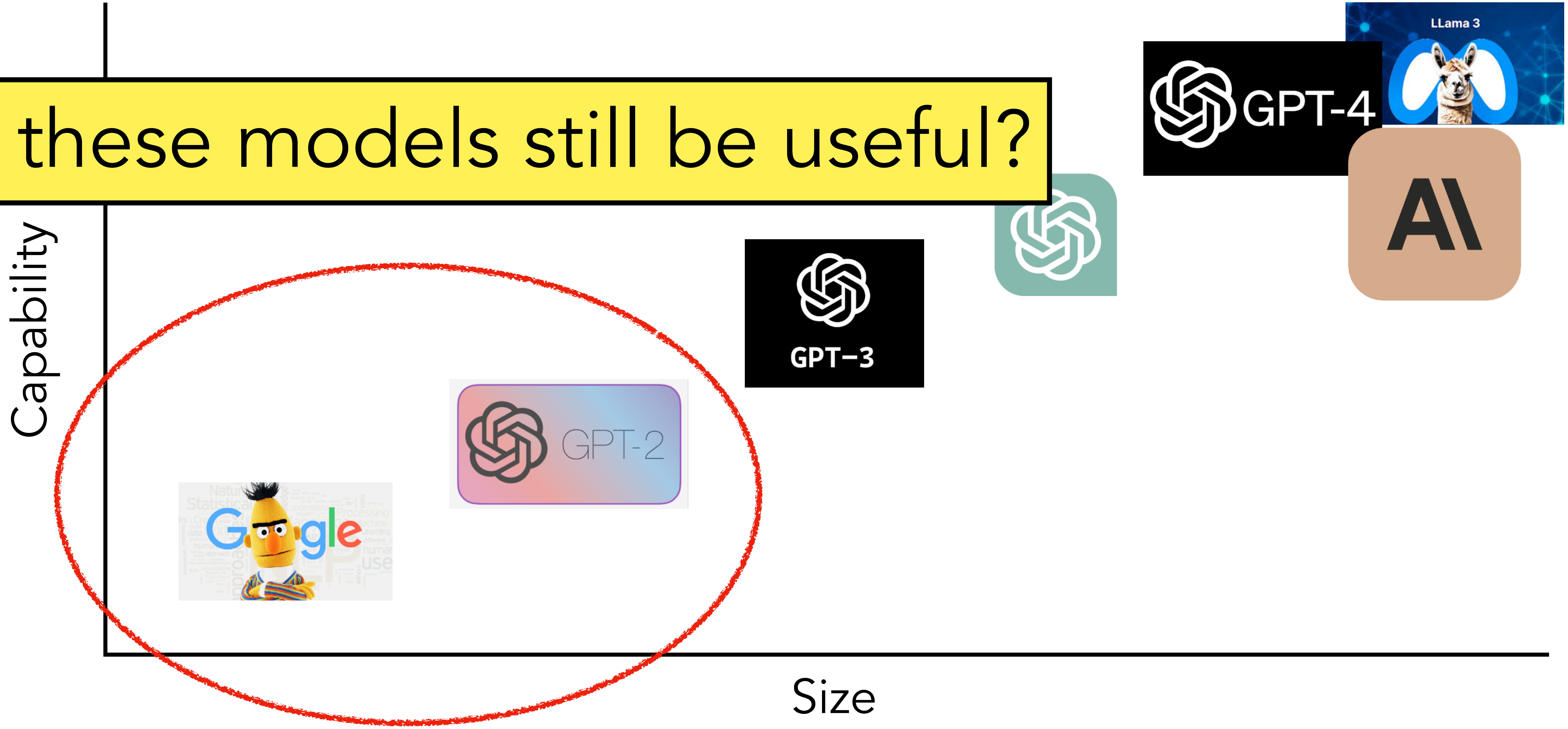
Small but MIGHTY

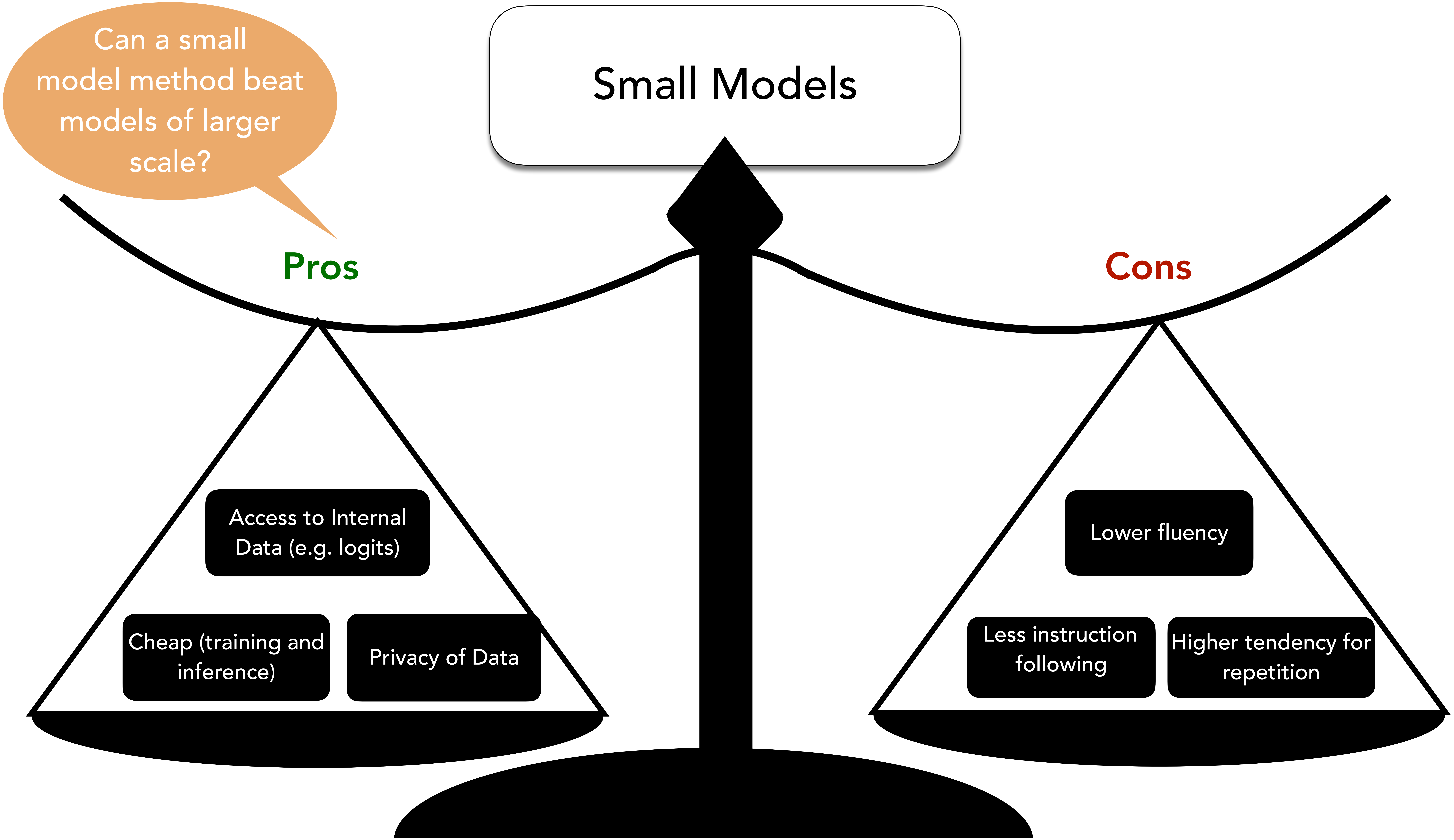
Empowering Small Language Models to Outperform Their Larger Counterparts

Presented by Jillian Fisher & Skyler Hallinan

Language Model Scaling

Can these models still be useful?

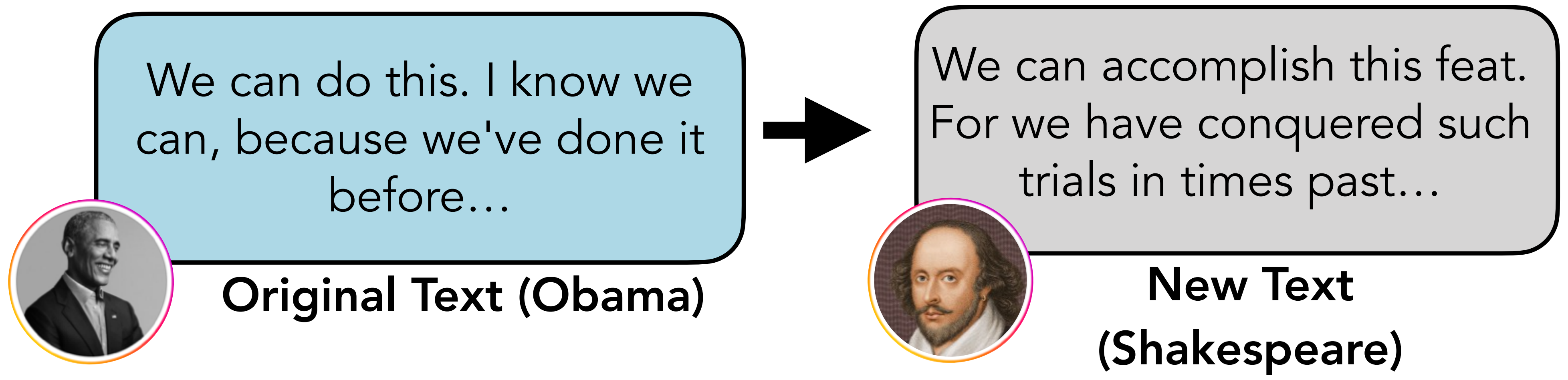



Can a small model method beat models of larger scale?
Small Models
Pros
Cons
Access to Internal Data (e.g. logits)
Cheap (training and inference)
Privacy of Data
Lower fluency
Less instruction following
Higher tendency for repetition

Improving on Text to Text Generation Tasks

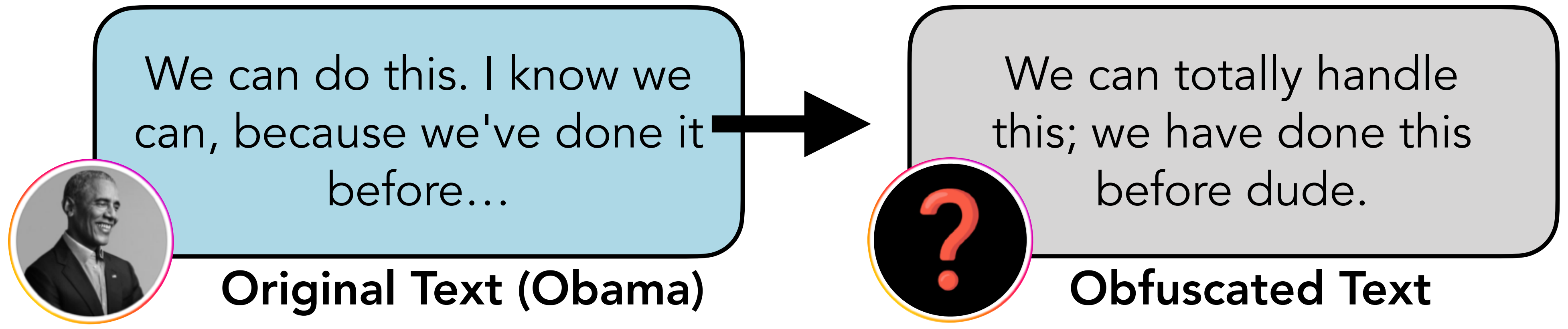
Style Transfer

Objective: *Target Style*



Authorship Obfuscation

Objective: *Not Original Author Style*



Improving on Text to Text Generation Tasks

Tasks:

Style Transfer

Authorship Obfuscation

Methods:

Inference Time Only Method

Expert Distillation Method

Knowledge Distillation + Inference Time Method

Improving on Text to Text Generation Tasks

Tasks:

Style Transfer

Authorship Obfuscation

Methods:

Inference Time Only Method

Expert Distillation Method

Knowledge Distillation + Inference Time Method

JAMDEC: Unsupervised Authorship Obfuscation using Constrained Decoding over Small Language Models

Jillian Fisher, Ximing Lu, Jaehun Jung, Liwei Jiang, Zaid Harchaoui, Yejin Choi

Findings of NAACL, 2024.

Authorship Obfuscation

What?

Rewriting text to obscure the original author's identity

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

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

Why?

Blind Review for Scientific Papers

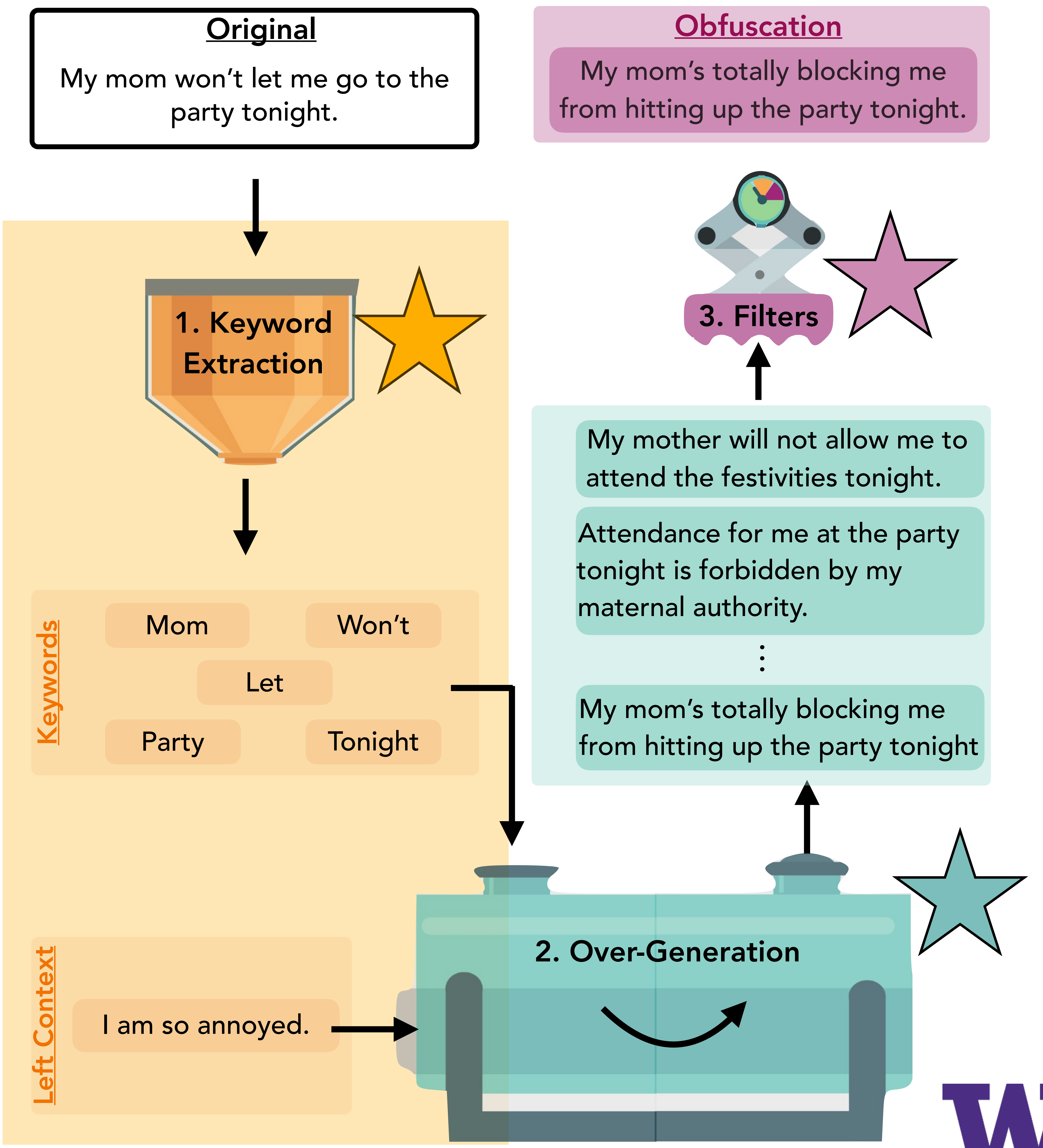
Interaction on Mental Health Forums

Anonymous Online Review

JAMDEC Decoding

- 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*: Extract keywords to maintain original content
  2. *Over-generation*: Generate many diverse outputs that include the keywords
  3. *Filters*: Maintain fluency and content preservation, +any user-specified control



Innovations: Keyword Extraction

- Current methods rely on word-embeddings with similar cosine similarity to whole phrase

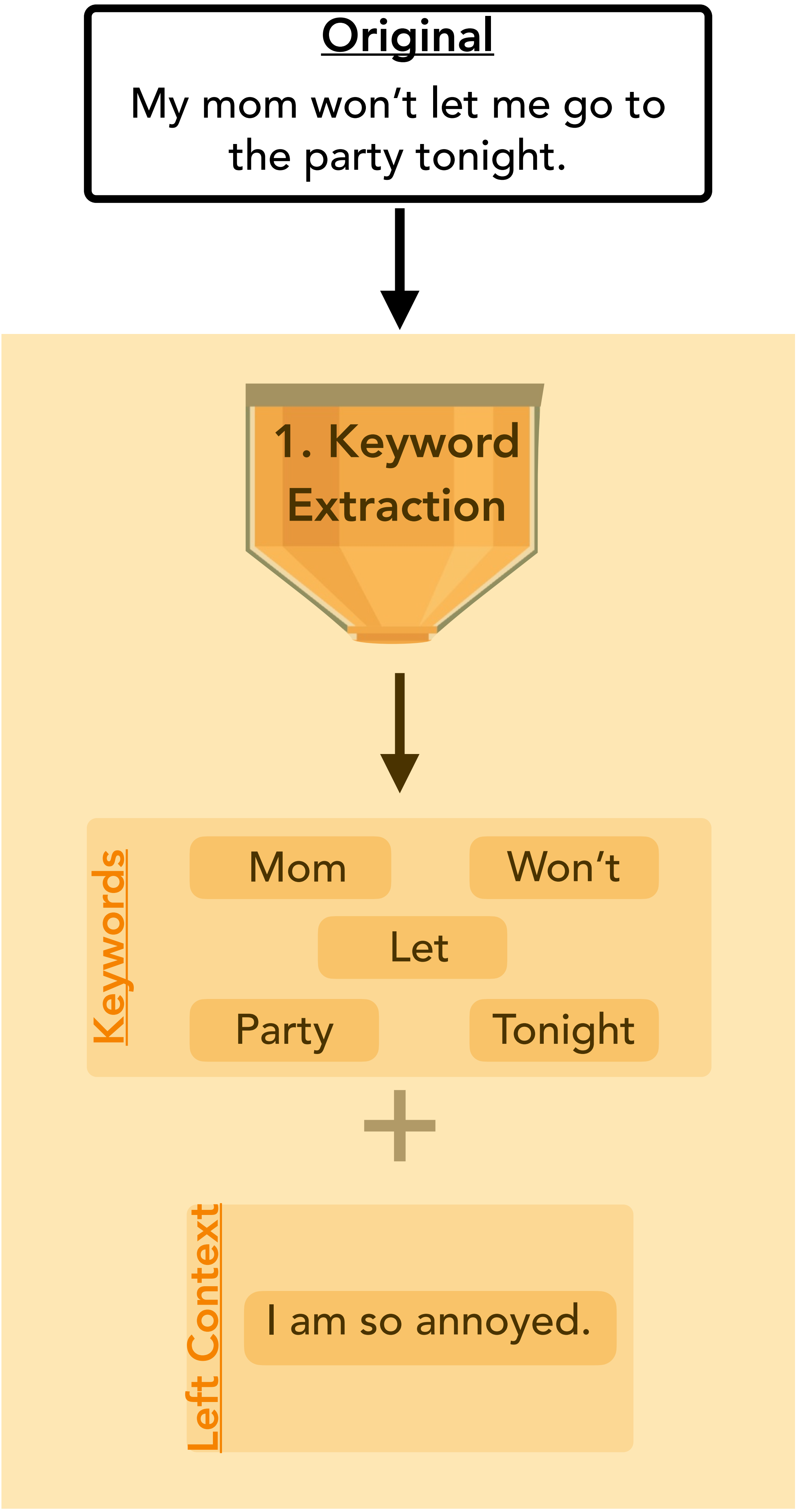
New Likelihood-based Method

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

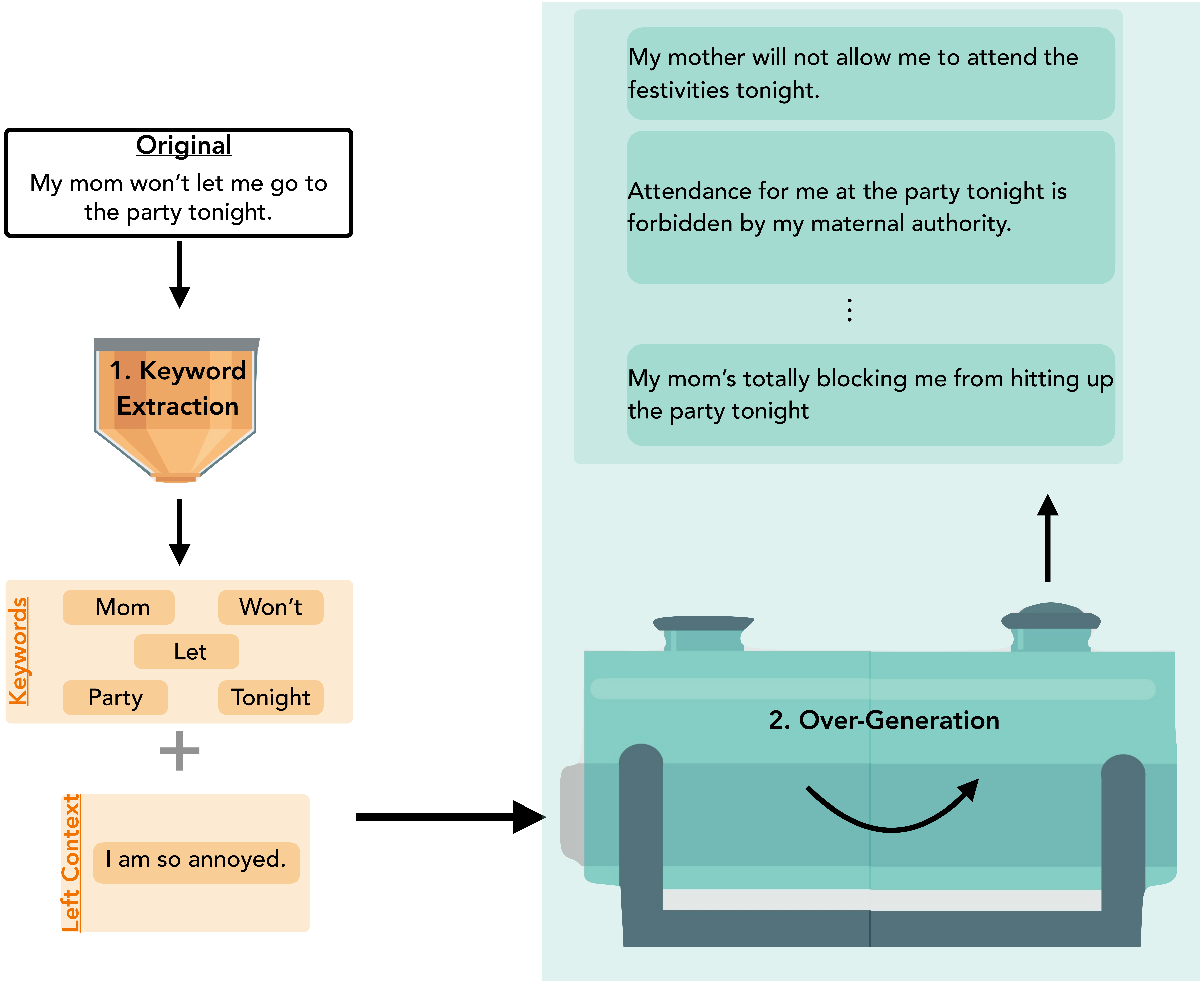
$$P(t_i \,|\, \text{context}_i)$$

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

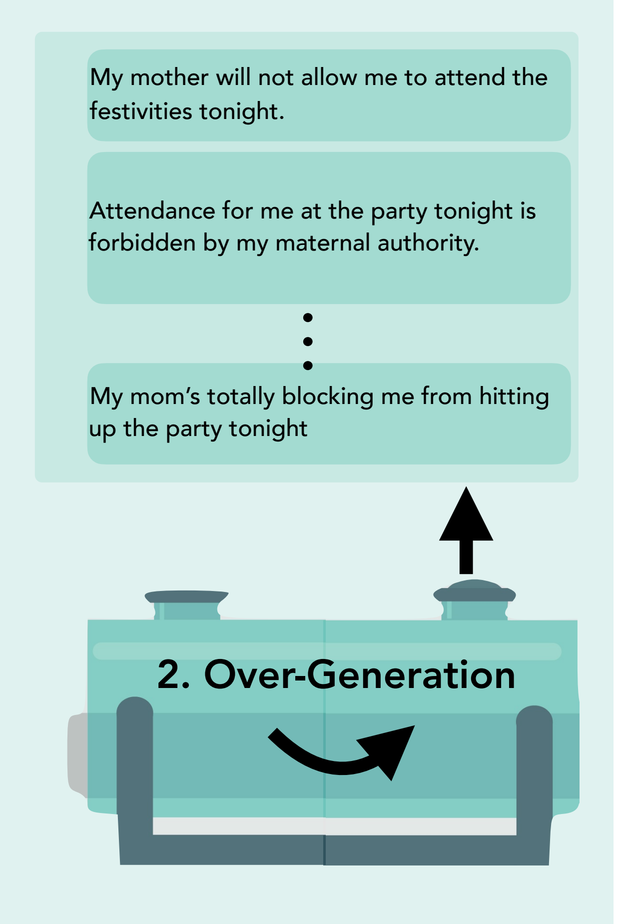
Auto-Regressive (GPT2)	$P(t_i \,\vert\, t_1, t_2, \ldots, t_{i-1})$
Text-to-Text (T5)	$P(t_i \,\vert\, t_1, \ldots, t_{i-1}, [MASK], t_{i+1}, \ldots, t_n)$



Innovations

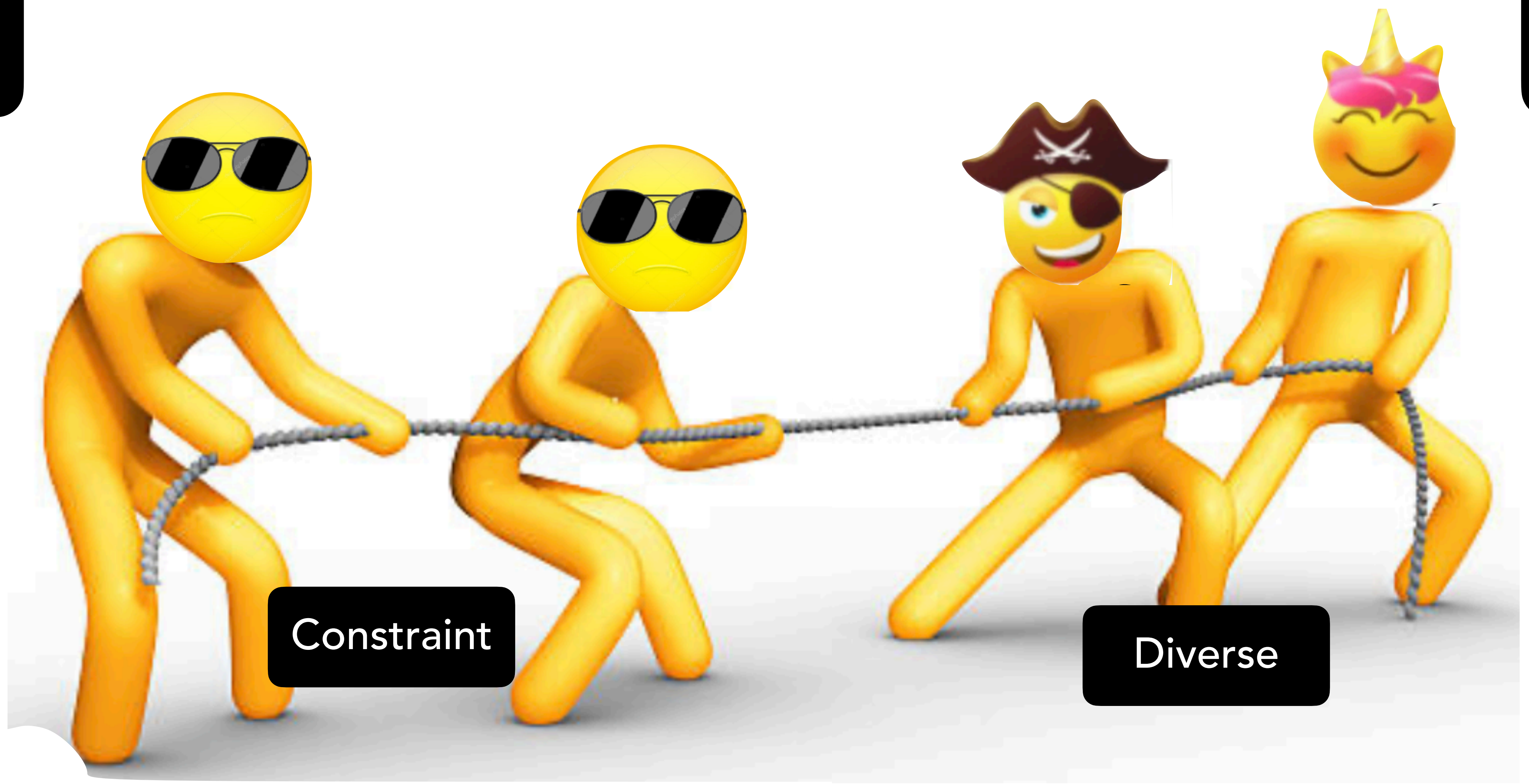


Innovations: Over-Generation



Constrain to original content

Create diverse authorship styles



Constrained + Diverse Beam Search (CoDi-BS)

Constrained + Diverse Beam Search (CoDi-BS)

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

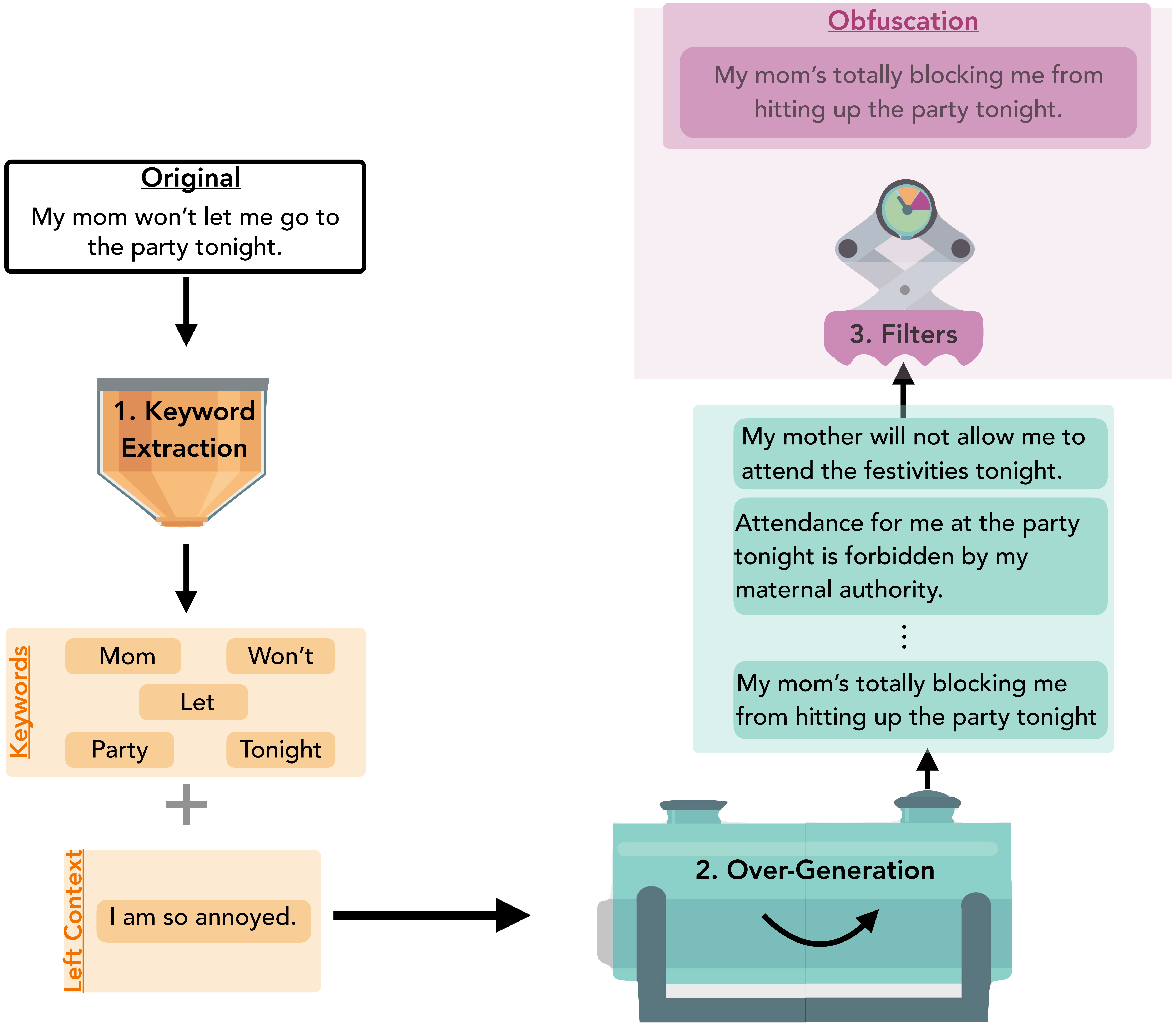
Where x is sequence of previous tokens, $y \in Y$ is the output sequence, and $\theta \in \Omega$ is the parameter vector.

Add Diversity

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

Where $L \in \mathbb{R}^v$ is the logits, $F \in \mathbb{R}^v$ is a vector of frequency of each token chosen in the previous beams, and λ is a hyperapramter

Innovations

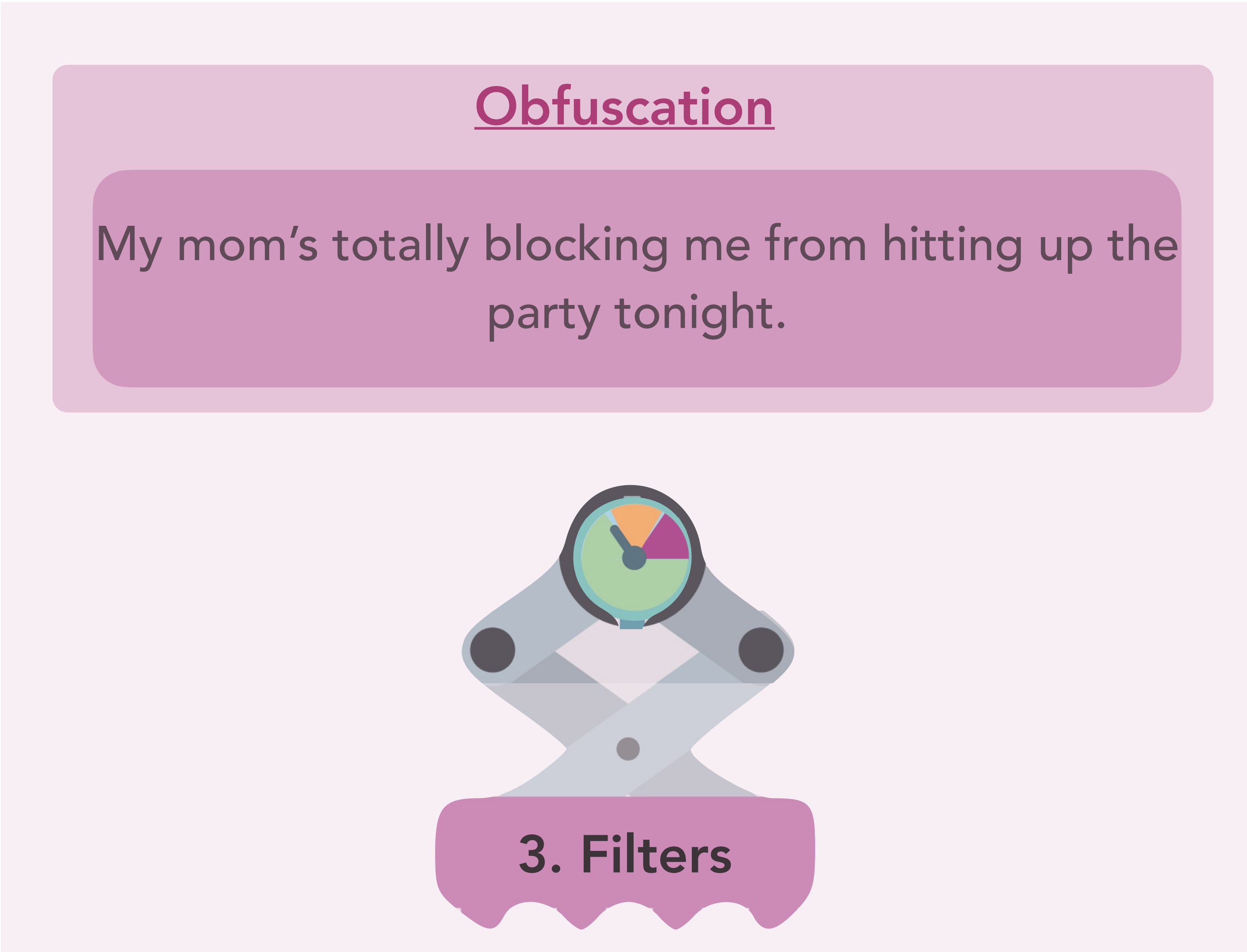


Innovations: Filtering

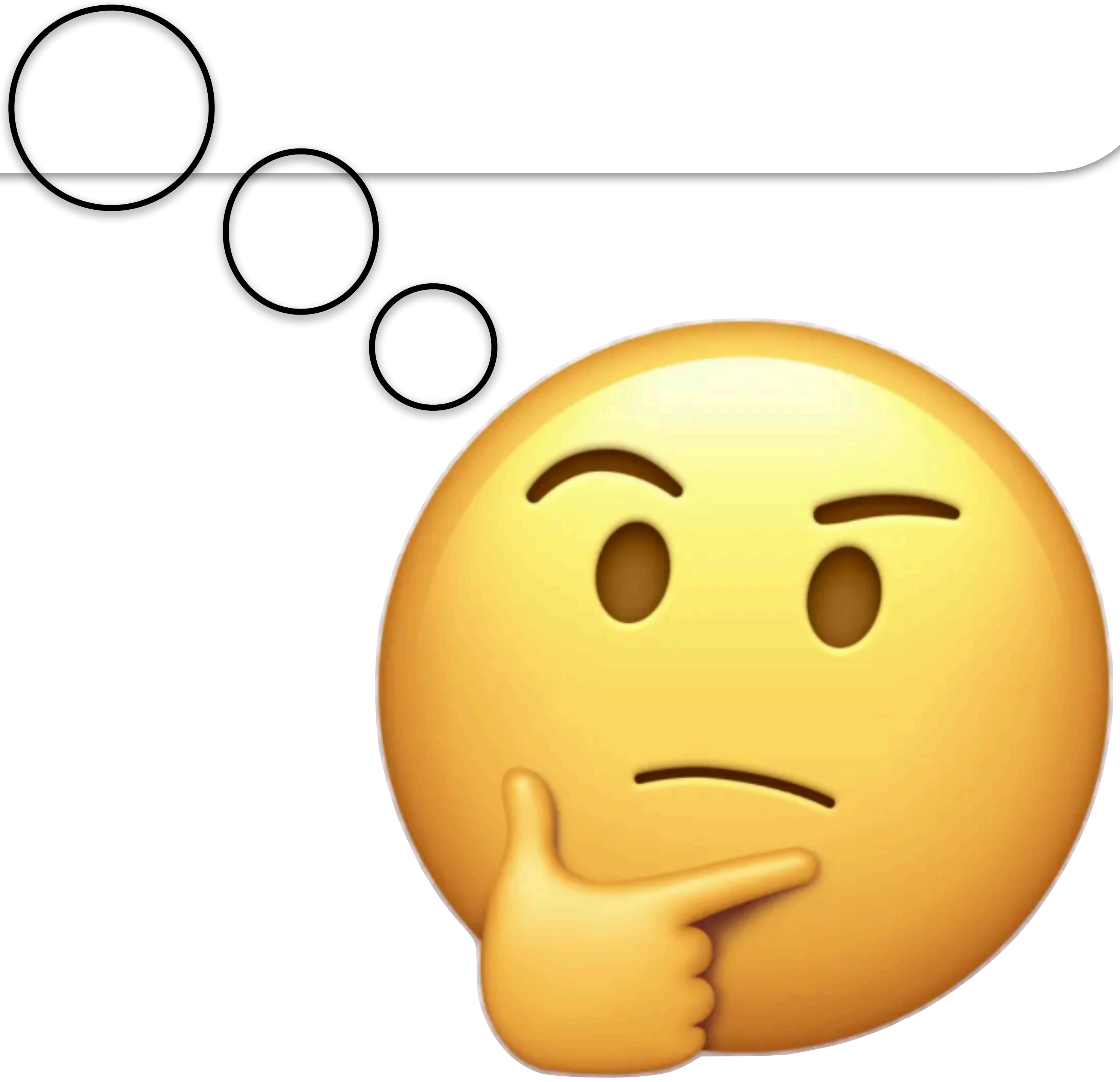
Filtering

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

- Customizable!
  - Length
  - Formality
  - Grade level



How does JAMDEC perform compared to other methods?



JAMDEC: Experimental Setup

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

- **<u>Baselines</u>**
  - *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
  - Paraphrase

JAMDEC: 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 (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**

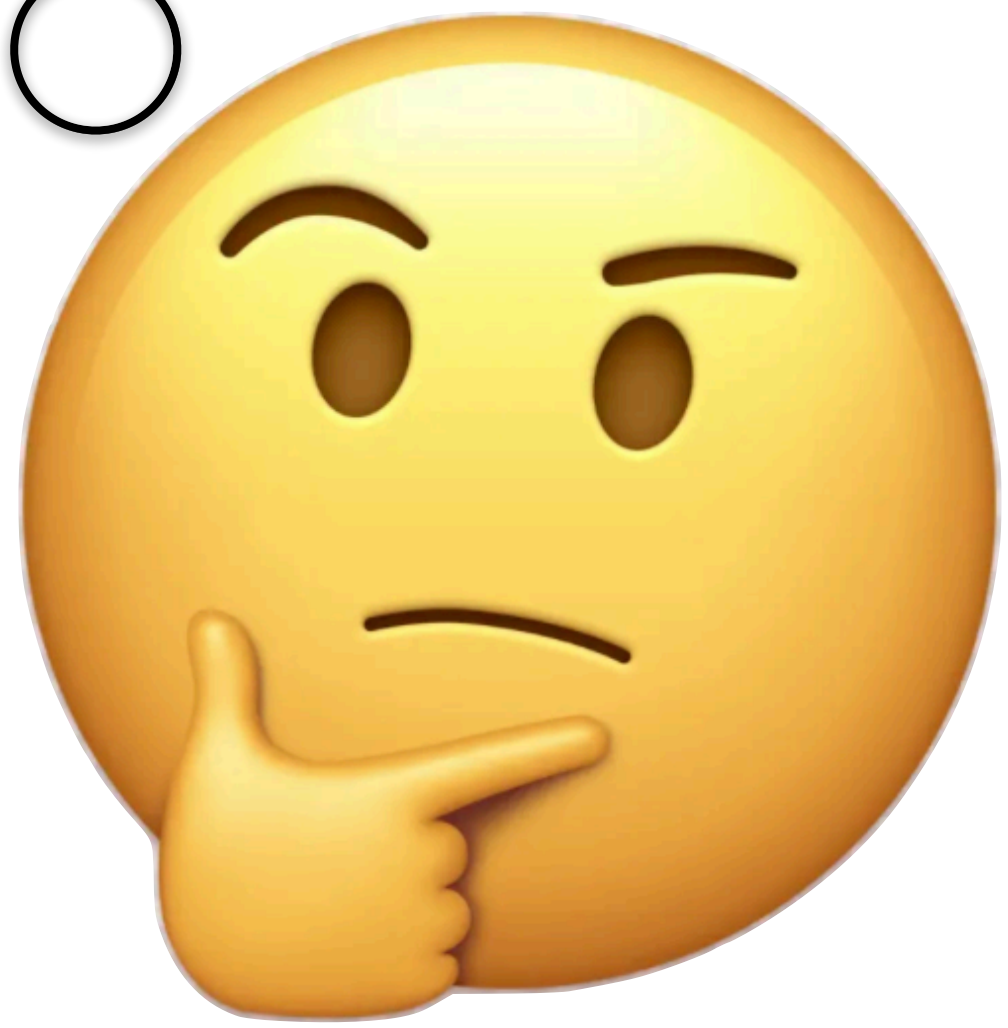
JAMDEC had the highest overall Task Score on every dataset!

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 similar to much larger models!

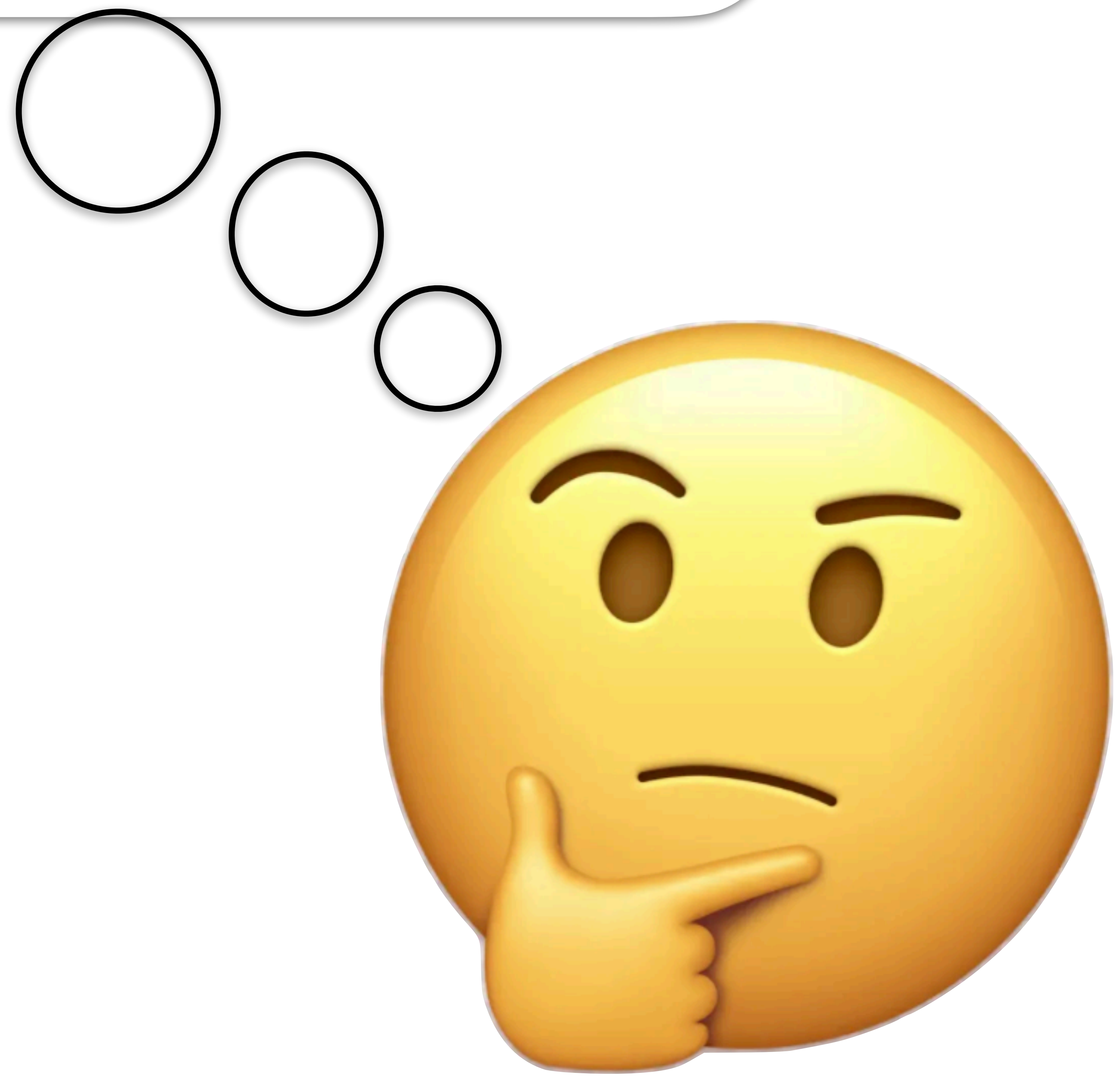
Would humans also agree that JAMDEC outperforms other methods?



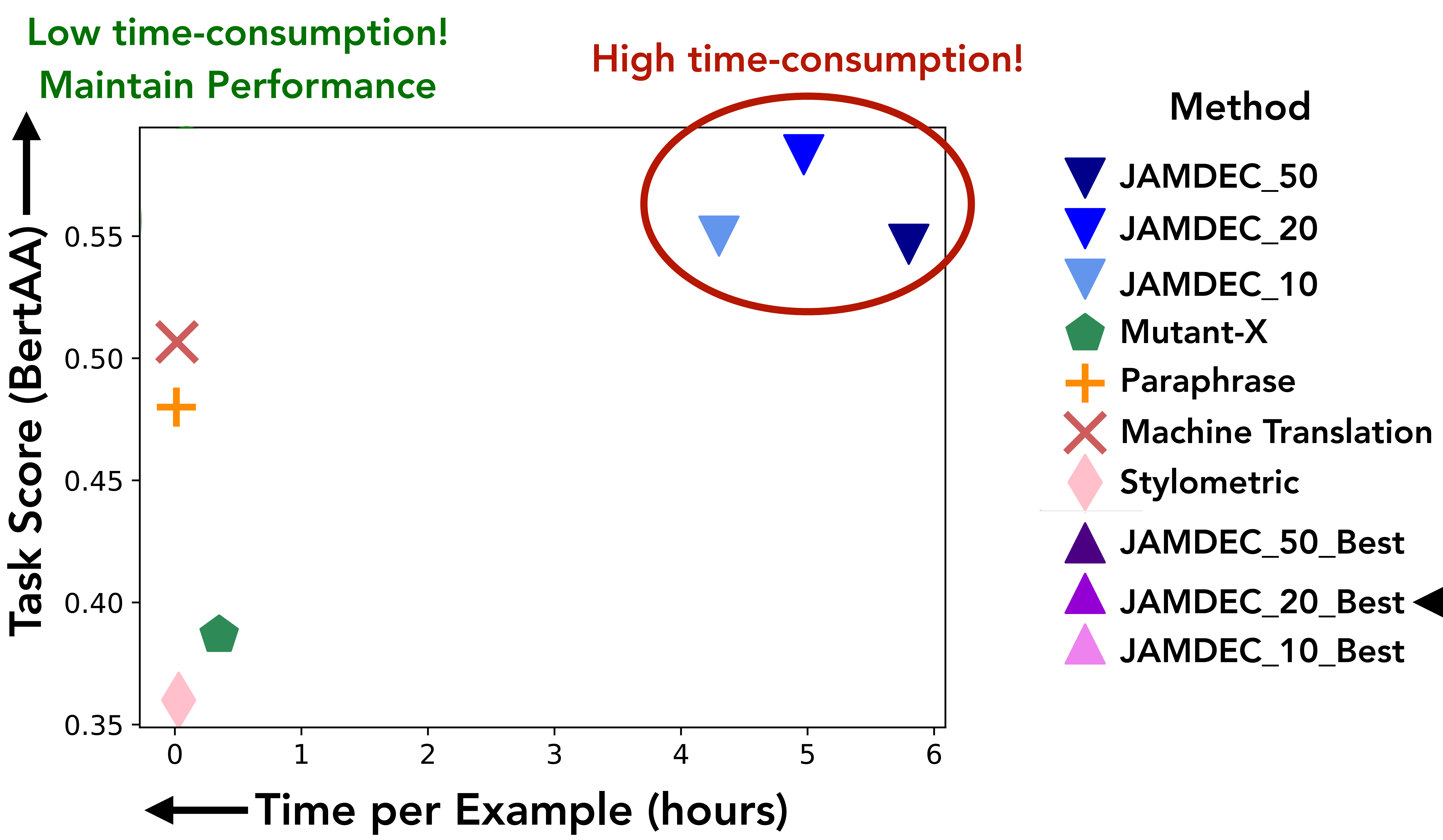
JAMDEC: Qualitative Results

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.**	Ungrammatical
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.	✓✓✓

Having to do over-generation seems like it would take <u>more time</u> than other methods



JAMDEC: Computational Time



Best keyword extraction, type of constraints, type of algorithm (sampling? Diversity?)

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

More in the Paper

- Comparison of trade-off between obfuscation, content-preservation, and grammaticality
- Ablation of JAMDEC Method (different beam width, with/without diversity, different filters, etc.)
- Comparison of "Style Transfer" methods
- Evaluation using "Adversarial Threat Models"
- Discussion of similarity to other tasks (paraphrasing, style transfer, authorship attribution, etc.)
- *And MORE*!

Improving on Text to Text Generation Tasks

Tasks:

Style Transfer

Authorship Obfuscation

Methods:

Inference Time Only Method

Expert Distillation Method

Knowledge Distillation + Inference Time Method

Improving on Text to Text Generation Tasks

Tasks:

Style Transfer

Authorship Obfuscation

Methods:

Inference Time Only Method

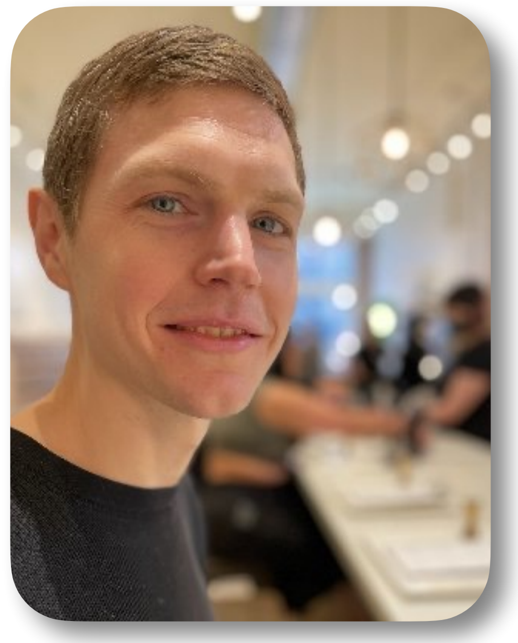
Expert Distillation Method

Knowledge Distillation + Inference Time Method

🐂 STEER: Unified Style Transfer with Expert Reinforcement



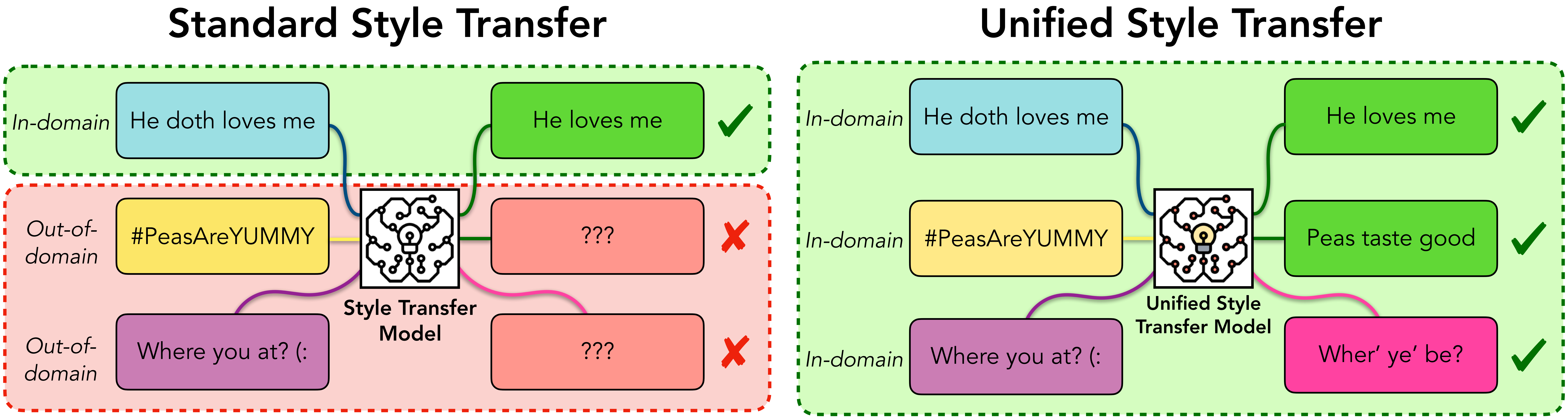





Skyler Hallinan, Faeze Brahman, Ximing Lu, Jaehun Jung, Sean Welleck, and Yejin Choi

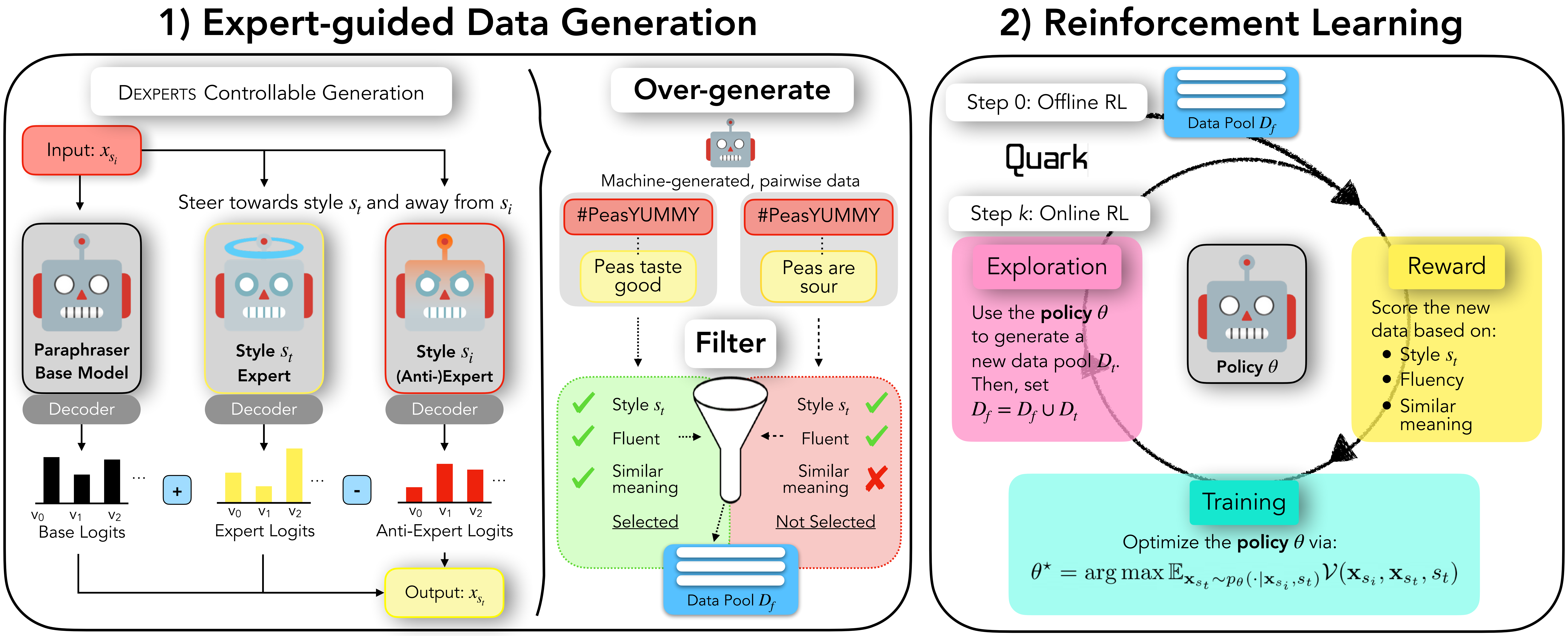
Findings of EMNLP, 2023. Presented at NILLI 2023.

Background: Style Transfer



Problem: No parallel data and a poor initial policy

Method: STEER



Dataset

- Training: the Corpus of Diverse Styles (CDS) [1]
  - 15 million sentences with minimal preprocessing
  - 11 diverse styles from multiple sources, including the web and literature
- Examples demonstrate the diversity of the corpus

Style	Size	Style	Size
Shakespeare	27.5K	Lyrics	5.1M
James Joyce	41.2K	1810-1830	216.0K
English Tweets	5.2M	1890-1910	1.3M
AAE Tweets	732.3K	1990-2010	2.0M
Romantic Poetry	29.8K	Bible	34.8K
Switchboard	148.8K		

What, are you busy, ho?

But, as I said, On Lammas Eve at night shall she be fourteen.

Shakespeare

if y- you know instead of

and uh cranberry sauce i- i could eat just that and be satisfied

Switchboard

[1] Krishna, K., Wieting, J., & Iyyer, M. (2020). Reformulating Unsupervised Style Transfer as Paraphrase Generation. ArXiv, abs/2010.05700.

Evaluation

- Style transfer traditionally evaluated on:
    - **Target Style Strength**: *How well does the style transfer fit in the target style?*
    - **Fluency**: *How understandable is the text?*
    - **Meaning Similarity:** *How similar in meaning is the generation to the original text?*
- Style transfer metrics can be assessed with automatic classifiers
- Following previous work [1], we take an **aggregate** of the three metrics, to get a single score representing the **overall quality** of style transfer

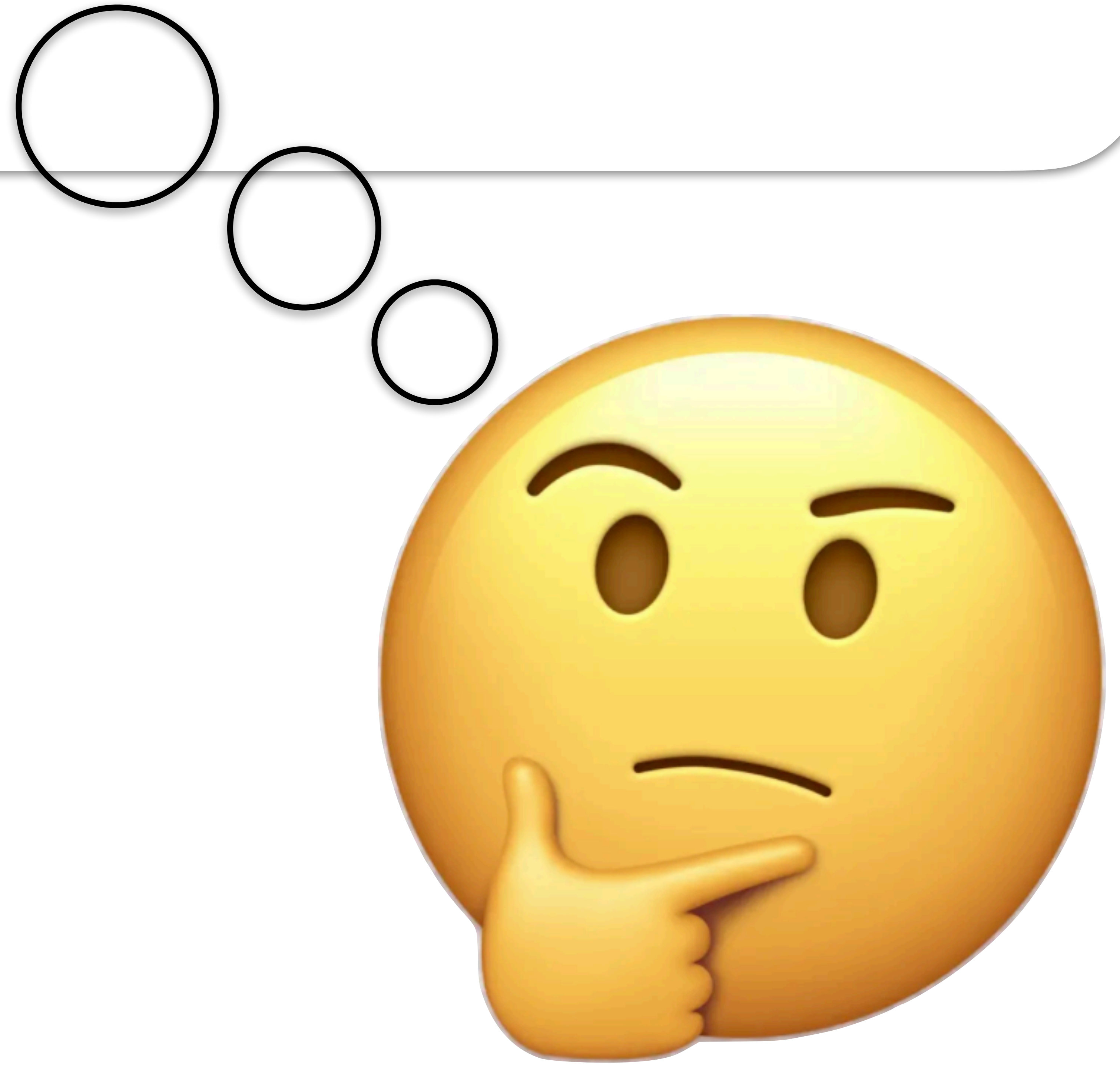
[1] Krishna, K., Wieting, J., & Iyyer, M. (2020). Reformulating Unsupervised Style Transfer as Paraphrase Generation. ArXiv, abs/2010.05700.

Experiments

- **In-Domain Evaluation:**
  - We generate a data pool with style transfer pairs from each of the 11 CDS styles to all other styles and train a GPT2-large policy using STEER.
  - For evaluation, we assess the performance of our model transferring to each of the 11 target styles with 1000 random sentences from all other styles
- **Out-of-Domain Evaluation:**
  - We evaluate the trained model from STEER on two styles **unseen** during training: the formal and informal styles from the GYAFC corpus [1]
- **Baselines:**
  - Instruction-tuned GPT3 (774M param), GPT2-large based methods: P-A-R [2] and STRAP [3]

[1] Rao, S., & Tetreault, J.R. (2018). Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer. North American Chapter of the Association for Computational Linguistics.
[2] Suzgun, M., Melas-Kyriazi, L., & Jurafsky, D. (2022). Prompt-and-Rerank: A Method for Zero-Shot and Few-Shot Arbitrary Textual Style Transfer with Small Language Models. ArXiv, abs/2205.11503.
[3] Krishna, K., Wieting, J., & Iyyer, M. (2020). Reformulating Unsupervised Style Transfer as Paraphrase Generation. ArXiv, abs/2010.05700.

How does STEER perform compared to other methods?

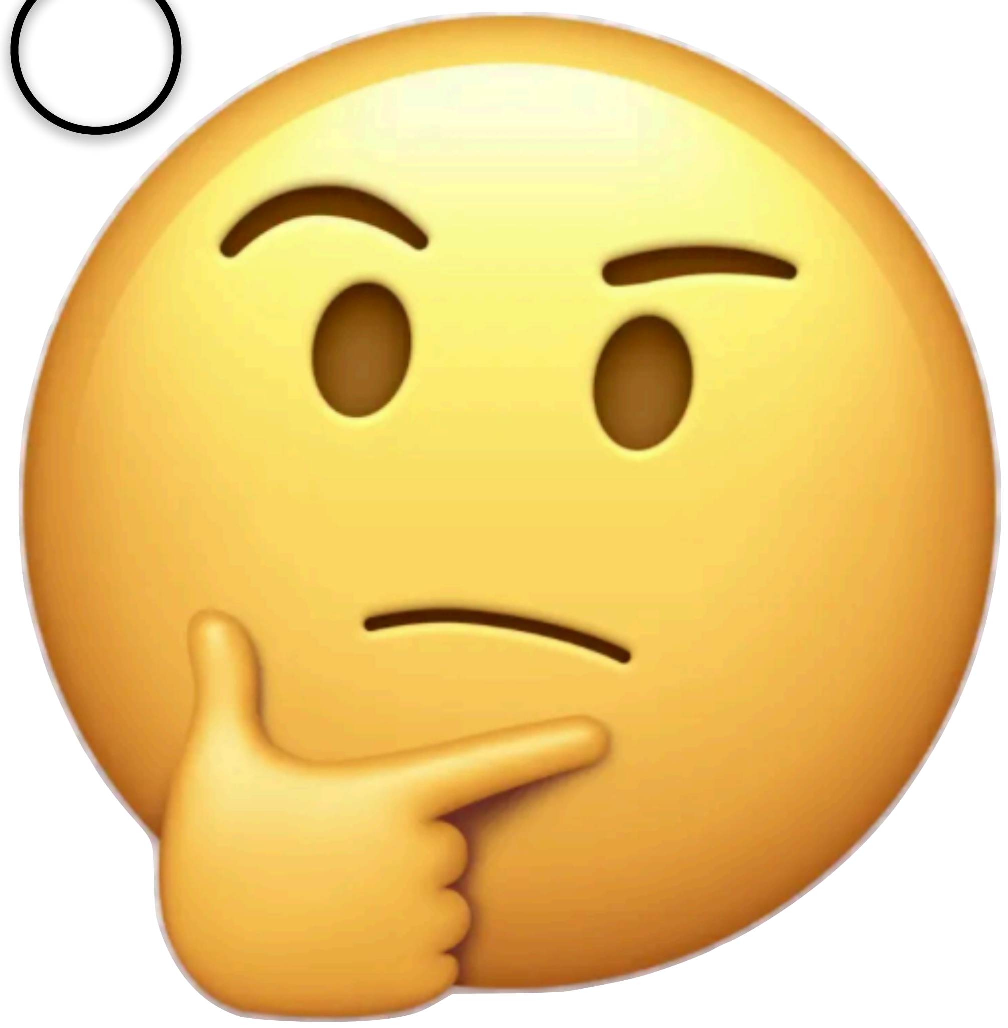


Results: In-domain

Target Style	GPT-2 Large			GPT-3 (`text-davinci-003`)			
	STEER	STRAP	P-A-R	$k=0$	$k=1$	$k=5$	$k=10$
AAE Twitter	**42.6**	7.4	3.8	23.2	11.2	<u>25.4</u>	22.7
Bible	**44.0**	<u>26.9</u>	6.6	5.2	16.0	20.2	21.0
1810-1820s	**30.2**	11.1	3.5	14.7	15.9	<u>17.4</u>	17.0
1890-1900s	**35.9**	<u>12.3</u>	4.4	8.6	9.1	10.4	10.1
1990-2000s	**42.3**	16.6	4.3	7.9	13.0	<u>17.5</u>	17.2
English Twitter	**41.2**	8.0	5.5	<u>35.0</u>	23.6	32.0	29.5
James Joyce	**20.4**	<u>11.8</u>	5.4	3.4	1.3	1.6	2.6
Song Lyrics	**33.3**	<u>20.2</u>	7.7	12.2	15.4	11.2	13.2
Romantic Poetry	**20.4**	<u>15.7</u>	2.8	1.1	3.4	6.2	4.9
Shakespeare	**13.6**	9.1	2.5	9.6	<u>10.0</u>	9.7	9.7
Switchboard	**52.9**	<u>21.1</u>	1.7	0.1	0.3	5.3	13.7
Overall	**34.3**	14.6	4.4	11.0	10.8	14.3	14.7

Table 1: Comparison of 11-way style transfer on the CDS dataset measured by aggregate score $\mathcal{V}$ with different methods, including STRAP (Krishna et al., 2020) and P-A-R (Suzgun et al., 2022), using GPT-2 Large (774M), and GPT-3 (175B). **Bold** and <u>underline</u> denote the highest and the second-highest score respectively in each row.

What about for styles that are out-of-domain?



Results: Out-of-domain

	GPT2-Large						GPT-3 (`text-davincii-003`)							
	STEER		STRAP		P-A-R		$k=0$		$k=1$		$k=5$		$k=10$	
Target Style	**Inf.**	**For.**	**Inf.**	**For.**	**Inf.**	**For.**	**Inf.**	**For.**	**Inf.**	**For.**	**Inf.**	**For.**	**Inf.**	**For.**
AAE Twitter	**44.0**	**47.7**	18.7	13.2	25.6	10.6	31.7	29.2	21.5	17.9	30	28.8	30.2	27.6
Bible	**36.1**	**38.8**	22	22.9	0.3	1.6	4.3	4.4	15.7	15.9	18.0	19.0	19.8	19.5
1810-1820s	**26.3**	**29.5**	5.9	10.0	1.2	4.7	12.4	15.6	14.3	16.9	17.6	21.6	16.9	20.1
1890-1900s	**33.5**	**34.7**	10.0	13.4	4.4	11.0	9.9	11.8	13.9	13.8	14.6	14.4	13.8	13.3
1990-200s	**50.2**	**56.2**	22.6	32.1	11.8	31.4	16.7	20.7	28.5	32.5	31.5	34.7	28.4	32.8
English Twitter	**46.1**	**54.1**	20.1	22.1	32.4	33.5	37.4	41.8	30.1	29.5	34.9	36.4	32.5	35.0
James Joyce	**22.3**	**22.8**	10.9	13.2	3.2	7.9	2.9	3.3	2.7	2.3	3.1	2.5	3.3	2.8
Song Lyrics	**42.6**	**40.5**	22.1	23.2	10.3	12.4	19.3	12.9	22.3	18.4	19.3	16.2	24.2	20.1
Romantic Poetry	**13.5**	**12.9**	8.9	10.8	0.8	0.9	2.0	1.1	5.2	4.3	7.0	4.7	6.0	3.9
Shakespeare	11.8	11.6	11.1	10.4	1.3	4.1	12.9	15.1	**15.3**	14.7	13.4	**15.2**	13.8	**15.2**
Switchboard	**54.6**	**59.3**	29.7	35.1	5.2	6.1	0.1	0.1	0.3	0.1	9.7	13.4	15.6	23.0
Overall	**34.6**	**37.1**	16.5	18.8	8.8	11.3	13.6	14.2	15.4	15.1	18.1	18.8	18.6	19.4

Table 2: Comparison of style transfer to each of the 11 styles in the CDS dataset measured by aggregate score $\mathcal{V}$ from two out-of-domain styles from the GYAFC corpus. For. and Inf. denote the formal and informal styles respectively. **Bold** and underline denote the highest and the second-highest score respectively in each row.

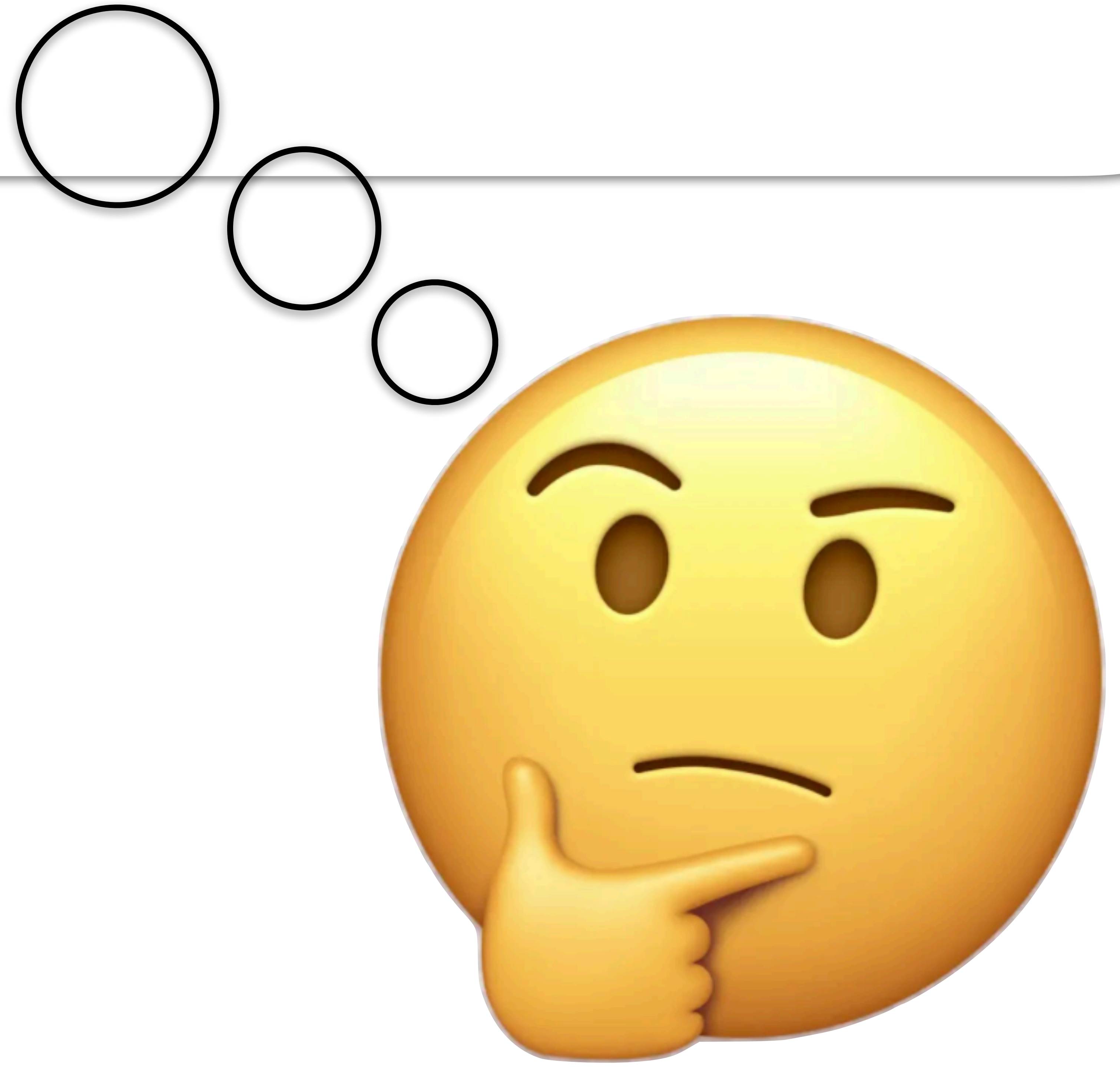
Examples

- We demonstrate examples of STEER vs other methods

Input: Can't sleep at all. Smh.	Transfer: AAE → 1990s-2000s
STEER	I mean, I can't sleep at all.
GPT-3	I am unable to obtain any rest; shaking my head in disbelief.
STRAP	I don't want to sleep.
P-A-R	Can't sleep at all, smh.
Input: Yeah one one way or another	Transfer: switchboard → rom. poetry
STEER	One way, or another, or both
GPT-3	Aye, one mayhap will find the way.
STRAP	And one way or the other
P-A-R	Yeah one one way or another
Input: In his fear, he dare not face me	Transfer: lyrics → bible
STEER	And he will not dare to face me: for fear of me is in his eyes.
GPT-3	And his fear was great, so that he could not stand before me.
STRAP	For he that is afraid of me is of me; but he that is of me is of him.
P-A-R	In fear he came and hid himself, because God was near to him

Table 3: Examples of style transfer pairs generated by STEER and other methods. GPT-3 is run with 10-shot.

Would <u>humans</u> also agree that STEER outperforms other methods?



Human Evaluation

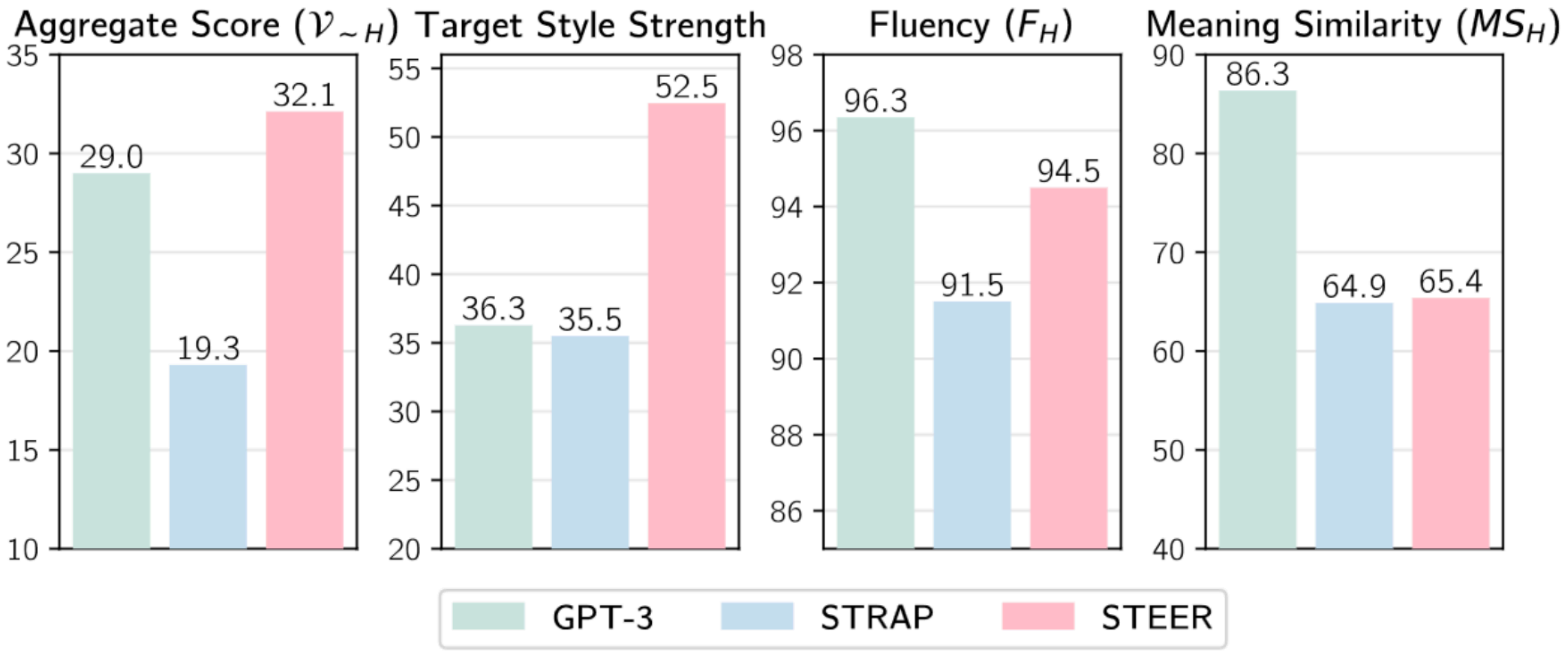


Figure 3: Style transfer quality $\mathcal{V}_{\sim\mathrm{H}}$ on CDS, averaged across all 11 styles, with fluency and meaning similarity human evaluation. **TSS** is automatically computed.[10]

Improving on Text to Text Generation Tasks

Tasks:

Style Transfer

Authorship Obfuscation

Methods:

Inference Time Only Method

Expert Distillation Method

Knowledge Distillation + Inference Time Method

Improving on Text to Text Generation Tasks

Tasks:

Style Transfer

Authorship Obfuscation

Methods:

Inference Time Only Method

Expert Distillation Method

Knowledge Distillation + Inference Time Method

StyleRemix

Interpretable Authorship Obfuscation via Distillation and Perturbation of Style Elements






Jillian Fisher*, Skyler Hallinan*, Ximing Lu, Mitchell Gordon, Zaid Harchaoui, Yejin Choi

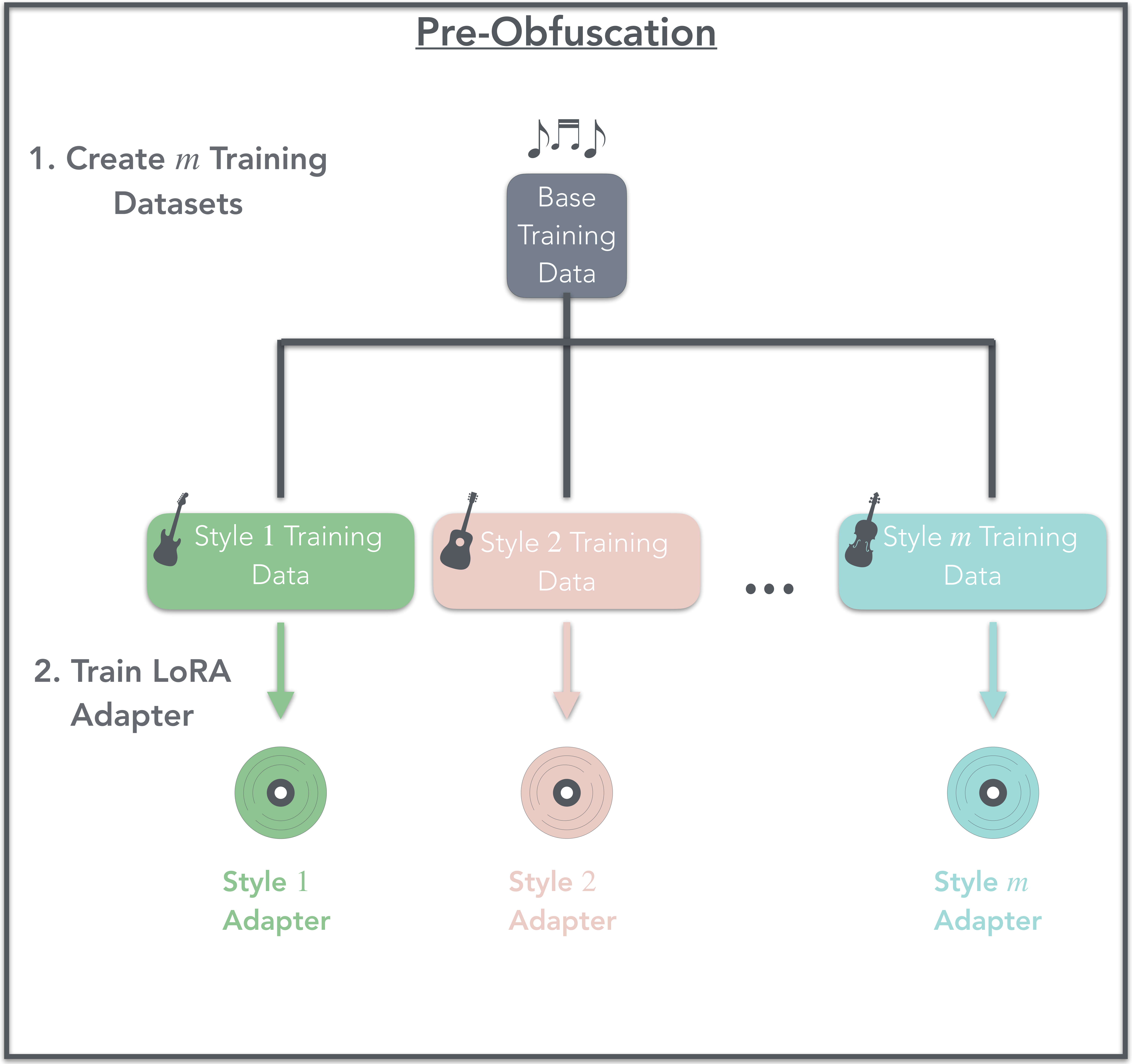
EMNLP 2024

*Co-First Authors

StyleRemix

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

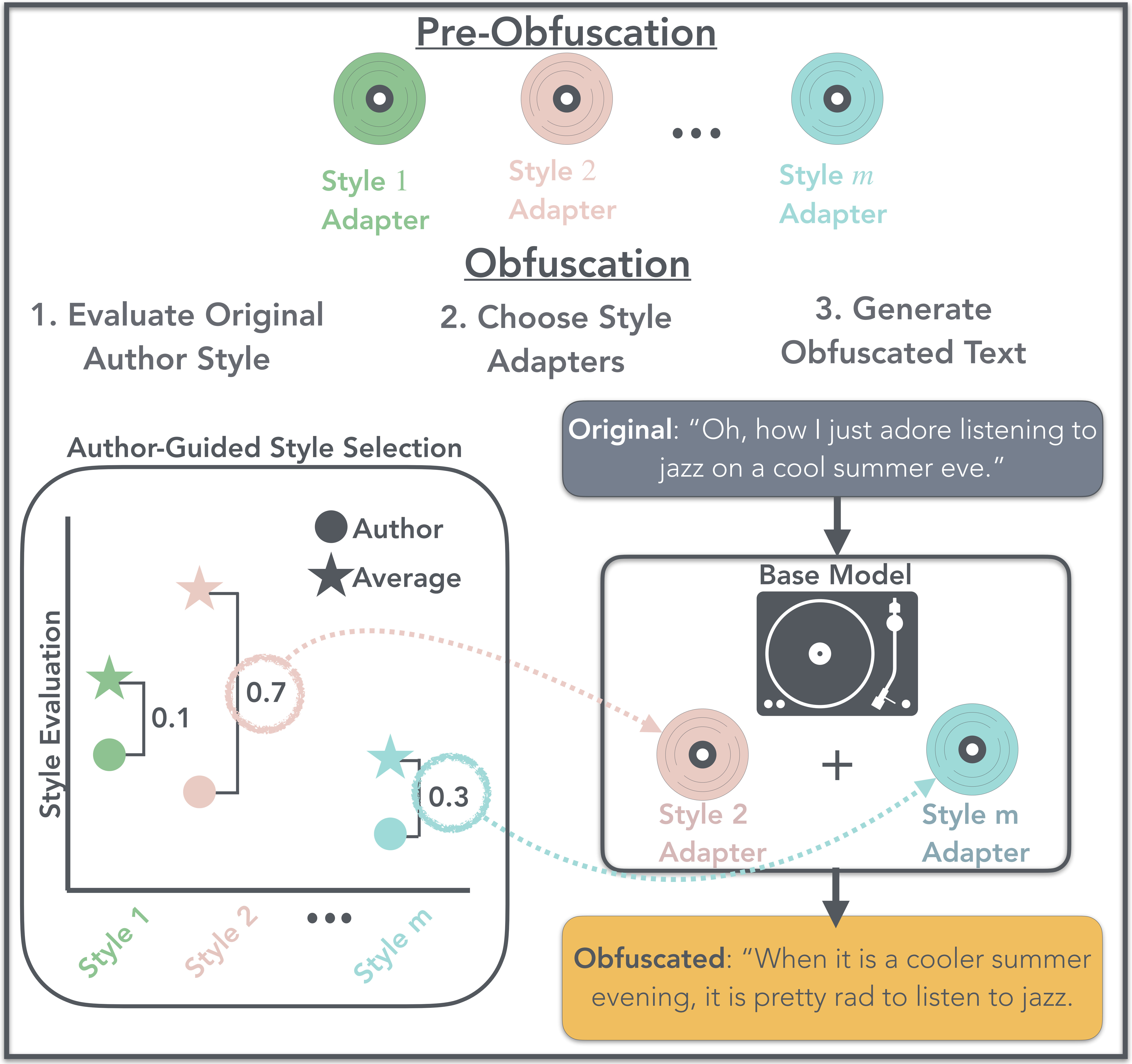
- **Pre-Obfuscation:**
  1. *Generate Training Data for each m style*
  2. *Train Low-Rank Adapters (LoRA Adapter)*



StyleRemix

- 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*




Pre-Obfuscation
Style 1 Adapter
Style 2 Adapter
Style m Adapter
Which style axis should we use??
Stylometry
NLP
Rule Based
Length
Function Words
Grade Level
Classifier — Formality
Sarcasm
Voice
Writing Intent
Classifier
Do these styles differentiate authors?
PCA Analysis of Authorship Style
PCA Dim 2
PCA Dim 1
Novels
Hemingway
Fitzgerald
Woolf
Speeches
Trump
Obama
Bush
Scholar
Scholar-H
Scholar-PP
Scholar-QQ
Blog
Blog-1
Blog-2
Blog-3
Blog-4
Blog-5

Pre-Obfuscation: Adapter Training Set

Style Axes

Length	Sarcasm
Function Words	Voice
Grade Level	Writing Intent
Formality	

Base Training Dataset

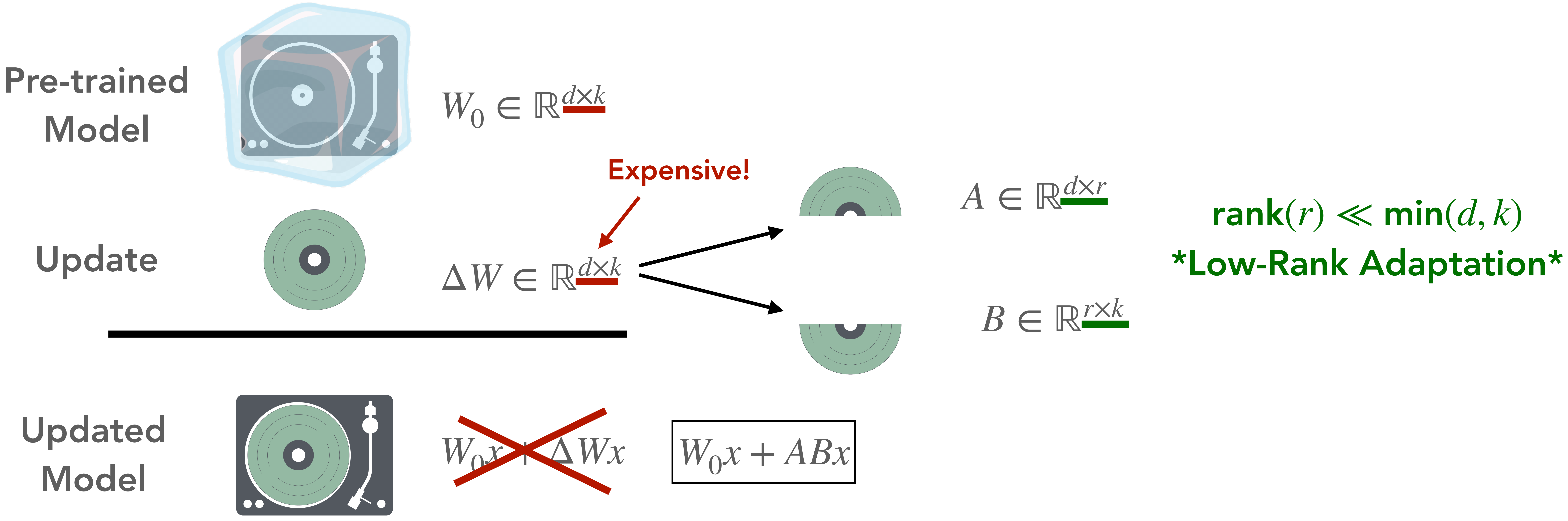
Wikipedia	Books +Plays	Blog

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-Obfuscation: Train LoRA Adapter

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

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

Original Text (Obama)

1. Evaluate Author Style

2. Choose k Style Axis (and direction)

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

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

Original Text (Obama)

1. Evaluate Author Style

2. Choose k Style Axis (and direction)

Function Words (Higher)	Formality (Lower)

3. Choose weights of style Axes

3.a) Static Weight Selection

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

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

3.b) Dynamic Weight Selection

Optimization of loss based on style axis evaluations

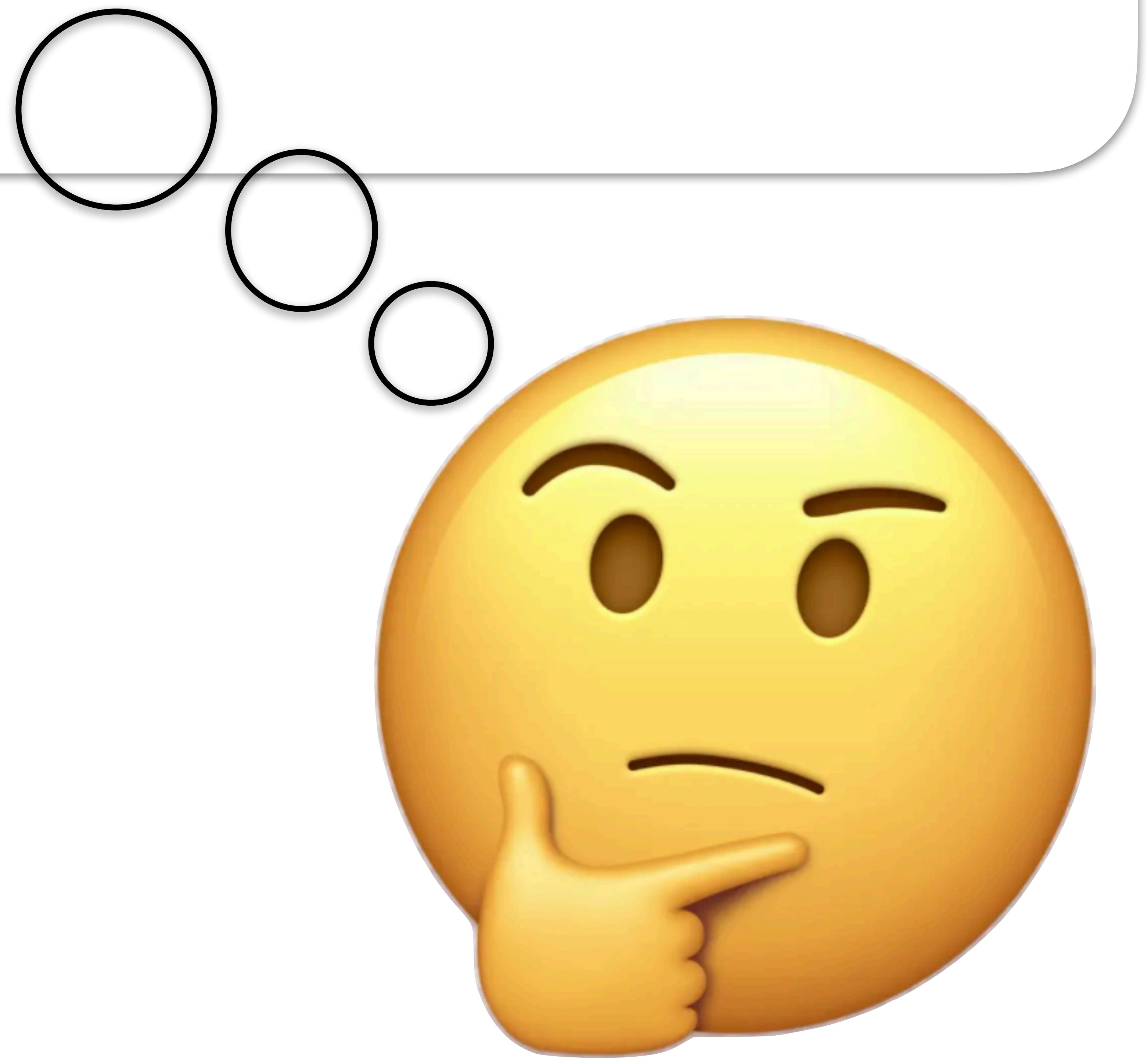
$$L = \sum_{v_i \in \{v_1, v_2\}} \begin{cases} v_i, & \text{if higher} \\ 1 - v_i, & \text{if lower} \end{cases} + \alpha \cdot f$$

v_i = Average style score of test set

f = fluency score

Obfuscation: Select Style Axes 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)
    - 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
    - Paraphrase
    - JAMDEC
    - Instruction-tuned LLMs

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 (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: *Cosine similarity of word embeddings*

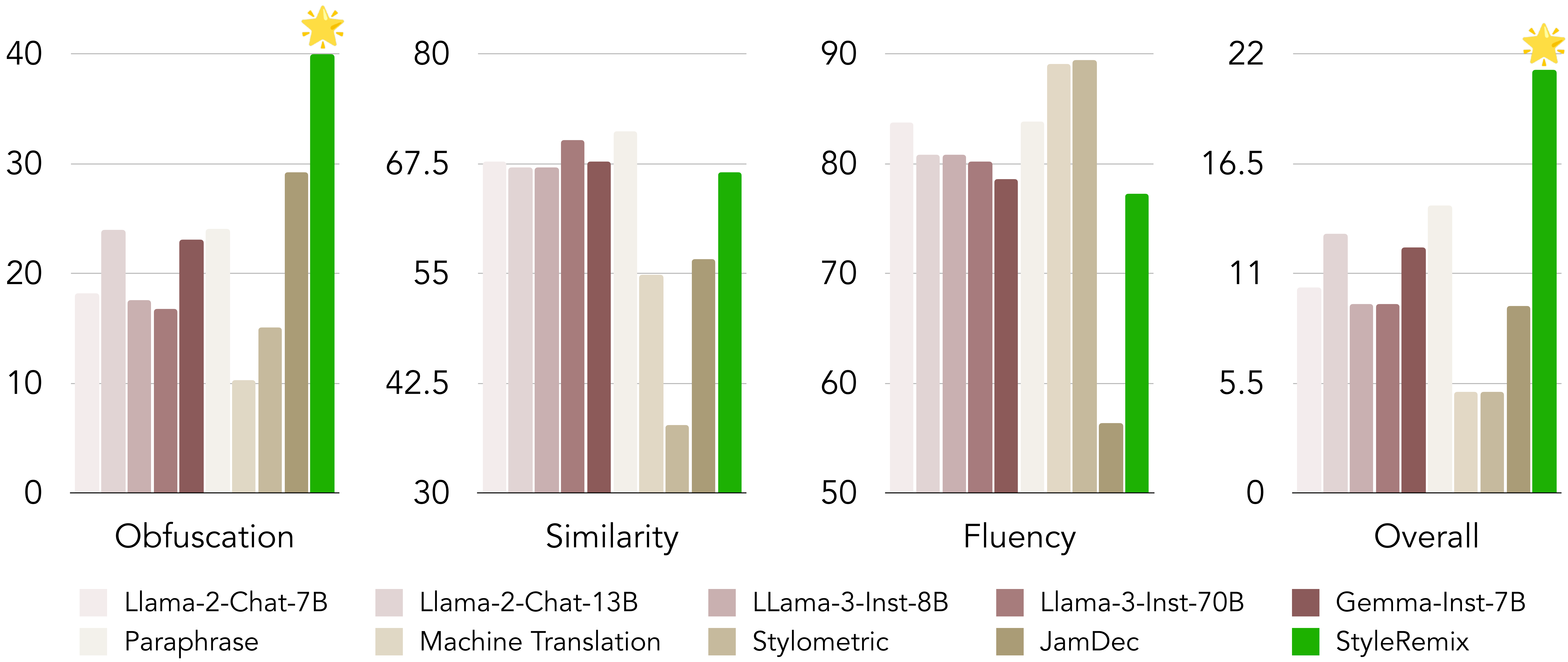
- Overall Task Score: **average** of the three metrics

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

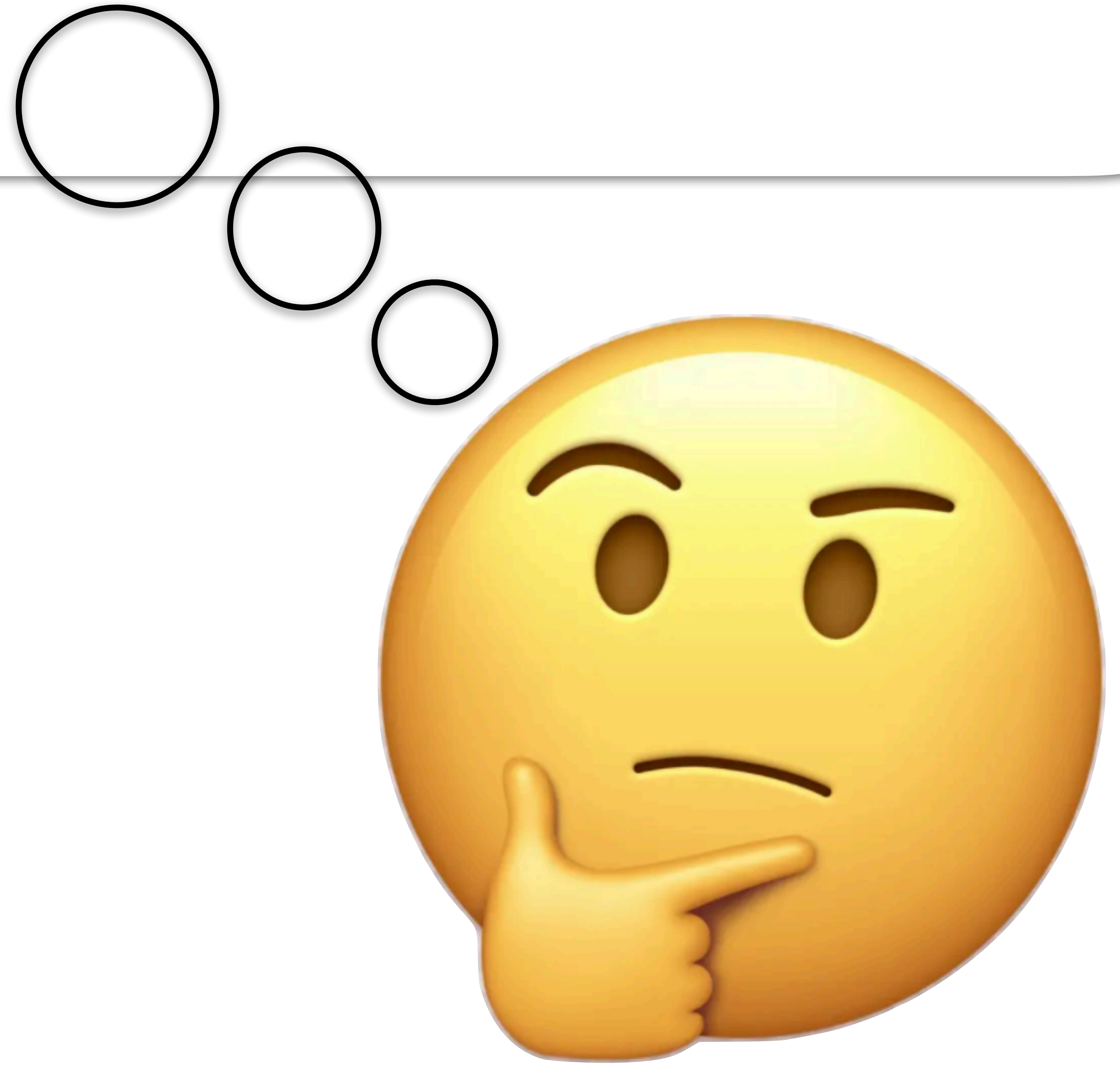
Results

AuthorMix - Blog (Auto.)

StyleRemix outperforms all baselines in obfuscation and overall quality!



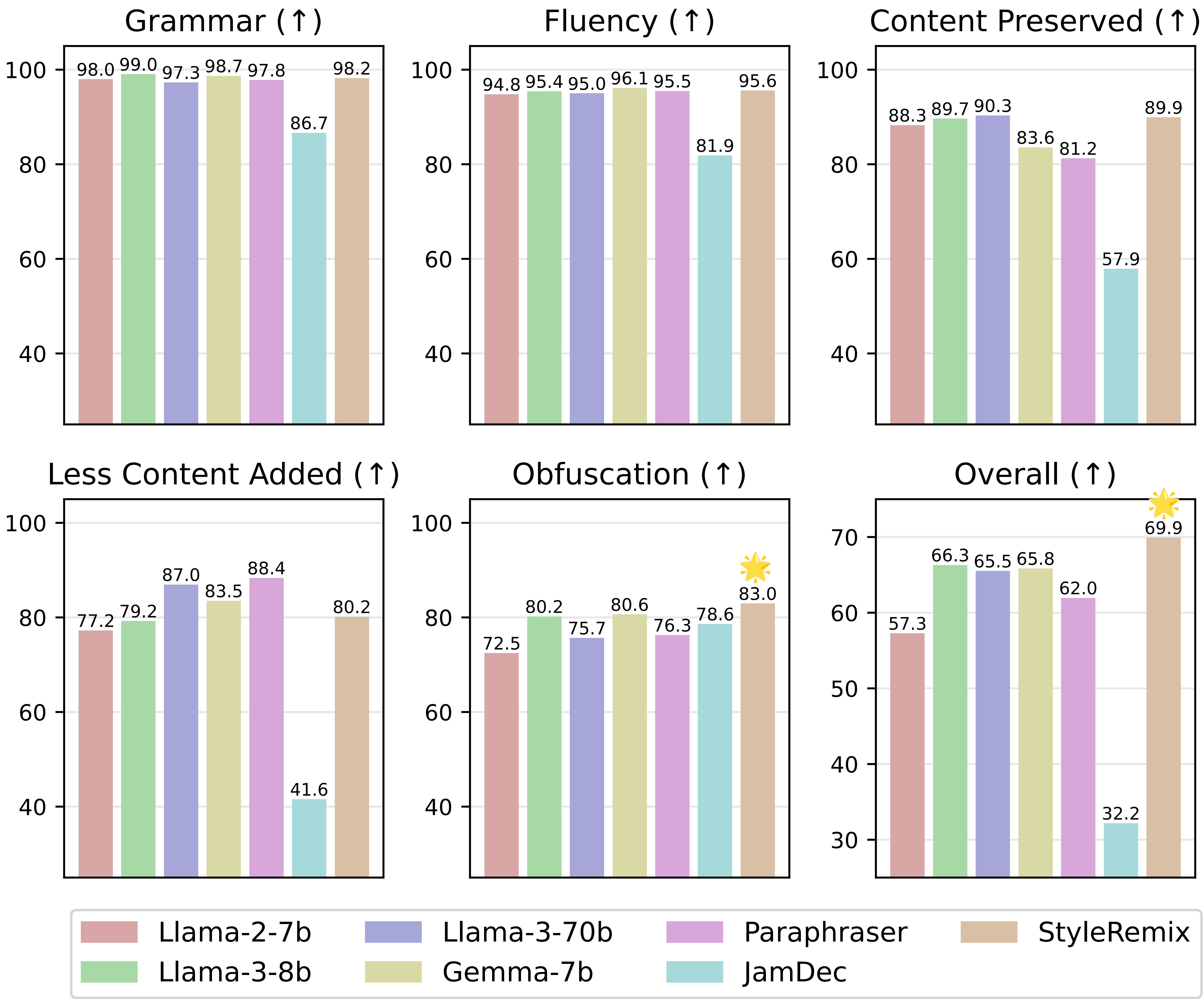
Would humans also agree that StyleRemix outperforms other methods?



Results

Human Evaluation

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



StyleRemix: Qualitative Results

Blog

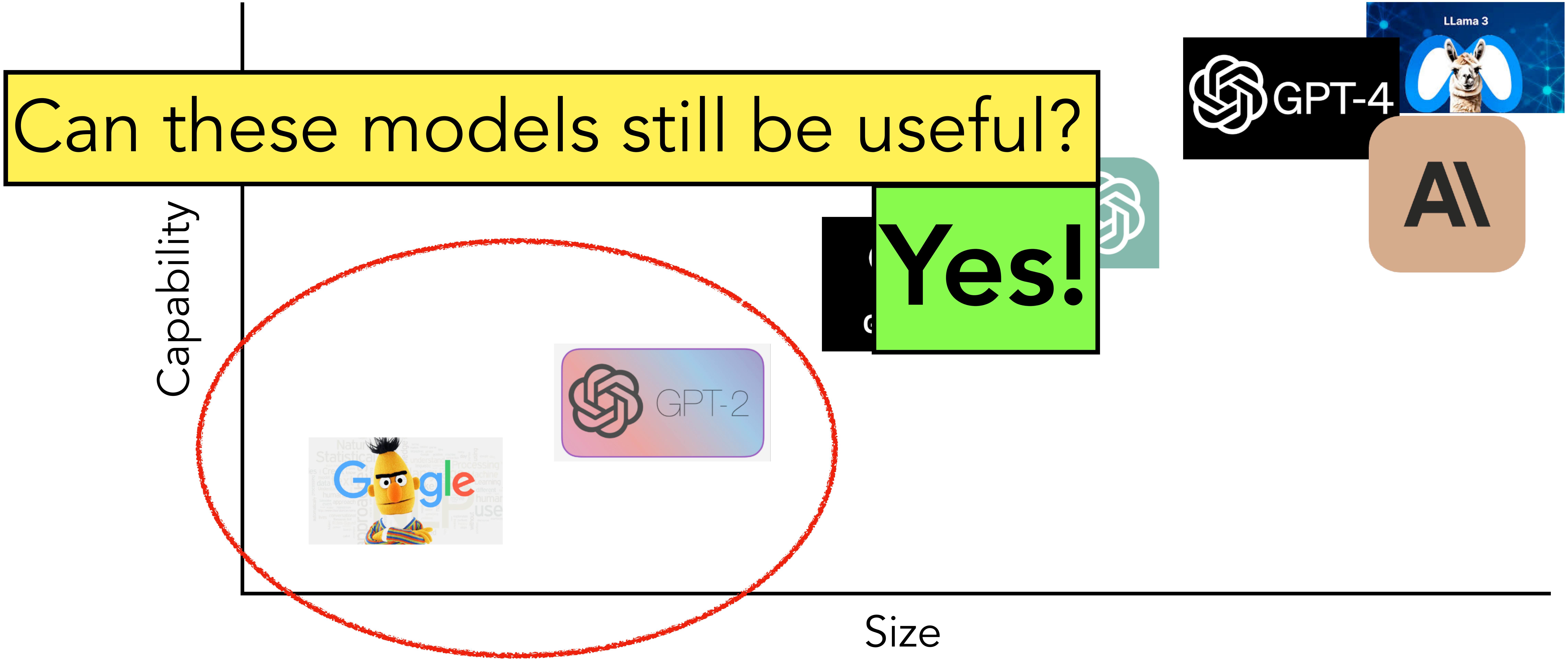
Method	Generation	
Original	100%! I was surprised, but not complaining lol. But yeah Mr. Curphey's now on crutches, hobbling around the classroom and still teaching us like his usual self.	
Paraphrase	Mr. Curphey is currently using crutches in the classroom, but he is still teaching us with his usual enthusiasm and enthusiasm. I was taken aback, but I didn't feel any remorse.	Ungrammatical
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.	Ungrammatical
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.	Ungrammatical
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.	Ungrammatical
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 ac- knowledge 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.usual style of instruction, significantly.	✓✓✓

Language Model Scaling



Takeways

- Small models can be incredibly capable!
  - …with thoughtful inference time algorithms
  - …with high-quality data curation (also from small models!)
  - …with plug-and-play inference-time adapters
- Why small models?
  - Accessibility
  - Customizability
  - Cheaper training and inference
- Let’s keep innovating beyond purely scale!

JAMDEC: Unsupervised Authorship Obfuscation using Constrained Decoding over Small Language Models

Jillian Fisher, Ximing Lu, Jaehun Jung, Liwei Jiang, Zaid Harchaoui, Yejin Choi

Findings of NAACL, 2024.

🐂 STEER: Unified Style Transfer with Expert Reinforcement

Skyler Hallinan, Faeze Brahman, Ximing Lu, Jaehun Jung, Sean Welleck, and Yejin Choi

Findings of EMNLP, 2023. Presented at NILLI 2023.

💿 StyleRemix

Interpretable Authorship Obfuscation via Distillation and Perturbation of Style Elements

Jillian Fisher*, Skyler Hallinan*, Ximing Lu, Mitchell Gordon, Zaid Harchaoui, Yejin Choi

EMNLP 2024

*Co-First Authors

Thank You! ✨

JAMDEC

https://arxiv.org/abs/2402.08761

STEER

https://arxiv.org/abs/2408.15666v1

StyleRemix

https://arxiv.org/abs/2408.15666v1

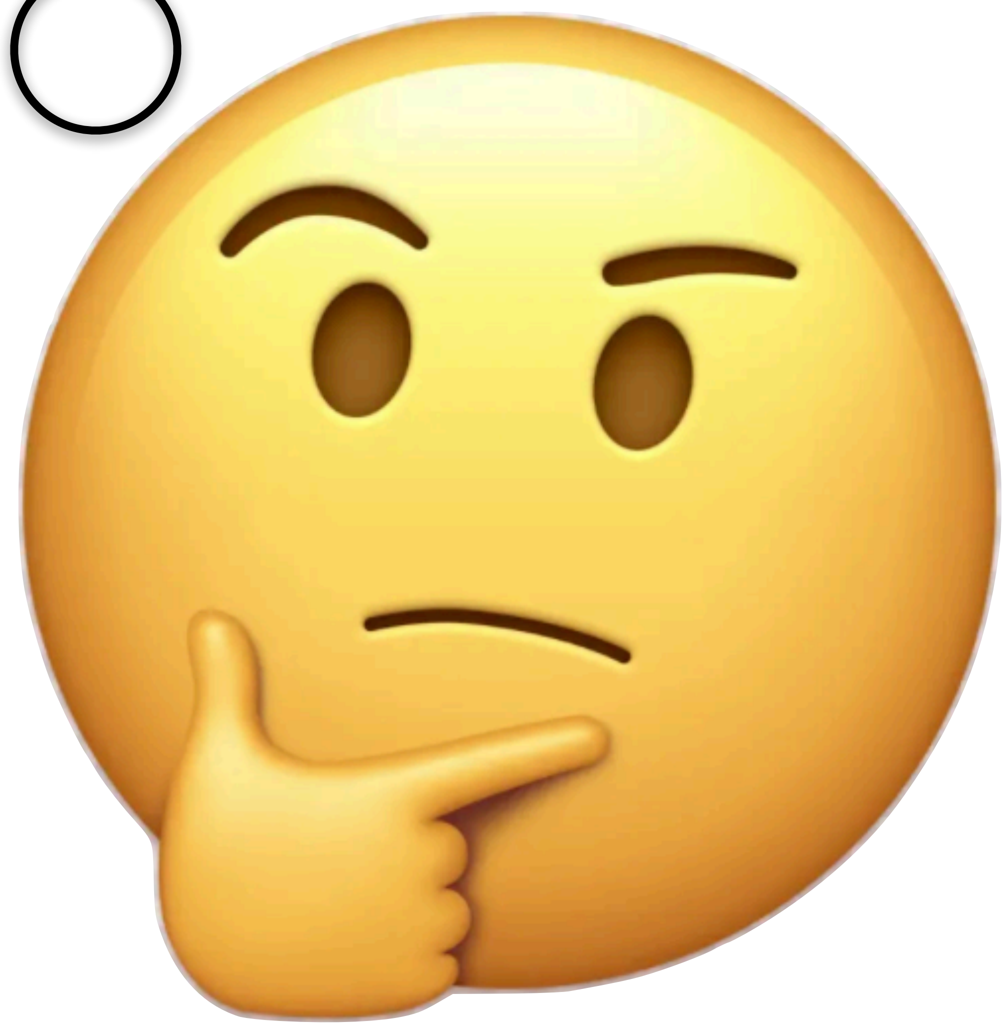
Contact Jillian Fisher & Skyler Hallinan at jrfish@uw.edu and shallina@usc.edu

Appendix

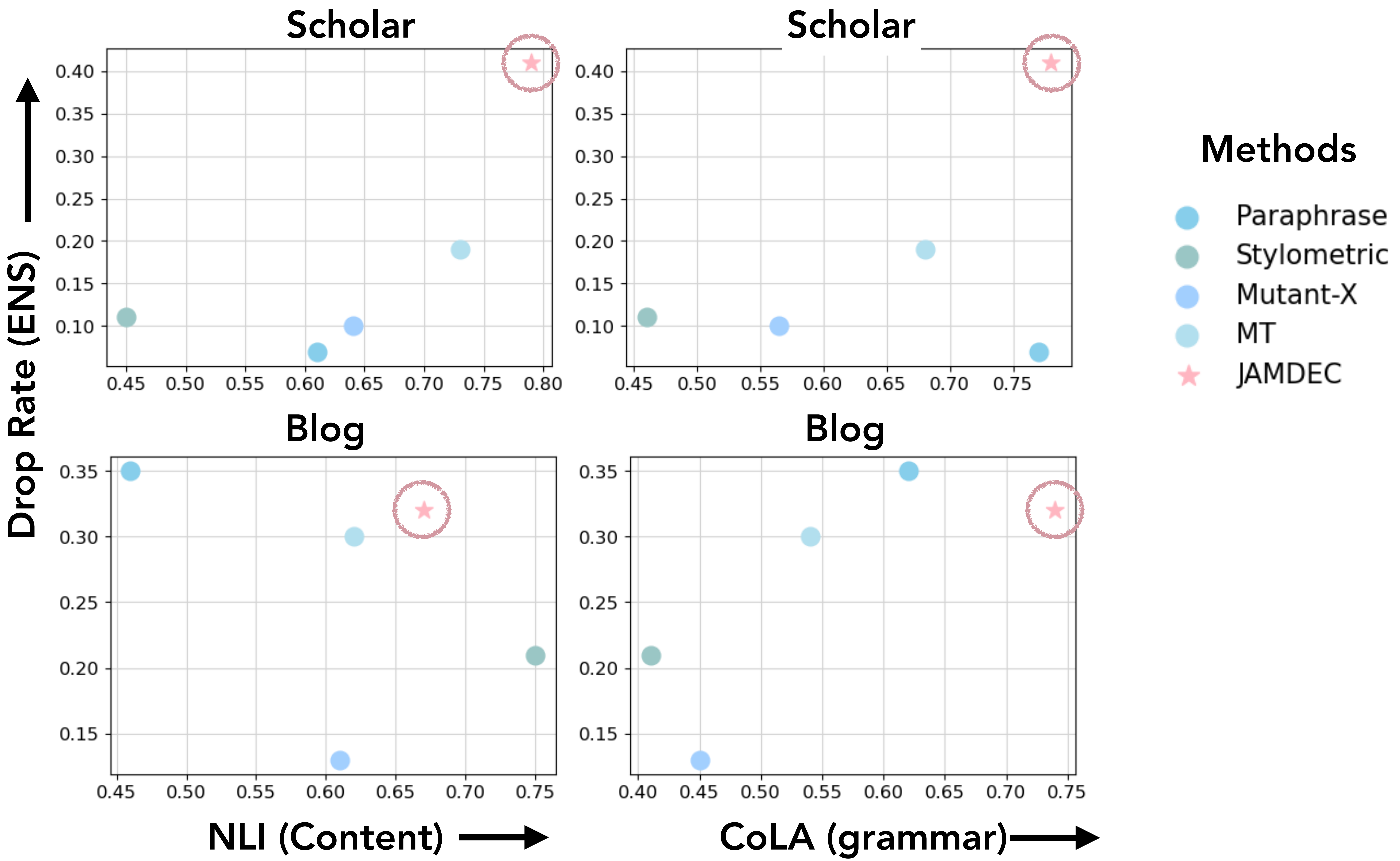
Appendix

Extra JAMDEC Results

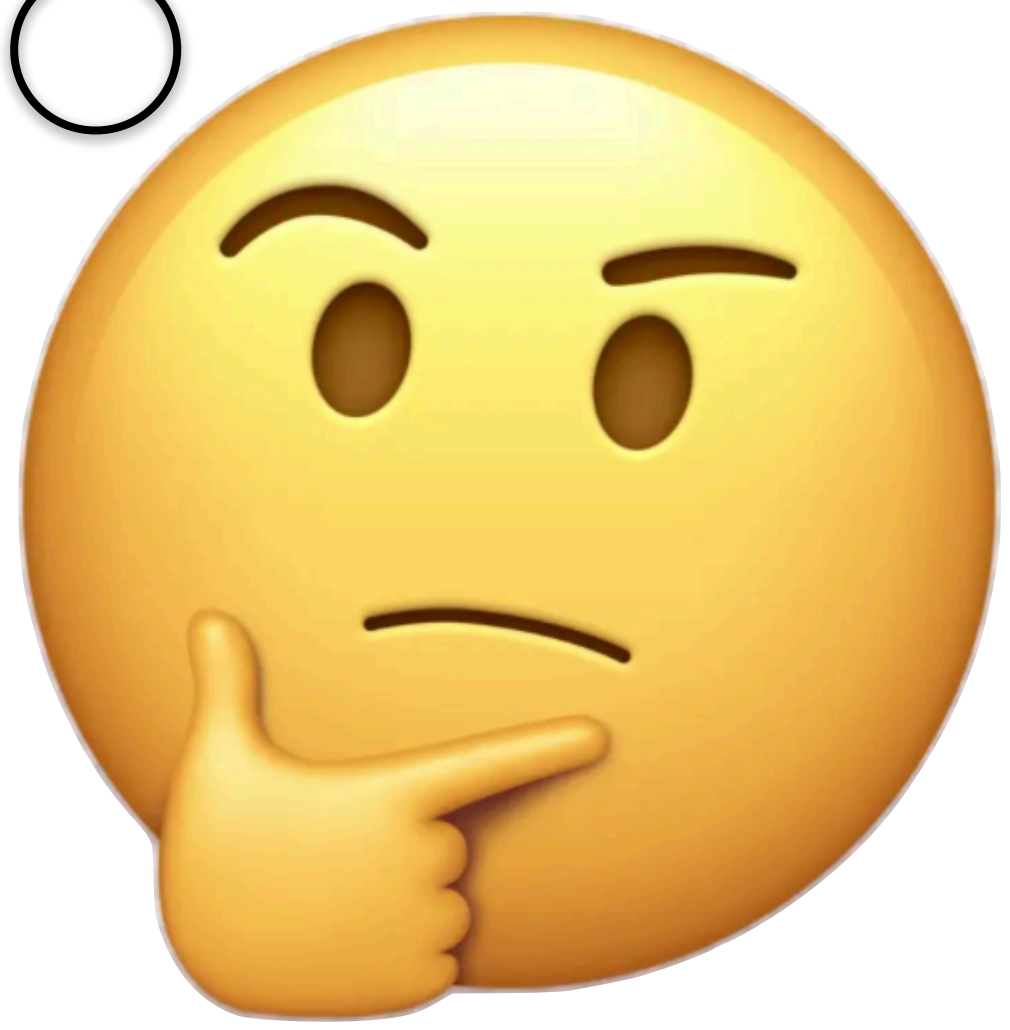
It seems like there might be a tradeoff between obfuscation, content preservation, and fluency…



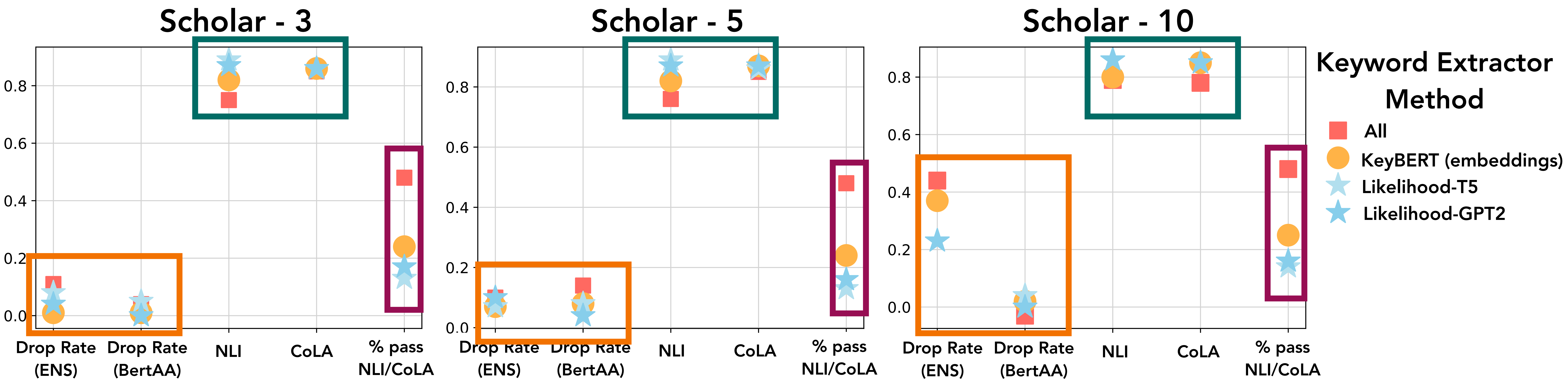
JAMDEC: Inherent Tradeoff



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**

Appendix

Extra StyleRemix Results

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		