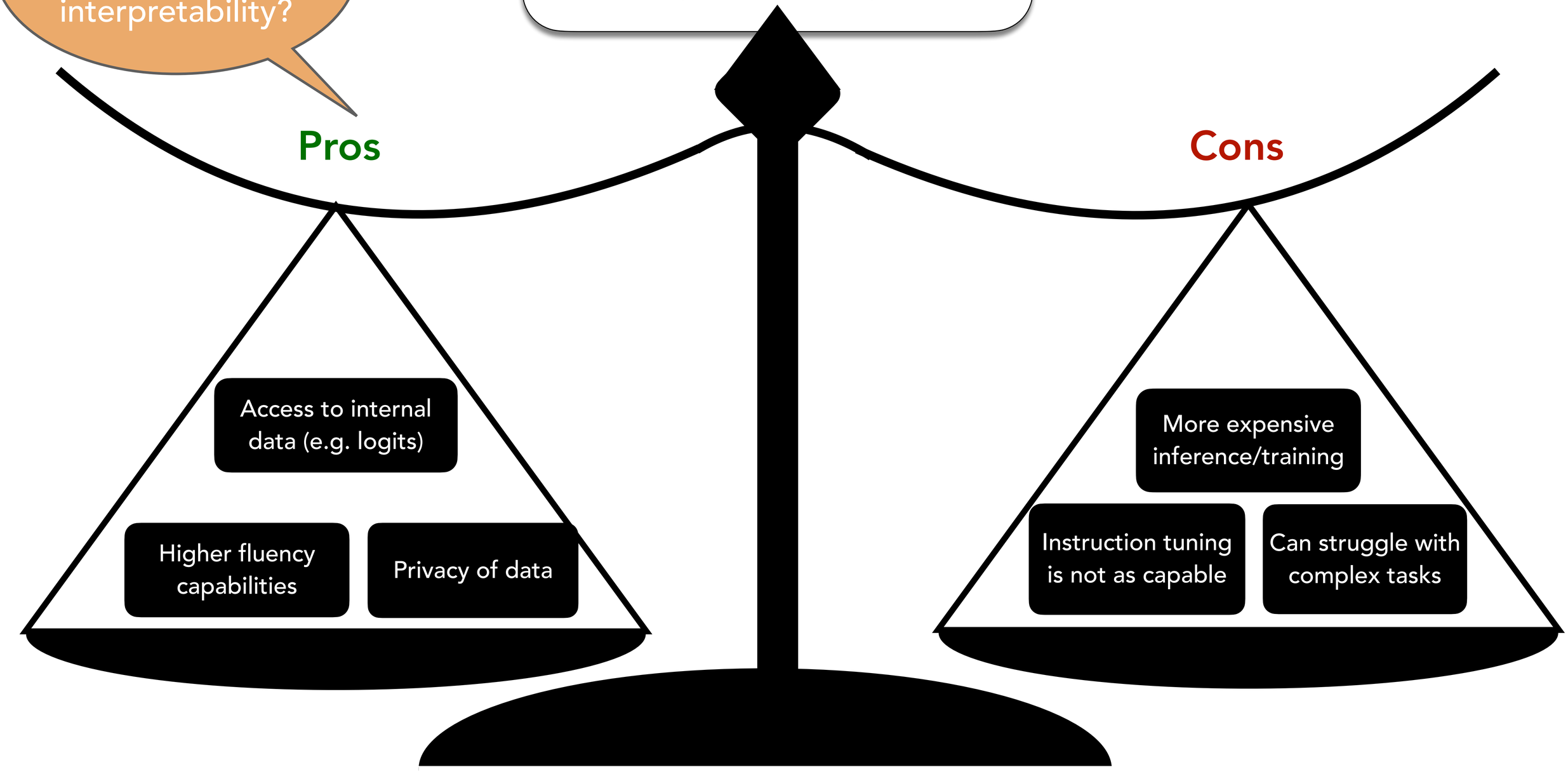
Closed Large Models
(>10B)



Model Size

Medium Models (2B - 10B)

Can we add more control and interpretability?



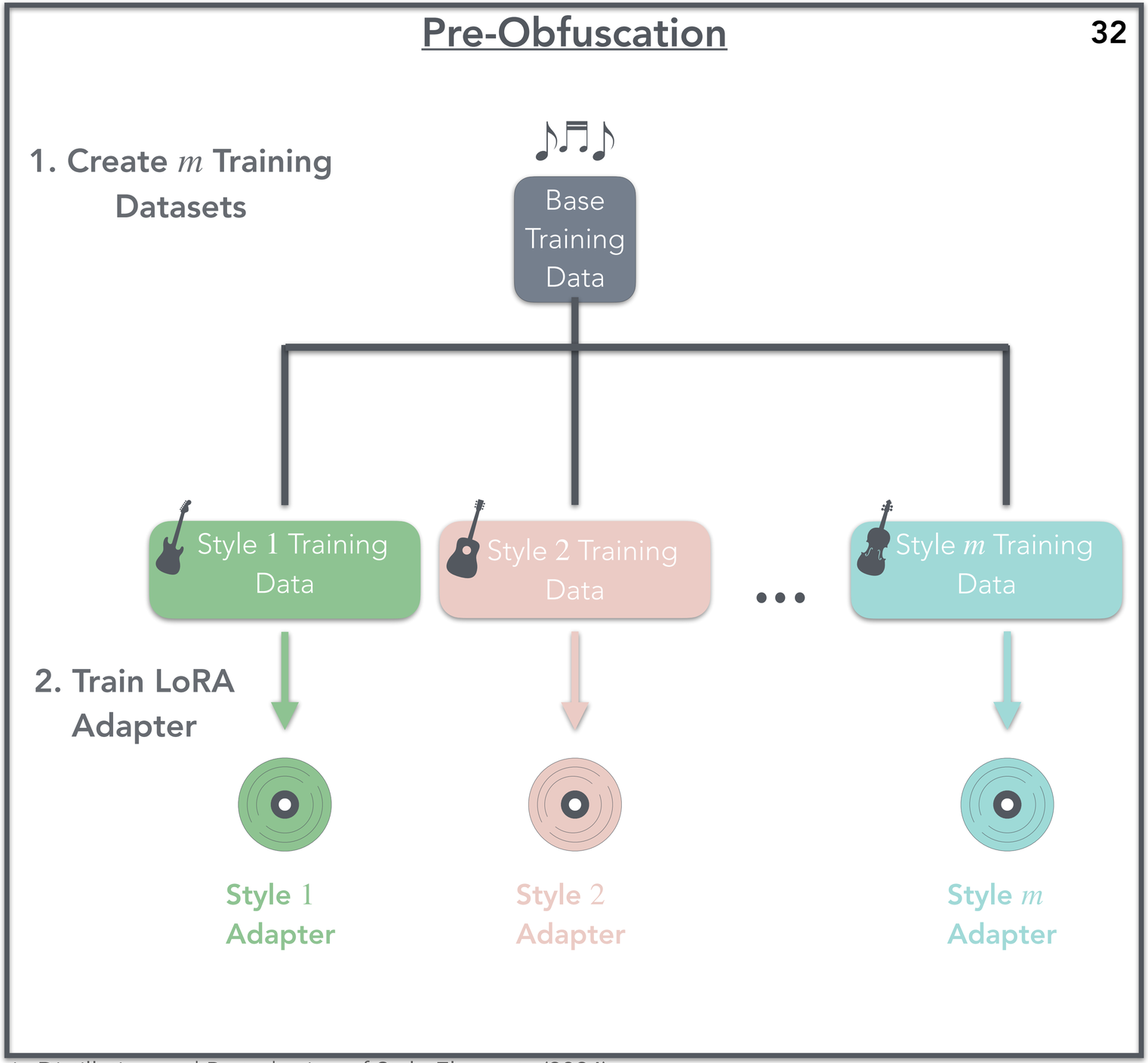
StyleRemix

Contributions:

- New interpretable obfuscation method
- 2 datasets: AUTHORMIX + DiSC
- 16 new style LoRA adapters
- 3 new style classifiers

- An adaptive and interpretable obfuscation method that perturbs specific, fine-grained style elements of the original input text

- **Pre-Obfuscation:**
  1. *Generate Training Data for each m style*
  2. *Train Low-Rank Adapters (LoRA Adapter)*



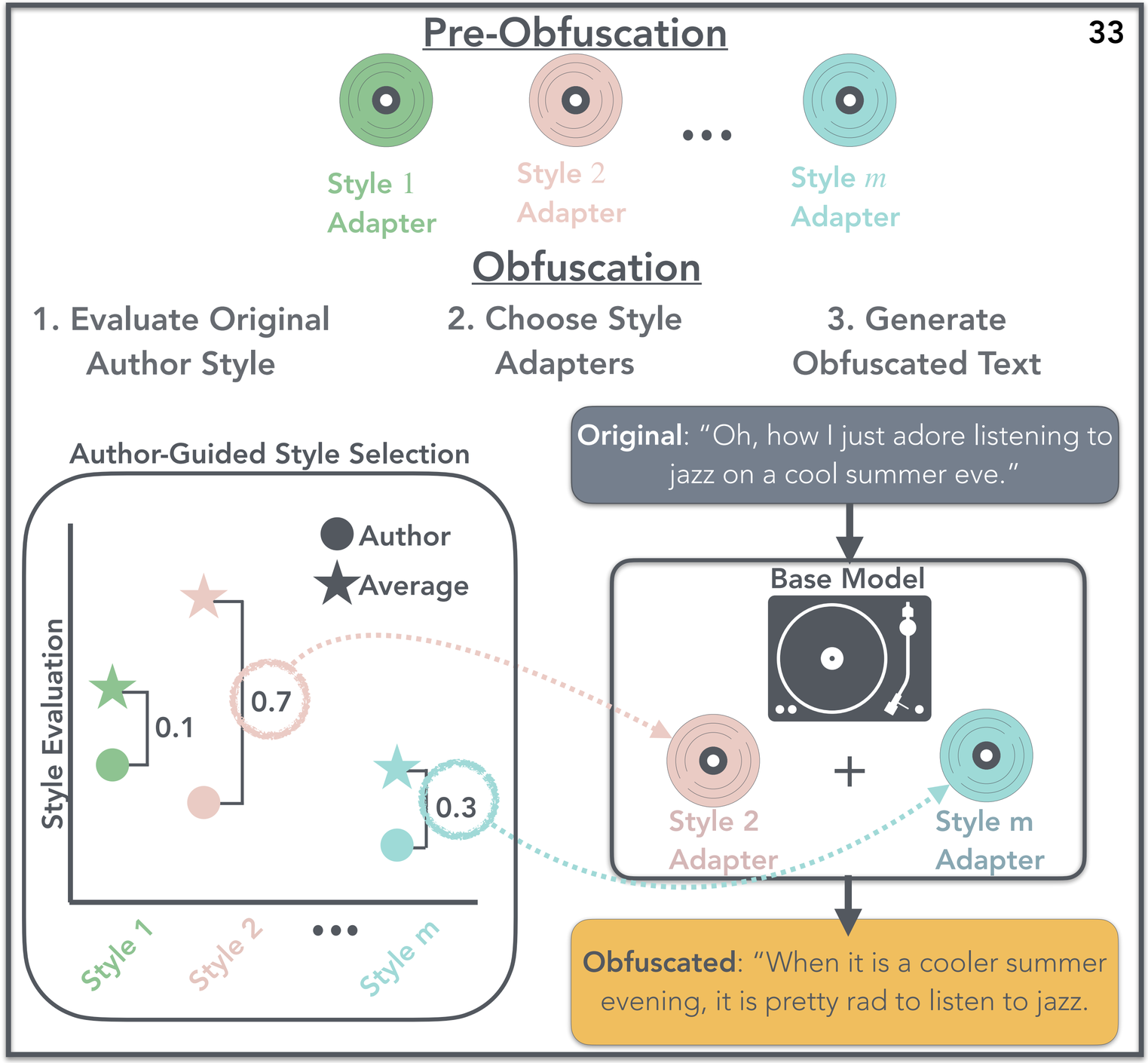
StyleRemix

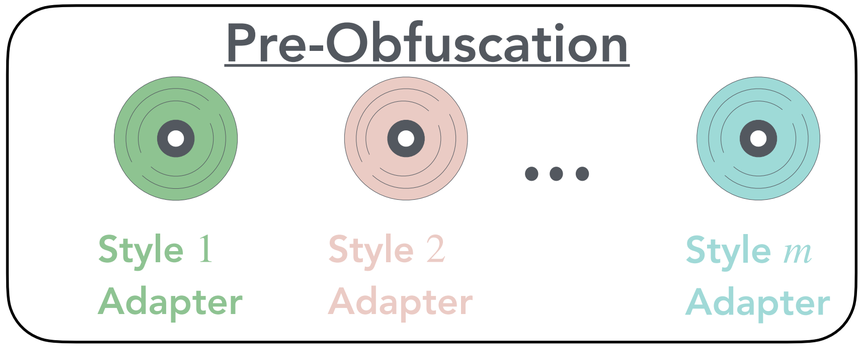
Contributions:

- New interpretable obfuscation method
- 2 datasets: AUTHORMIX + DiSC
- 16 new style LoRA adapters
- 3 new style classifiers

- An adaptive and interpretable obfuscation method that perturbs specific, fine-grained style elements of the original input text

- **Obfuscation**
  1. *Evaluate Original Author Style*
  2. *Choose Style Adapters*
  3. *Generate Obfuscated Text*

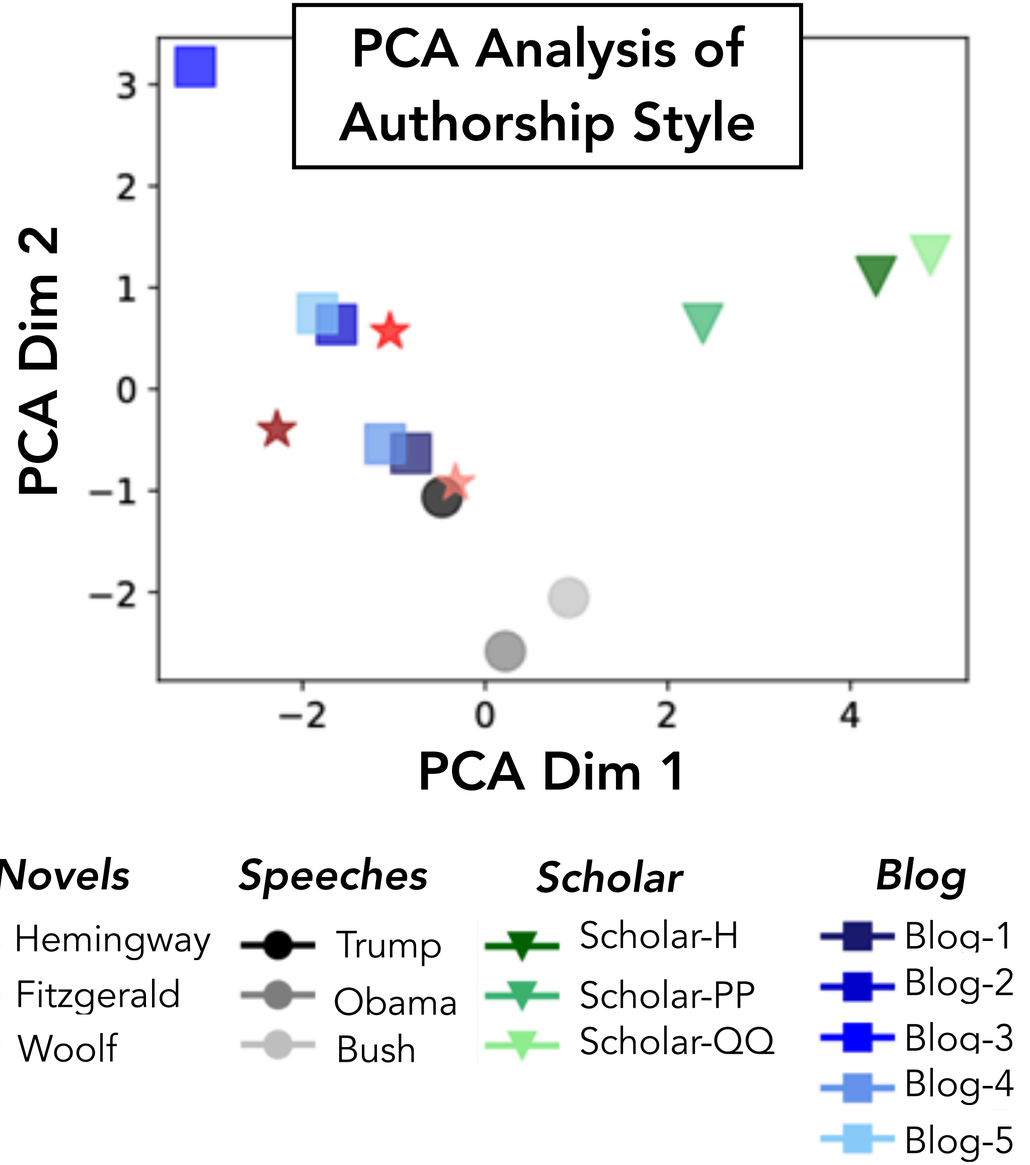




Which style axes should we use?

Writing Intent = Writing Style: Persuasion, Descriptive, Narrative, Expository

Do these styles differentiate authors?



Pre-Obfuscation: Adapter Training Set

Style Axes

- Length
- Function Words
- Grade Level
- Formality
- Sarcasm
- Voice
- Writing Intent

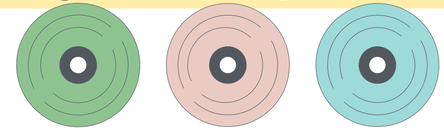
Base Training Dataset

- Wikipedia
- Books + Plays
- Blog

Distilled Style Components Dataset (DiSC)

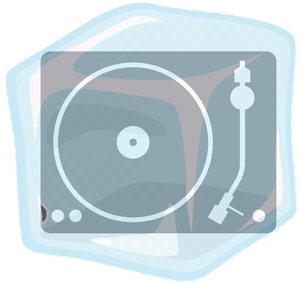
- A set of web, book, and blog texts rewritten towards **16 distinct style** directions across seven style axes

→ **Used to train style adapters!**



Pre-Obfuscation: Train LoRA Adapter

Pre-trained Model



$W_0 \in \mathbb{R}^{d \times k}$

Update

Expensive!

$\Delta W \in \mathbb{R}^{d \times k}$

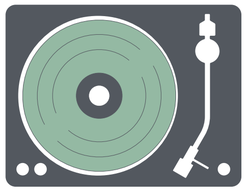
$A \in \mathbb{R}^{d \times r}$

$B \in \mathbb{R}^{r \times k}$

$\text{rank}(r) \ll \min(d, k)$

Low-Rank Adaptation

Updated Model



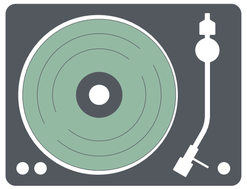
~~$W_0 x + \Delta W x$~~

$W_0 x + ABx$

Hu et. Al., LoRA: Low-Rank Adaptation of Large Language Models. 2021

Pre-Obfuscation: Train LoRA Adapter

Updated Model



One for each Style Axis

Original (Trump)
And I will now sign the executive action. And then, if you want, we can ask a few questions.

Grade Level: Higher
I am currently initiating my actions by formally aligning them. Subsequently, it might be an option for us, aligning to your preferences, to engage and might find ourselves addressing various issues by effectively initiating our various activities.

Length: Higher
And at this point, I intend to execute the executive order. Following that, if you are interested, we can proceed to ask a few questions.

Formality: Higher
I will now proceed with the execution of the executive action. Subsequently, should you wish, we may engage in a brief question-and-answer session.

Sarcasm: Higher
Oh, and let's not forget the pièce de résistance: the executive action. After that, if you're so inclined, we can move on to the fun part—questions.

Intent: Narrative
I reached for the pen, ready to sign the executive action that had been on my desk all morning. "And then," I continued, "if you have any questions, we can address them after this."

Function Words: Higher
And I will now sign the executive action. And then, if you wish, we can ask a few questions.

Voice: Passive
The executive action will now be signed by me. And then, if you wish, a few questions can be asked by us.

Pre-Obfuscation: Train LoRA Adapter

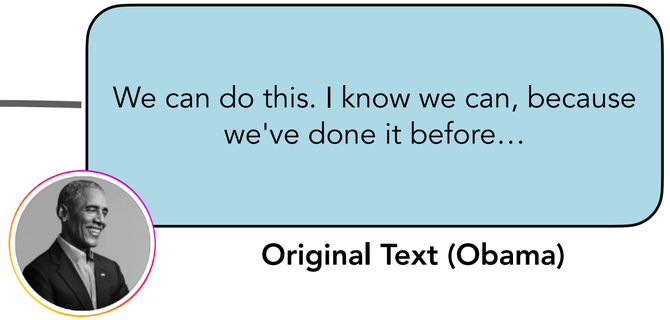
Style Axis *(metric)*	**Original**	**More**	**Less**
Length *(words/sent)*	18.87	**23.04**	18.24
Function Words *(# func. words)*	40.08	**55.19**	21.47
Grade Level *(avg. of 3)*	9.45	**11.08**	6.72
Formality *(model score)*	0.68	**0.97**	0.43
	Accuracy *(human evaluation)*		
Sarcasm	97.7		
Voice	93.7		
Writing Intent *(4 classes)*	77.7		

How do we select the LoRA adapters?

We can do this. I know we can, because we've done it before…

Original Text (Obama)

Obfuscation: Select Style Axes

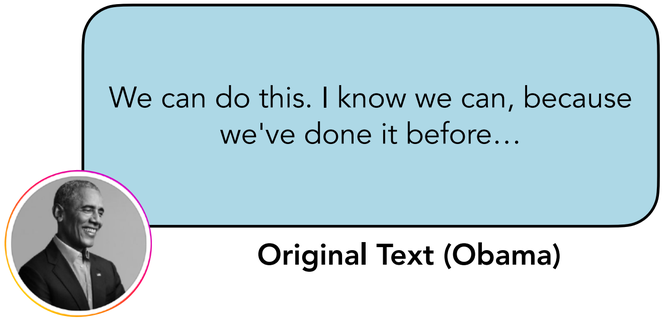


1. Evaluate Author Style

2. Choose k Style Axes (and directions)

Metric	Length	Function Words	Grade Level	Formality	Sarcasm	Voice	Intent*
Obama	0.8	0.4	0.6	0.8	0.4	0.2	0.5
Average	0.6	0.7	0.4	0.3	0.4	0.3	0.5
Diff.	0.2	-0.3	0.2	0.5	0.0	-0.1	0.0
		(Higher)		(Lower)			

Obfuscation: Select Style Axes Weights



Original Text (Obama)

1. Evaluate Author Style

2. Choose k Style Axis (and directions)

Function Words	Formality
(Higher)	(Lower)

→

3. Choose Weights of Style Axes

3.a) Static Weight Selection

of std. from the average: $\underline{\text{std}(\bar{x}_i)}$

$$w_i = \begin{cases} 0.7, & \text{if } \text{std}(\bar{x}_i) \leq 1 \\ 0.9, & \text{if } 1 < \text{std}(\bar{x}_i) \leq 2 \\ 1.2, & \text{if } 2 < \text{std}(\bar{x}_i) \leq 3 \\ 1.5, & \text{if } \text{std}(\bar{x}_i) > 3 \end{cases}$$

3.b) Dynamic Weight Selection

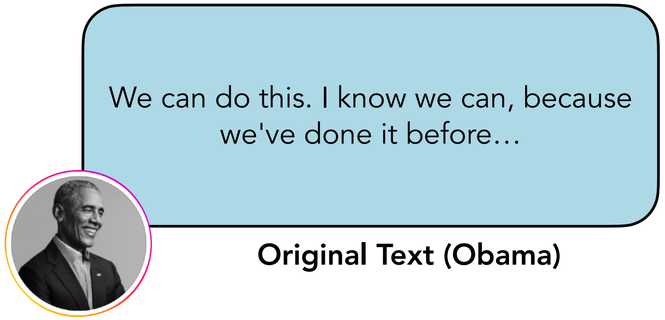
Optimization of loss based on style axes evaluations

$$L = \sum_{v_i \in \{v_1, v_2\}} \begin{cases} v_i, & \text{if higher} \\ 1 - v_i, & \text{if lower} \end{cases} + \alpha \cdot f$$

v_i = Average style score on test set

f = Fluency score

Obfuscation: Select Style Axes Merging



1. Evaluate Author Style

2. Choose k Style Axis (and directions)

Function Words (Higher)

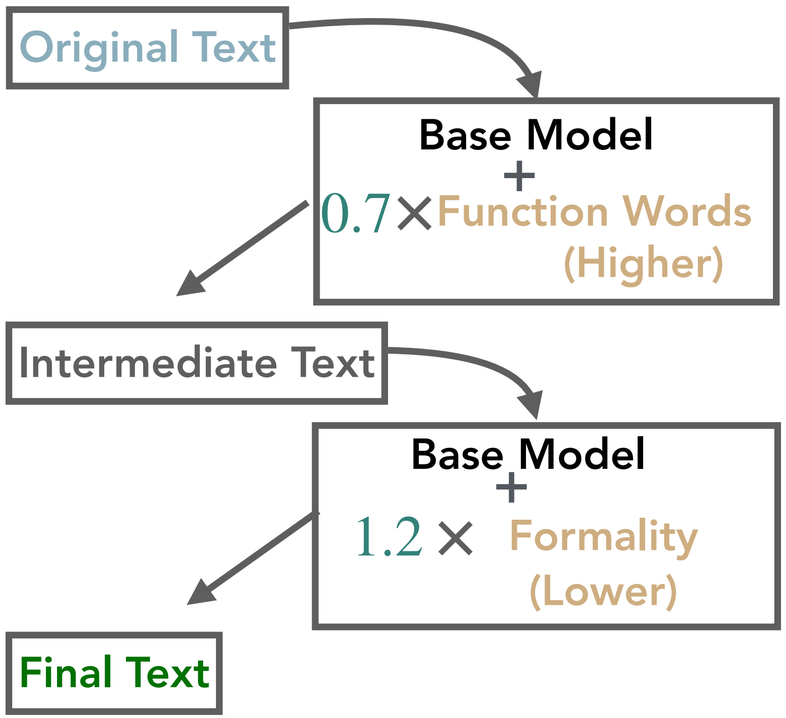
Formality (Lower)

3. Choose Weights of Style Axes

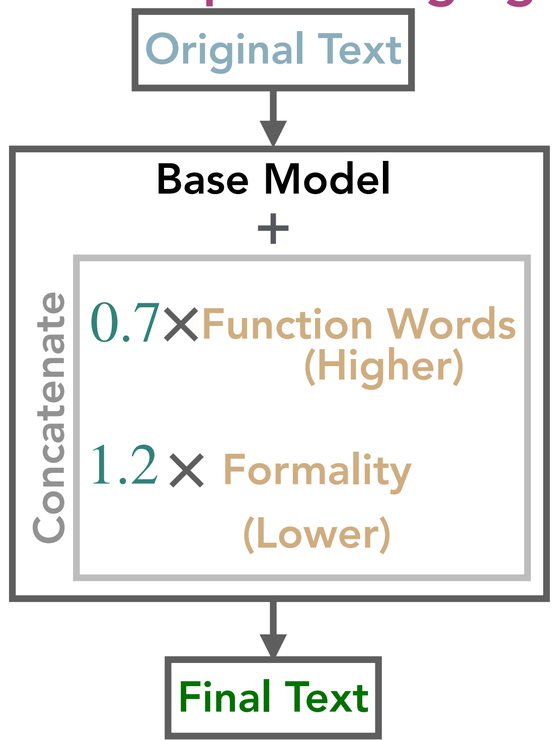
$0.7, 1.2$

4. Combine Style Adapters

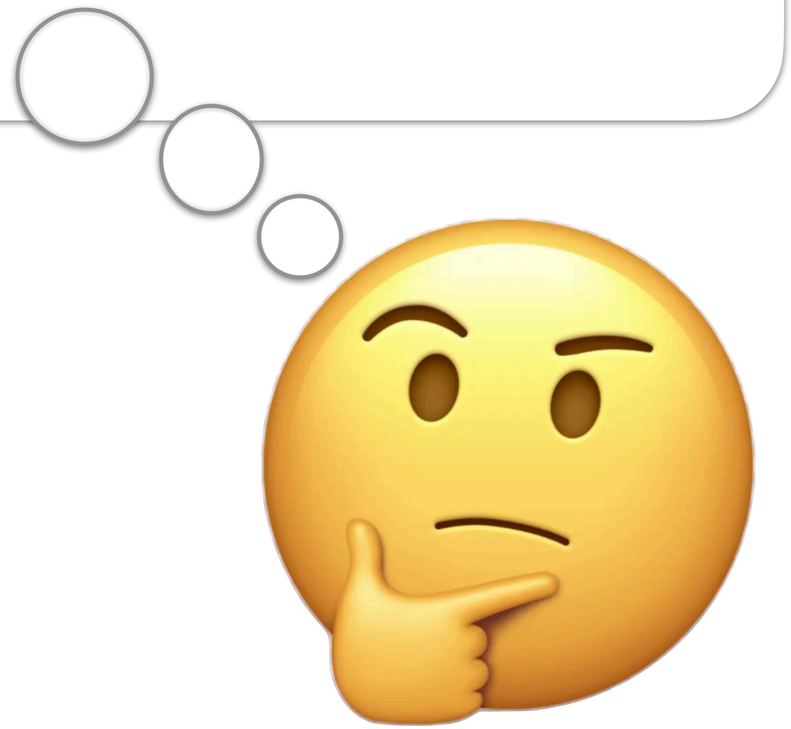
4.a Sequential



4.a Adapter Merging



How does StyleRemix perform compared to other methods?



StyleRemix: Experimental Setup




- **Four Datasets (AUTHORMIX)**
  1. Extended-Brennan-Greenstadt: collection of formal scholarly passages
  2. Blog Authorship Corpus: diary-style entries from blog.com
  3. Presidential Speeches: transcript of presidential speeches (Trump, Obama, Bush)
  4. Novels: 1900s fiction writers (Fitzgerald, Woolf, Hemingway)

  More distinct styles

  - Number of Authors: 3 or 5
- **Baselines**
  - *Stylometric*: rule-based changes such as synonyms, number of words, punctuation, etc.
  - *Round Trip Machine Translation*: English —> German —> French —> English
  - *Mutant-X*: Iteratively re-writes and combines randomly, uses internal classifier
  - Paraphrase
  - JAMDEC
  - Collection of LLM's

- **Base Model:** Llama-3 (8B)

StyleRemix: Evaluation Metrics

- Authorship obfuscation traditionally evaluated (automatically) on:

1. Obfuscation

How well does the rewritten text obfuscate the author style?

Metric: *Drop-Rate* using automatic authorship classifier (RoBERTa large)

2. Fluency

How understandable is the text?

Metric: *Probability of acceptable grammar* using CoLA model

3. Content Preservation

How similar in meaning is the generation to the original text?

Metric: *Cosine similarity of word embeddings*

Better for longer text

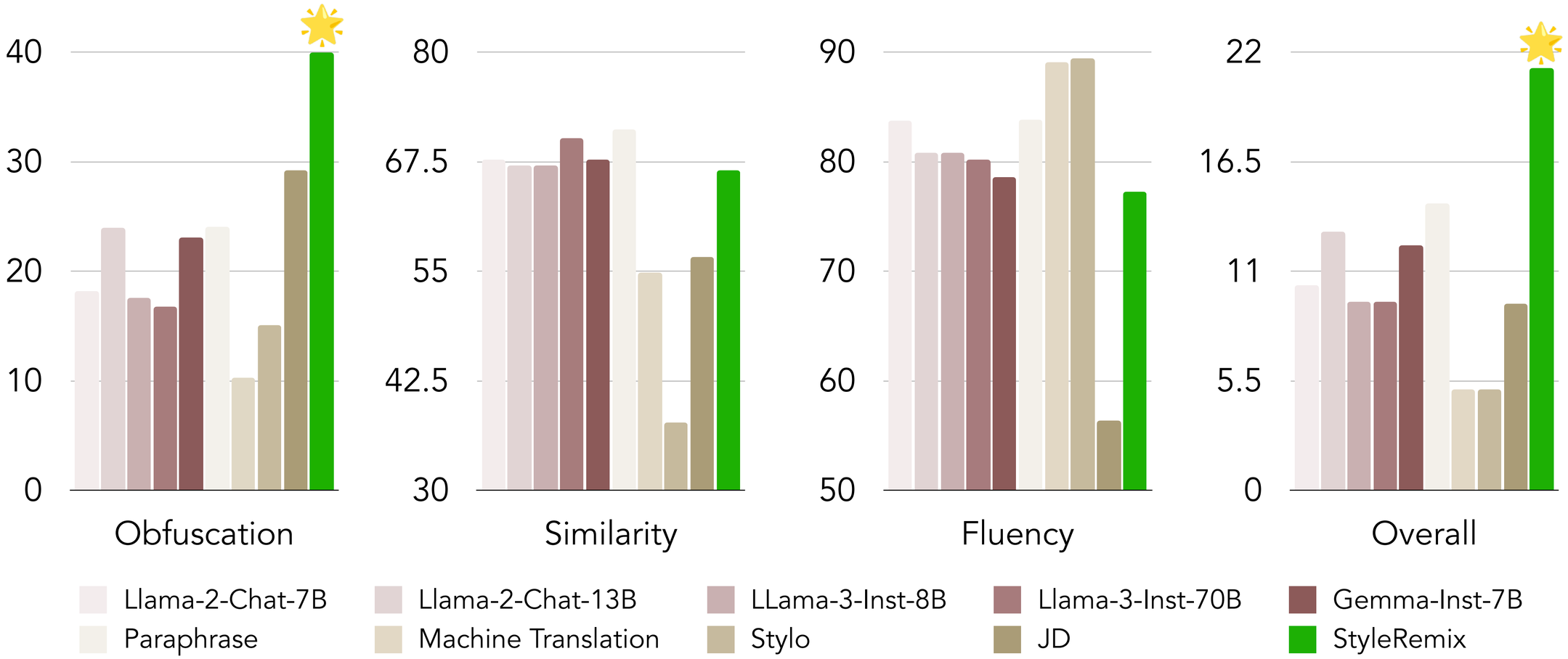
- Overall Task Score: **average** of the three metrics

$$\text{Task Score} = \frac{\text{Drop Rate} + \text{NLI} + \text{CoLA}}{3}$$

StyleRemix: Automatic Evaluation

StyleRemix outperforms all baselines in obfuscation and overall quality!

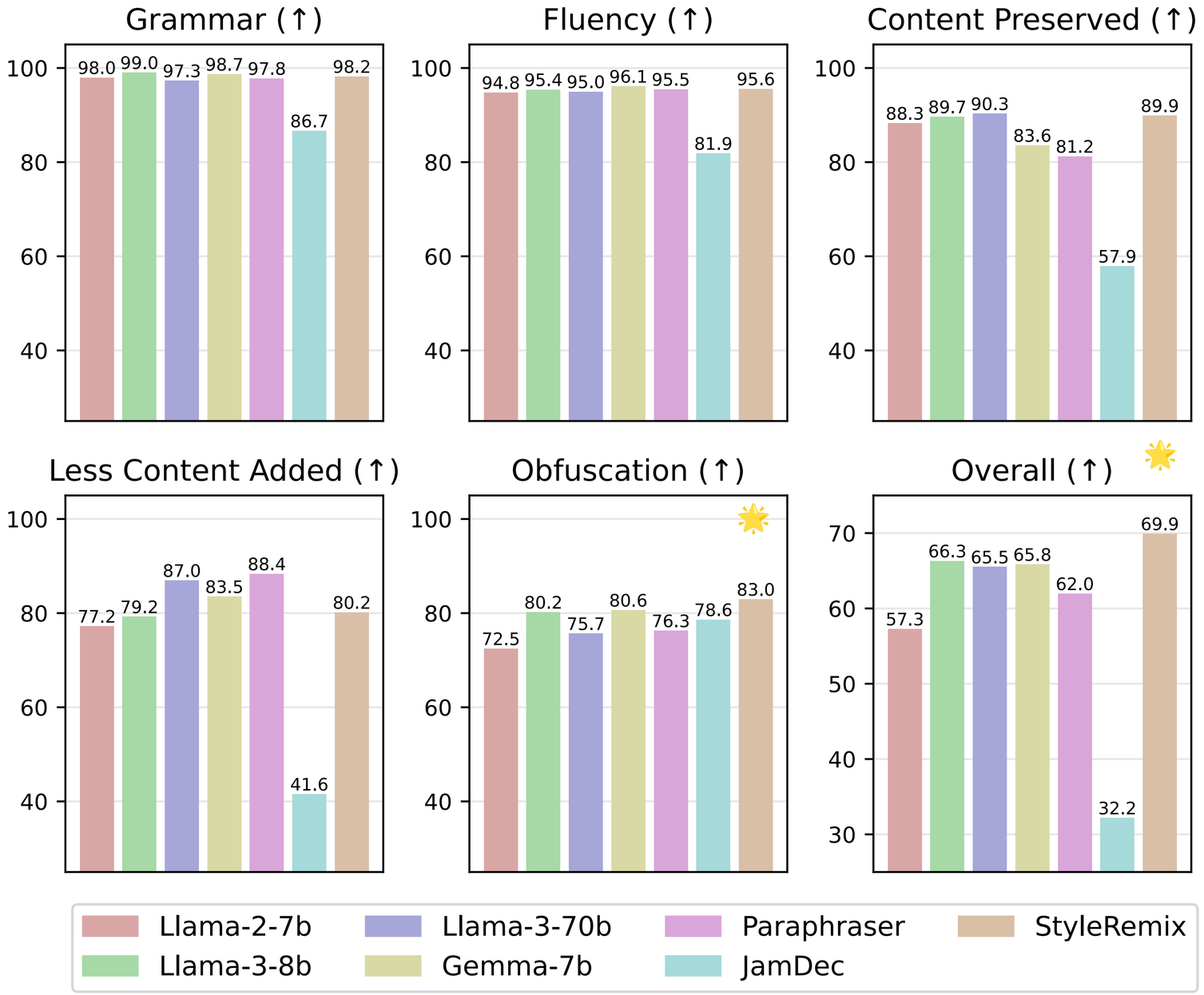
AuthorMix - Blog



Do humans agree that StyleRemix outperforms other methods?

StyleRemix: Human Evaluation

StyleRemix has best overall obfuscation quality, even compared to much larger models!



StyleRemix: Qualitative Results

Blog

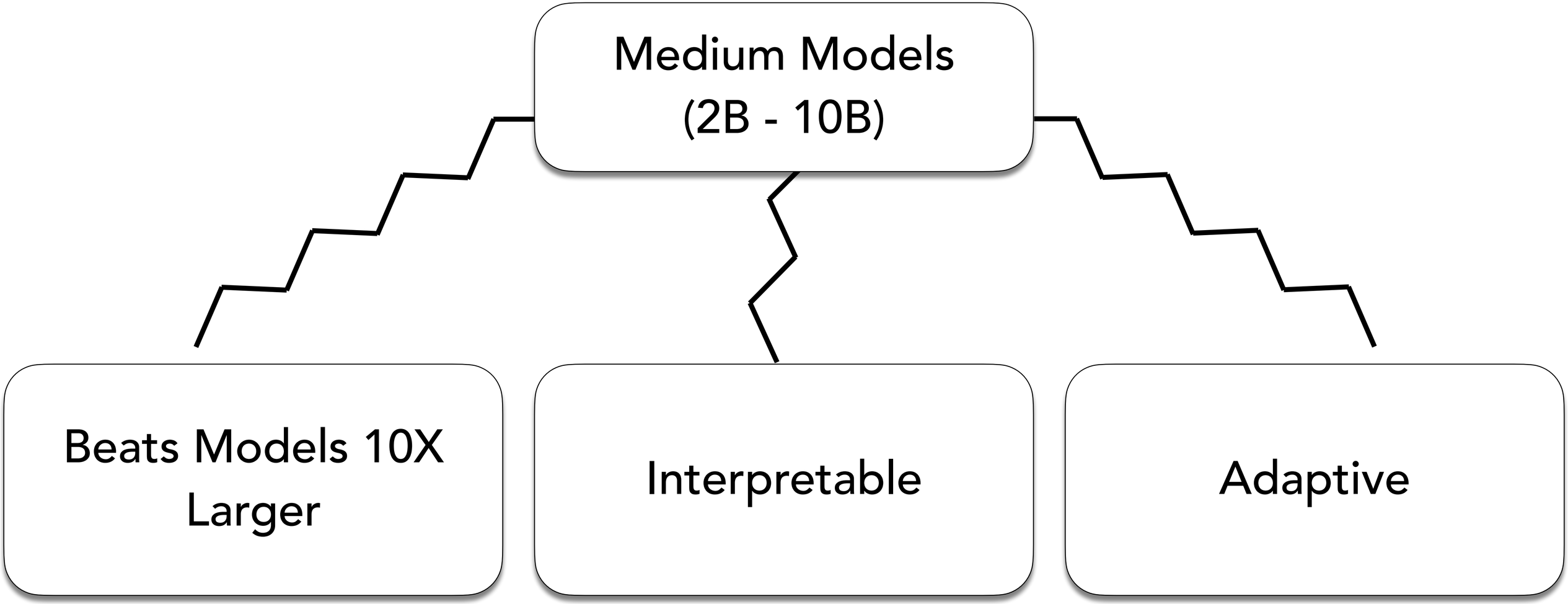
Method	Generation	
Original	100%! I was surprised, but not complaining lol. But yeah Mr. Curphey's now on crutches, hobbling around the classroom and still teaching us like his usual self.	
Paraphrase	Mr. Curphey is currently using crutches in the classroom, but he is still teaching us with his usual enthusiasm and enthusiasm. I was taken aback, but I didn't feel any remorse.	Not Grammatical
Machine Translation	100 per cent! I was surprised, but I don't complain lol. But yes, Mr. Curphey is now in the gardens to walk through the classrooms and still teach us as his ordinary self.	Not Grammatical
Stylometric	One hundred percent; i was surprised, but not complaining ll, as but yeah Mr. murphy 's now on crutches, sobbing around the classroom and always teaching us like his usual self.	Incorrect Content
JAMDEC	100%! I was surprised, but not complaining lol. But yeah Mr. Curphey's now on crutches, hobbling around the classroom and still teaching us like his usual self.	Unchanged
Llama-3 (8B)	100%! I was taken aback, but I wasn't about to grumble about it, haha! But yes, Mr. Curphey's now on crutches, limping around the classroom, and yet he's still teaching us as himself, business as usual.	Similar Style
StyleRemix	Initially, I experienced a notable degree of surprise, yet my responses remain devoid of significant complaint. Currently, Mr. Curphey, significantly, is navigating our classroom utilizing crutches, effectively maintaining his usual style of instruction, significantly.	✓ ✓ ✓

StyleRemix: Qualitative Results

Obama

Method	Generation	
Original	I recognize that times have changed since I first spoke to this convention. The times have changed, and so have I. I'm no longer just a candidate. I'm the President.	
Paraphrase	Since I first addressed this convention, I have seen significant changes. As a result, I have evolved from a candidate to the current President.	Not Grammatical
Machine Translation	I recognize that the times have changed since I first spoke to this convention. The times have changed, and that's how I have. I'm not just a candidate anymore. I am the President.	Incorrect Content
Stylometric	I acknowledge that times have changed since I 1st spoke to this convention, and the times have changed, and so have I. I am no longer just a candidate; and myself 'm the President.	Not Grammatical
JAMDEC	It's important to remember where we came from, where we're going, and what we're trying to do," he said, adding, "I think it's important for all of us to be. We are the people who built this country, and we have a responsibility to make sure that we preserve it for our children and our children's children, and that's what we're going to do,".	Added Information
Llama-3 (8B)	Since my initial address to this esteemed convention, I have come to acknowledge that the times have indeed changed. So too have I. Gone are the days when I stood before you as a mere candidate, eager to make my case. Today, I stand before you as the President, with a wealth of experience and insight gathered during my time in office.	Added Information
StyleRemix	Oh, how the world has transformed since I first addressed you all here! Indeed, the world has shifted, and so have I. It's not just about being a candidate anymore—it's about being the President.	✓ ✓ ✓

StyleRemix

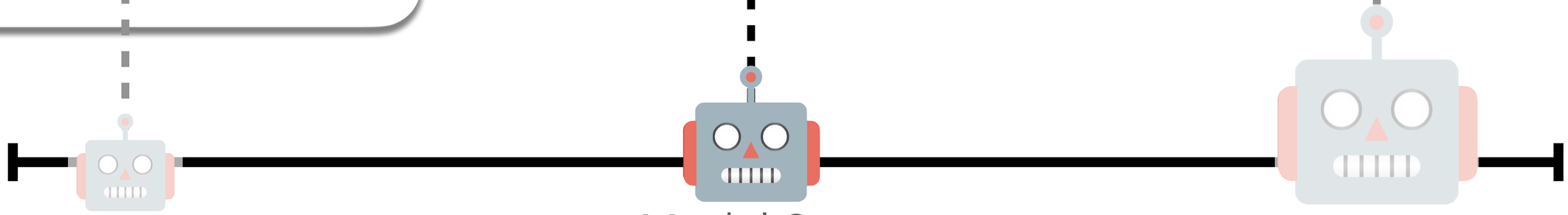

Medium Models
(2B - 10B)
Beats Models 10X Larger
Interpretable
Adaptive

Controllable Generation

Small Models
(<2B)

Medium Models
(2B - 10B)

Large Models
(>10B)



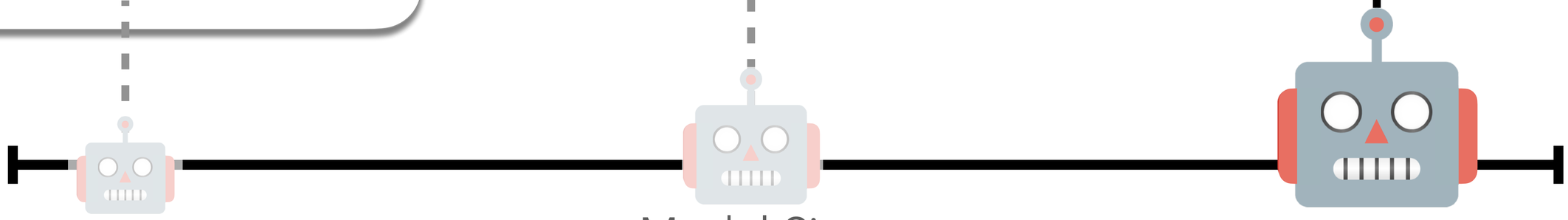
Model Size

Controllable Generation

Small Models
(<2B)

Medium Models
(2B - 10B)

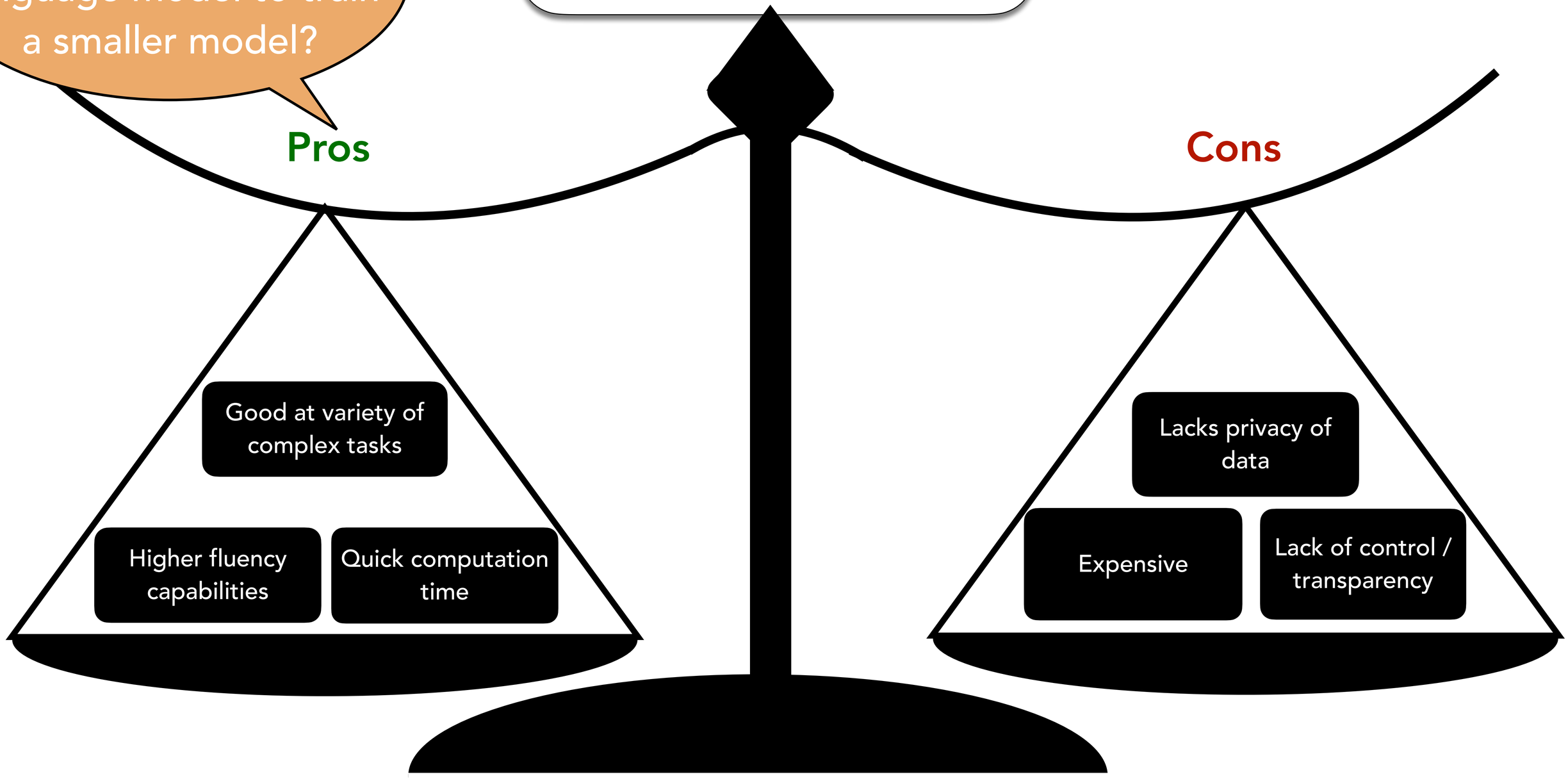
Large Models
(>10B)



Model Size

Can we leverage a large language model to train a smaller model?

Large Models (>10B)

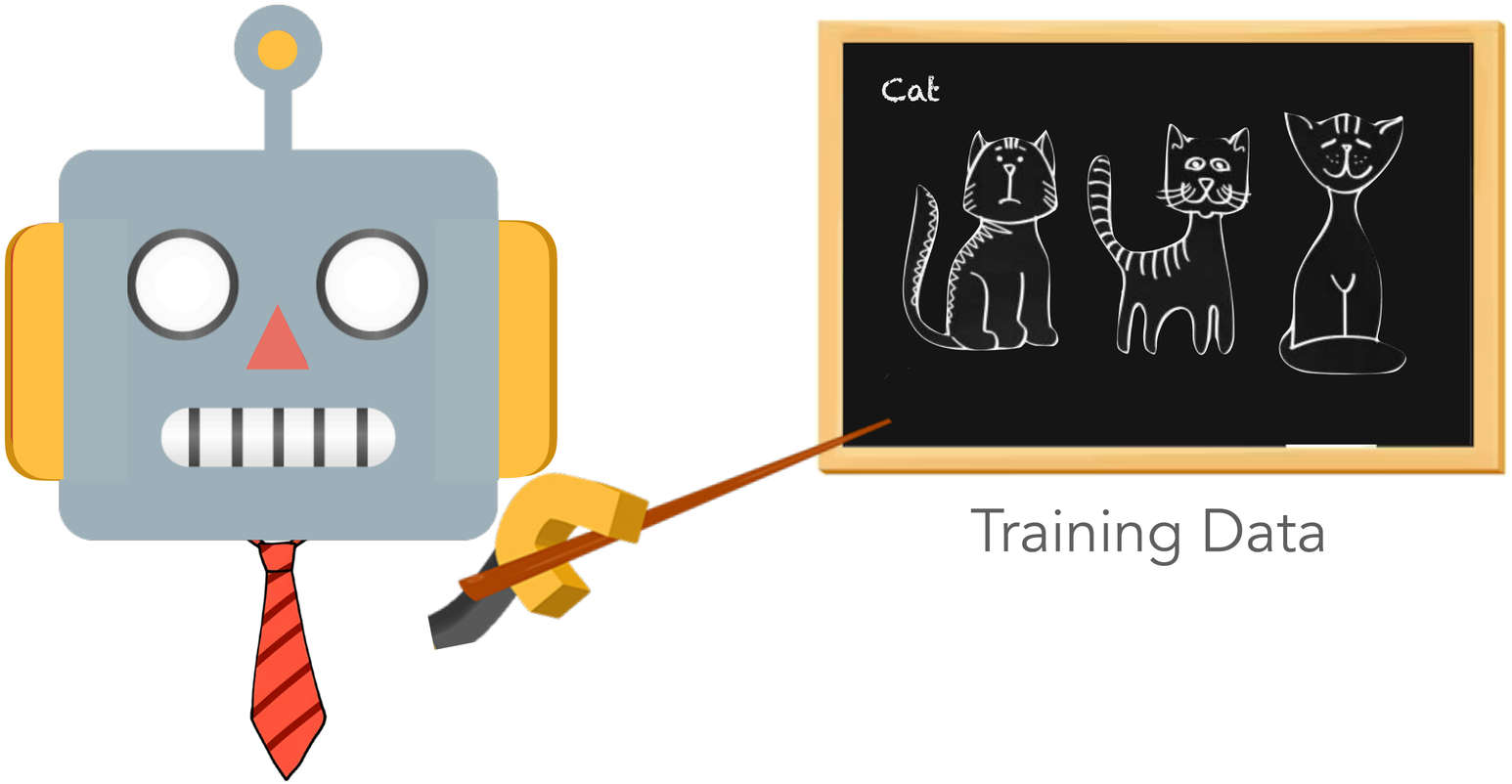


Large Models (>10B)

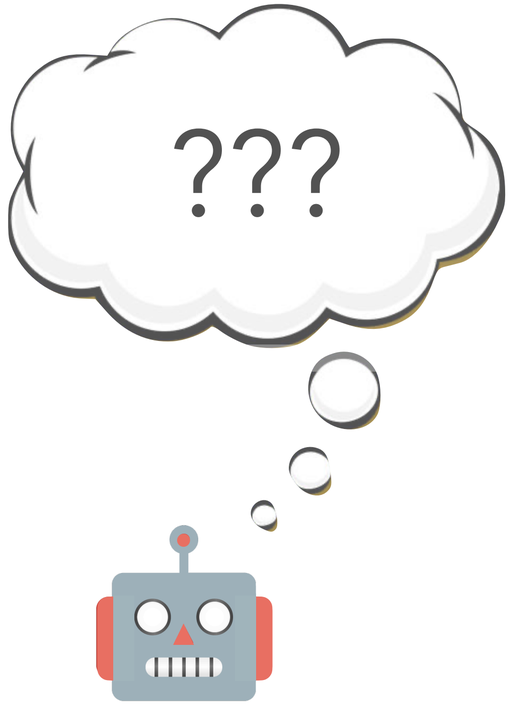
Can we leverage a large language model to train a smaller model?

Knowledge Distillation

Large Model
(Teacher)



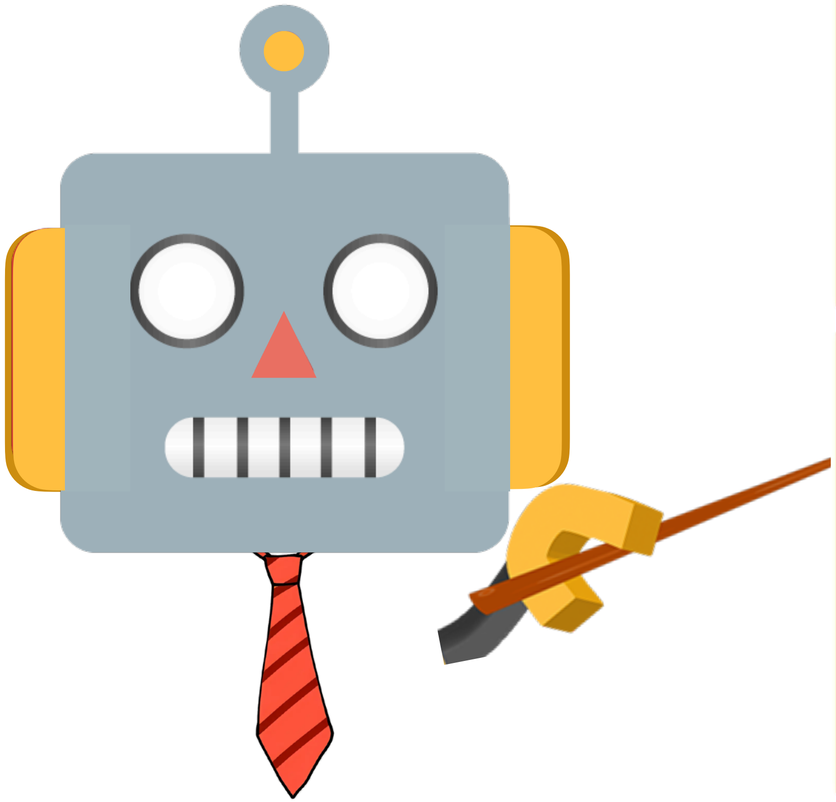
Training Data



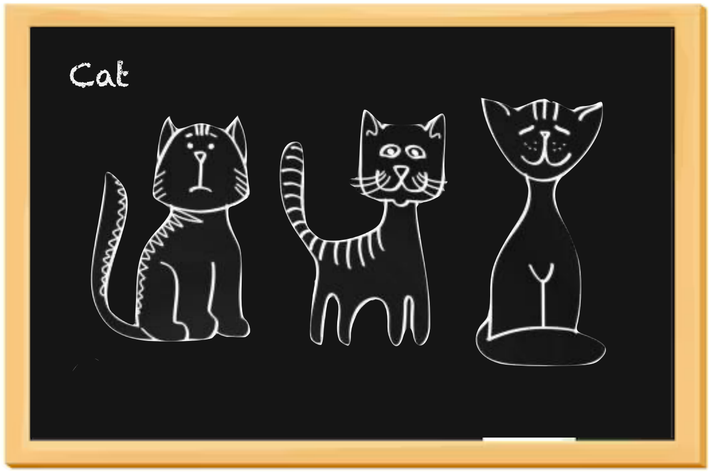
Small Model
(Student)

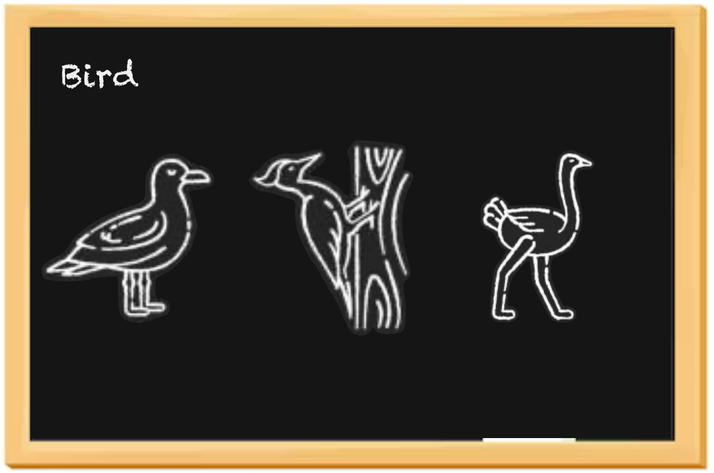
Knowledge Distillation

Large Model
(Teacher)



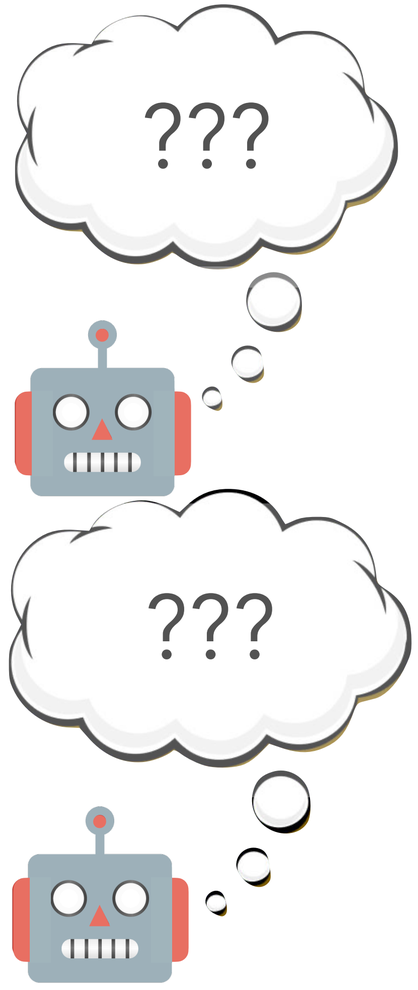
Training Data





Double the Resources!

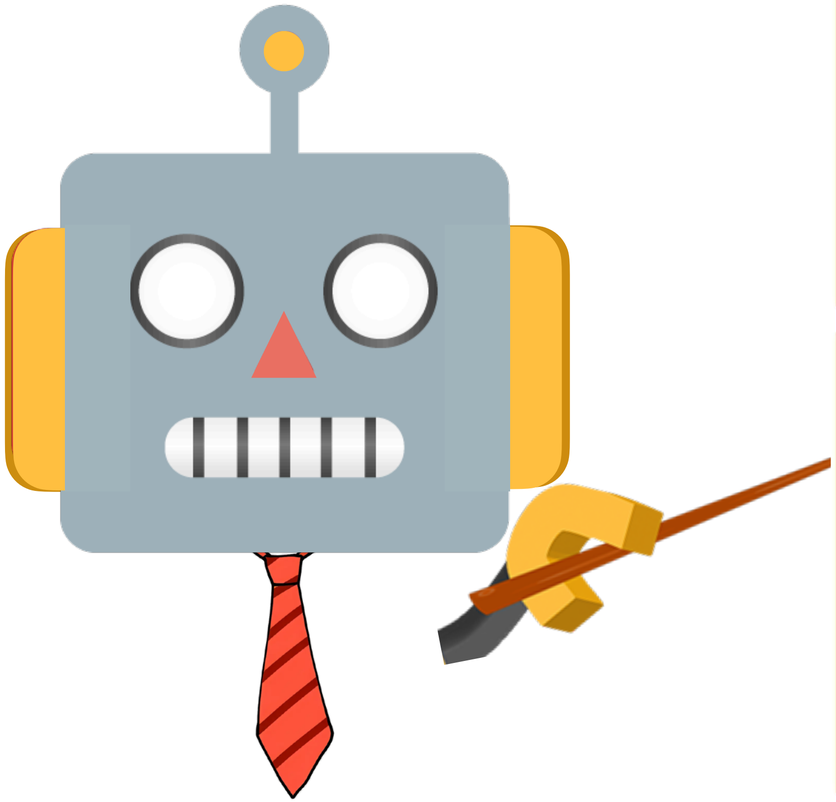
Doesn't utilize possible overlap!



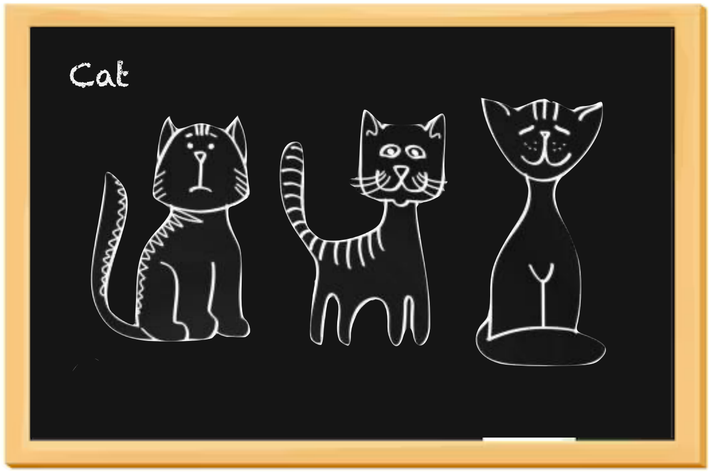
Small Model
(Student)

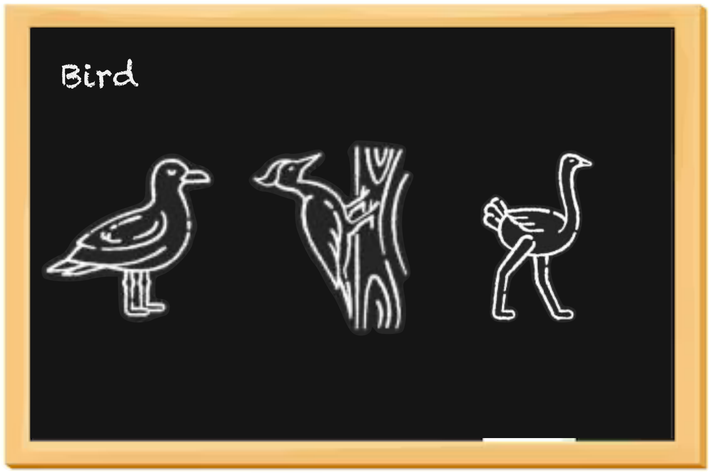
Knowledge Distillation+

Large Model
(Teacher)



Training Data

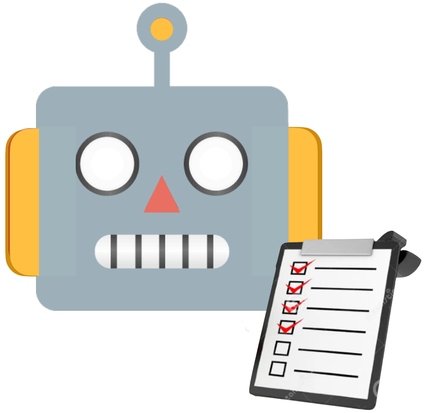


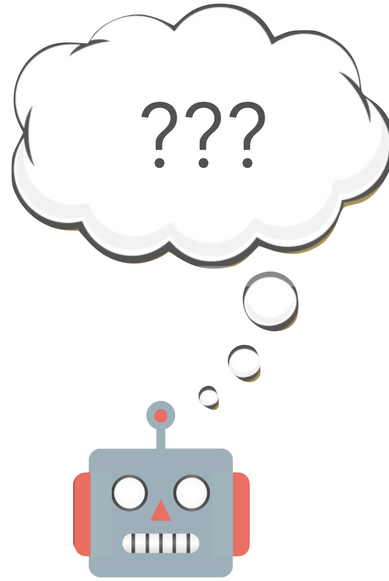


TA Responsibilities

- Supervise that student model learns all knowledge equally
- Adaptive Loss function: compares student to TA's response

Medium Model
(Teaching Assistant)





Small Model
(Student)

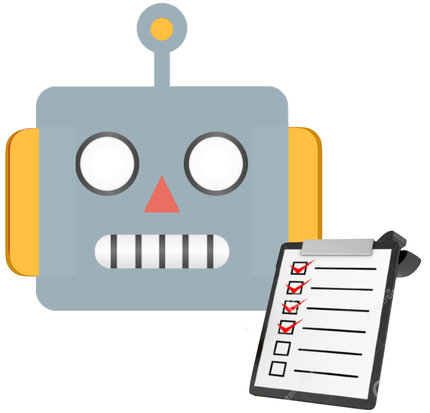
Knowledge Distillation+

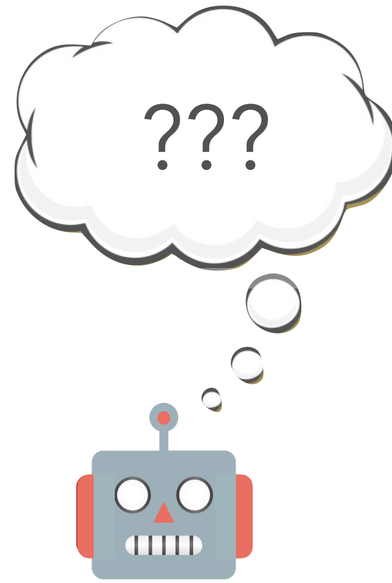
Large Model
(Teacher)

Training Data

TA Responsibilities
- Supervise that student model learns all [illegible] equally
- Adaptive loss function: [illegible] student to TA's [illegible]

Medium Model
(Teaching Assistant)





Small Model
(Student)

Coming Soon

Thank You!

JAMDEC

Paper

Code

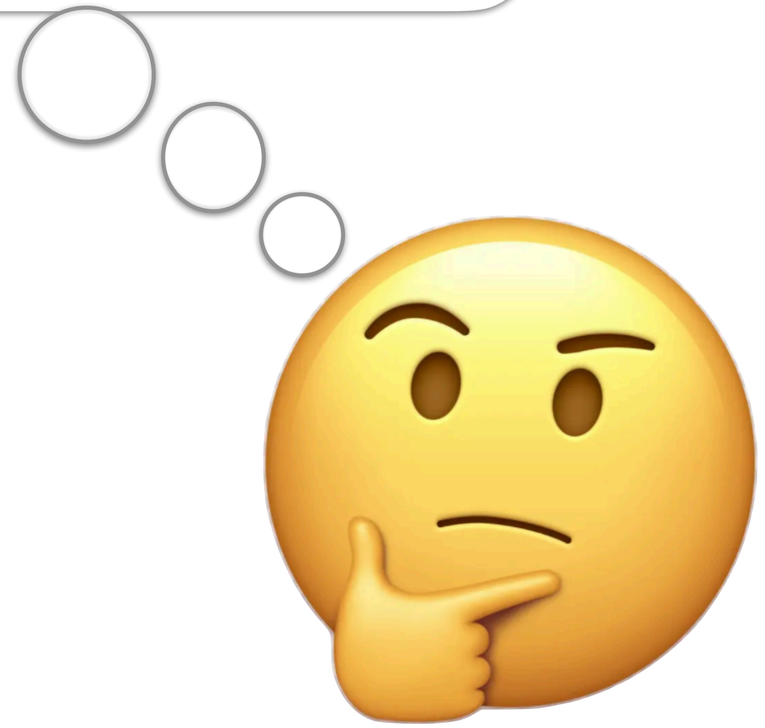
StyleRemix

Paper

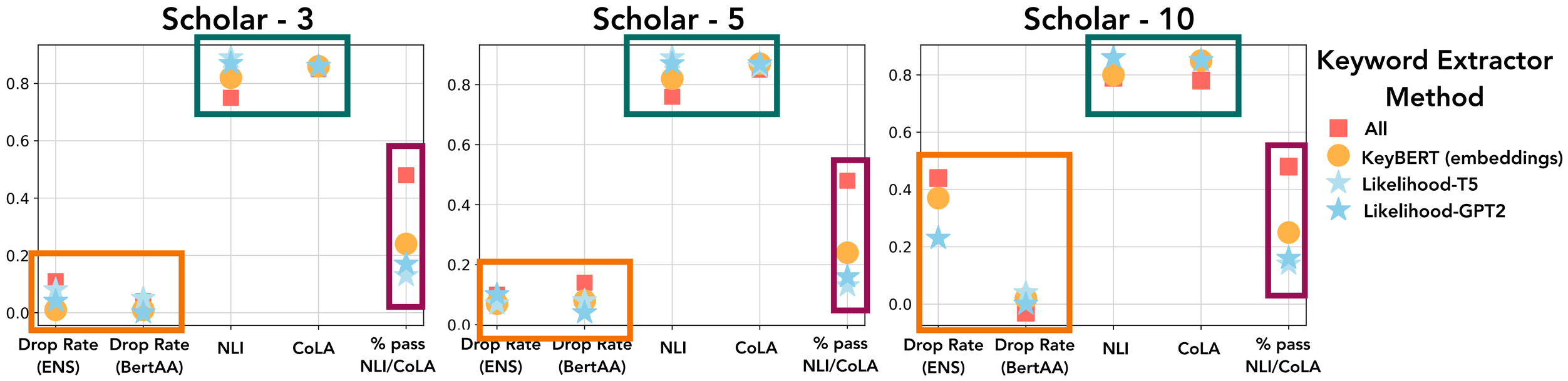
Code

Appendix

Does our innovation to the pipeline result in better downstream performance? Likelihood Keyword Extraction? Constrained-Diversity Beam search?



JAMDEC: Keyword Extraction Comparison



All methods have similar drop rate (Obfuscation)

Likelihood methods have higher NLI and similar CoLA (Fluency/Grammar)

Using all three results in higher % passing NLI/CoLA threshold

↳ Each method produces diverse set of keywords

JAMDEC: Diversity Results

		JAMDEC	
Dataset	**Metric**	**W/O Diversity**	**W/ Diversity**
Scholar - 3	Drop Rate (ENS)	0.01	**0.11**
	Drop Rate (BertAA)	0.08	0.04
	NLI	**0.87**	0.81
	CoLA	**0.86**	0.79
	Average Gen.	0.16	**0.52**
Scholar -5	Drop Rate (ENS)	0.1	**0.1**
	Drop Rate (BertAA)	0.01	**0.14**
	NLI	**0.87**	0.76
	CoLA	**0.87**	0.85
	Average Gen.	0.16	**0.48**

$\sim 5\,\%$ increase in Obfuscation

$\sim 6\,\%$ decrease in NLI/CoLA

$\sim 35\,\%$ increases in generations passing NLI/CoLA threshold