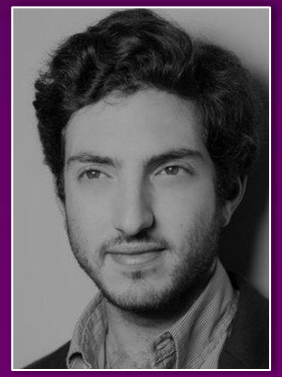
A Multifaceted Framework to Evaluate Evasion, Content Preservation, and Misattribution in Authorship Obfuscation Techniques

Malik Altakrori
McGill/Mila

Thomas Scialom
Meta AI

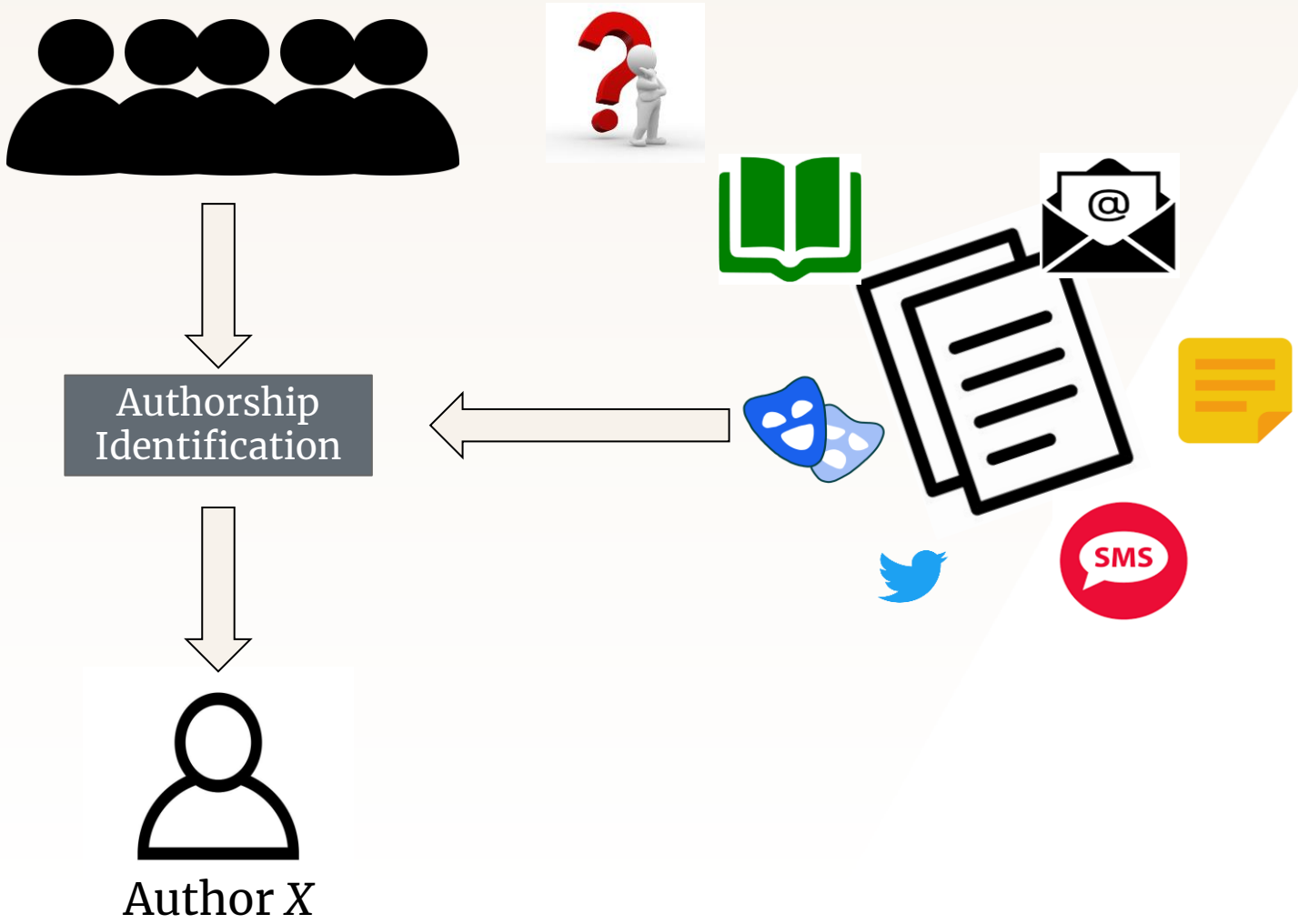
Benjamin Fung
McGill/Mila

Jackie Cheung
McGill/Mila

McGill Mila

EMNLP 2022, Abu Dhabi - UAE

Background: Authorship Identification



Applications of Authorship Identification

Potentially useful:

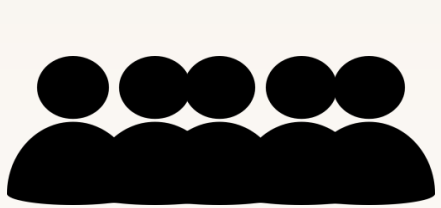
- Literature
  - Shakespeare vs. Marlow[1]
- Forensics
  - Hate speech.
  - Threatening messages.

Potentially harmful:

- Author of a negative scientific review.
- Prevent freedom of speech.

[1] https://www.theguardian.com/culture/2016/oct/23/christopher-marlowe-credited-as-one-of-shakespeares-co-writers

Identification vs. Obfuscation



Identification Technique

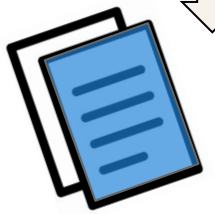
Author *X*

Obfuscation tools:

- Generic tools
- Obfuscation-specific tools

Original

Obfuscation Technique



Obfuscated

Authorship Identification

Authorship Obfuscation

Contribution

- We re-evaluate existing obfuscation techniques using a suite of measures.

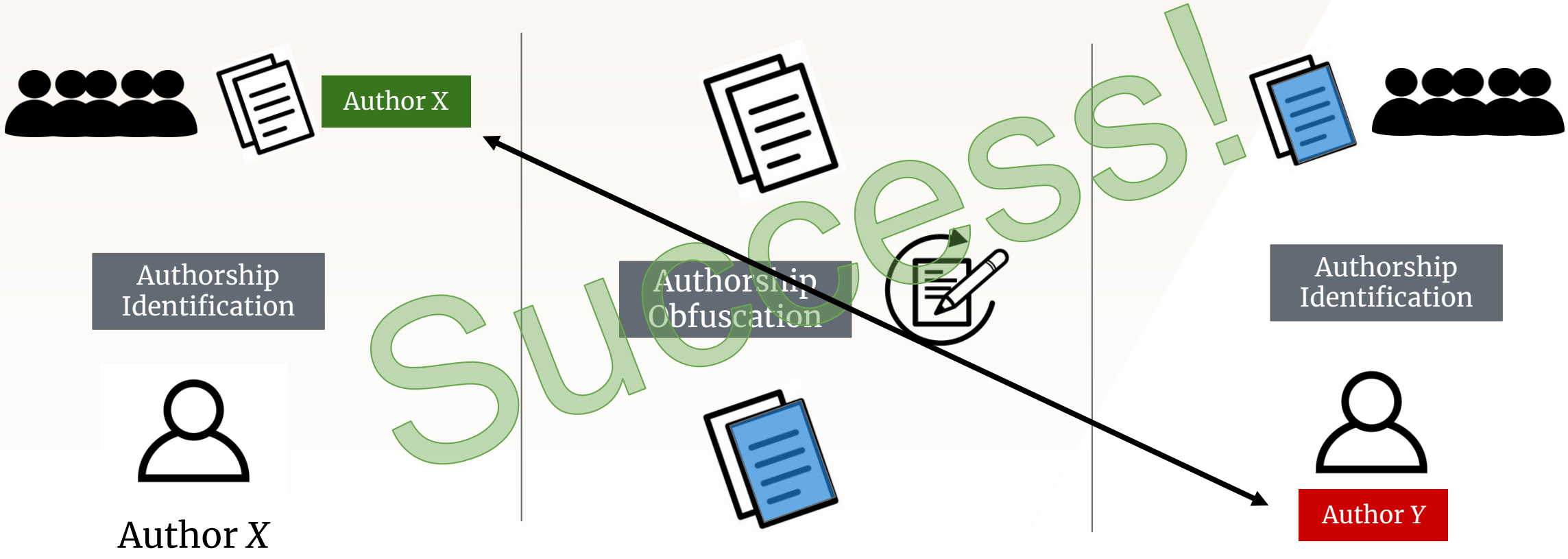
- Three dimensions:
  - Evade detection (Safety)
  - Convey the same message (Content preservation)
  - Do not implicate others (Fairness)

Question-Answering!

New!

Evasion

To evade detection by an identification technique.



Before Obfuscation

After Obfuscation

Content Preservation

To convey the same information as the original text.

Before Obfuscation

- EMNLP is in the **UAE**.
- Lyana took the candy **jar**.
- Hashem went to a **conference**.
- Omar **got** a scholarship.

After Obfuscation

- EMNLP is in the **United Arab Emirates**.
- Lyana took the candy **b**ar.
- Hashem went to a **journal**.
- Omar **played** a scholarship?

Misattribution Harm

The side-effect of evading detection.

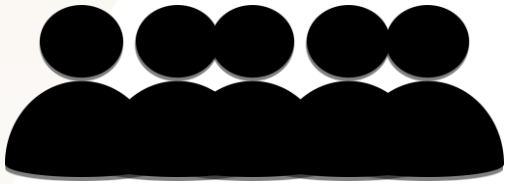
Before Obfuscation

After Obfuscation

You

Success!

Someone else

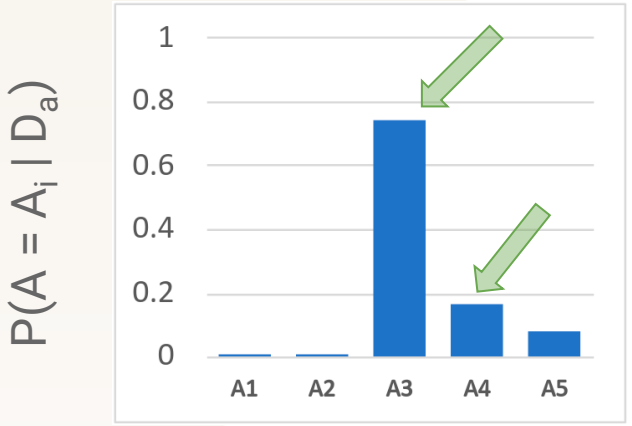


- Friend
- Classmate
- Colleague

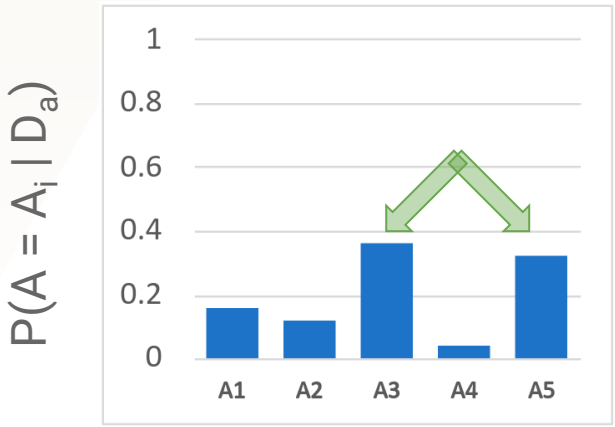
Characterizing Misattribution

- **Confidence** in the outcome of the authorship identification task.

- We can measure this using **entropy.**
  - Higher entropy -> Lower confidence



Case 1



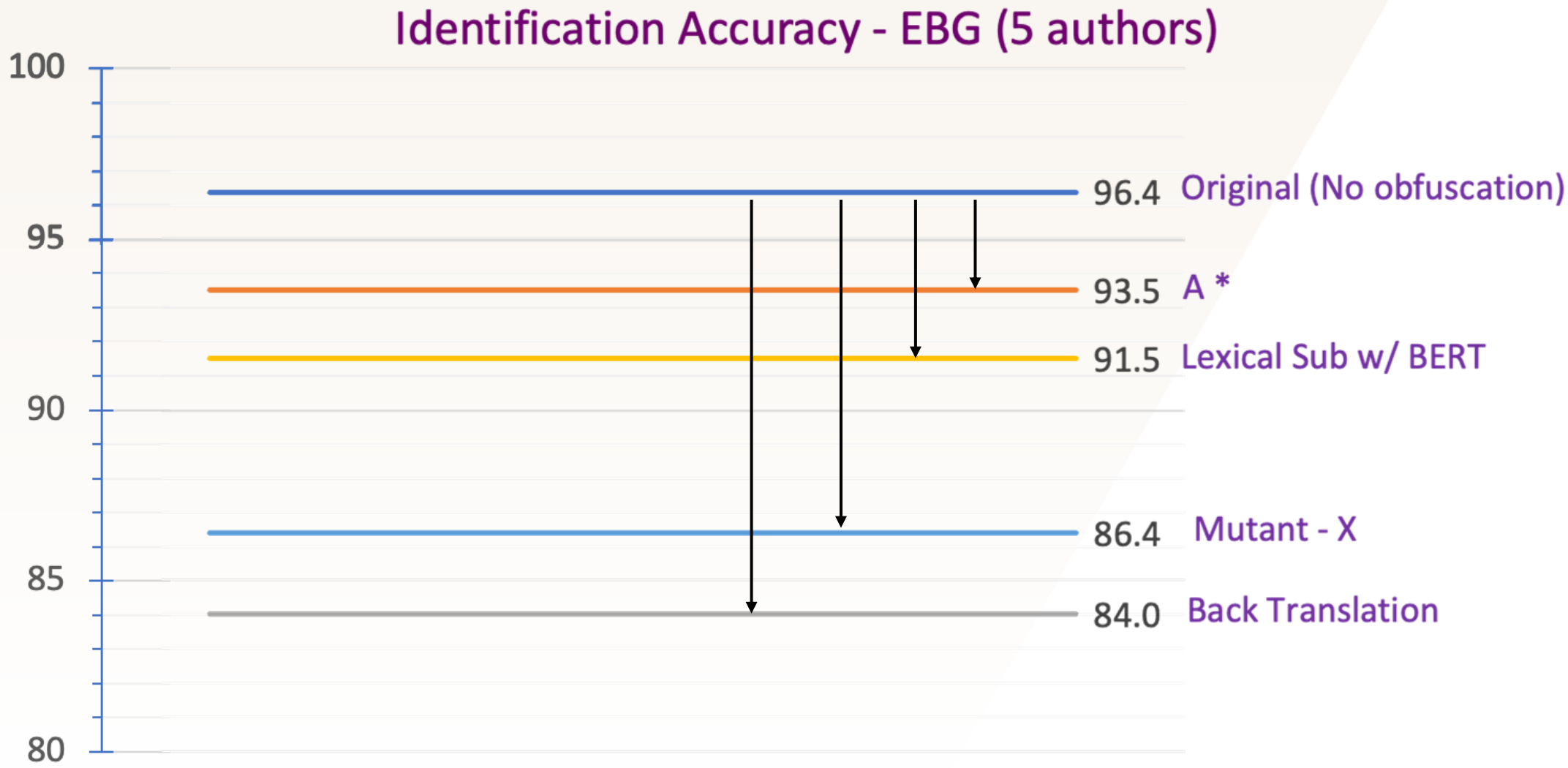
Case 2

Obfuscation Techniques

- **Generic tools**
  - Lexical Substitution with BERT (Mansoorizadeh et al., 2016)
  - Back Translation (Meta AI M2M-100) (Schwenk et al., 2021)

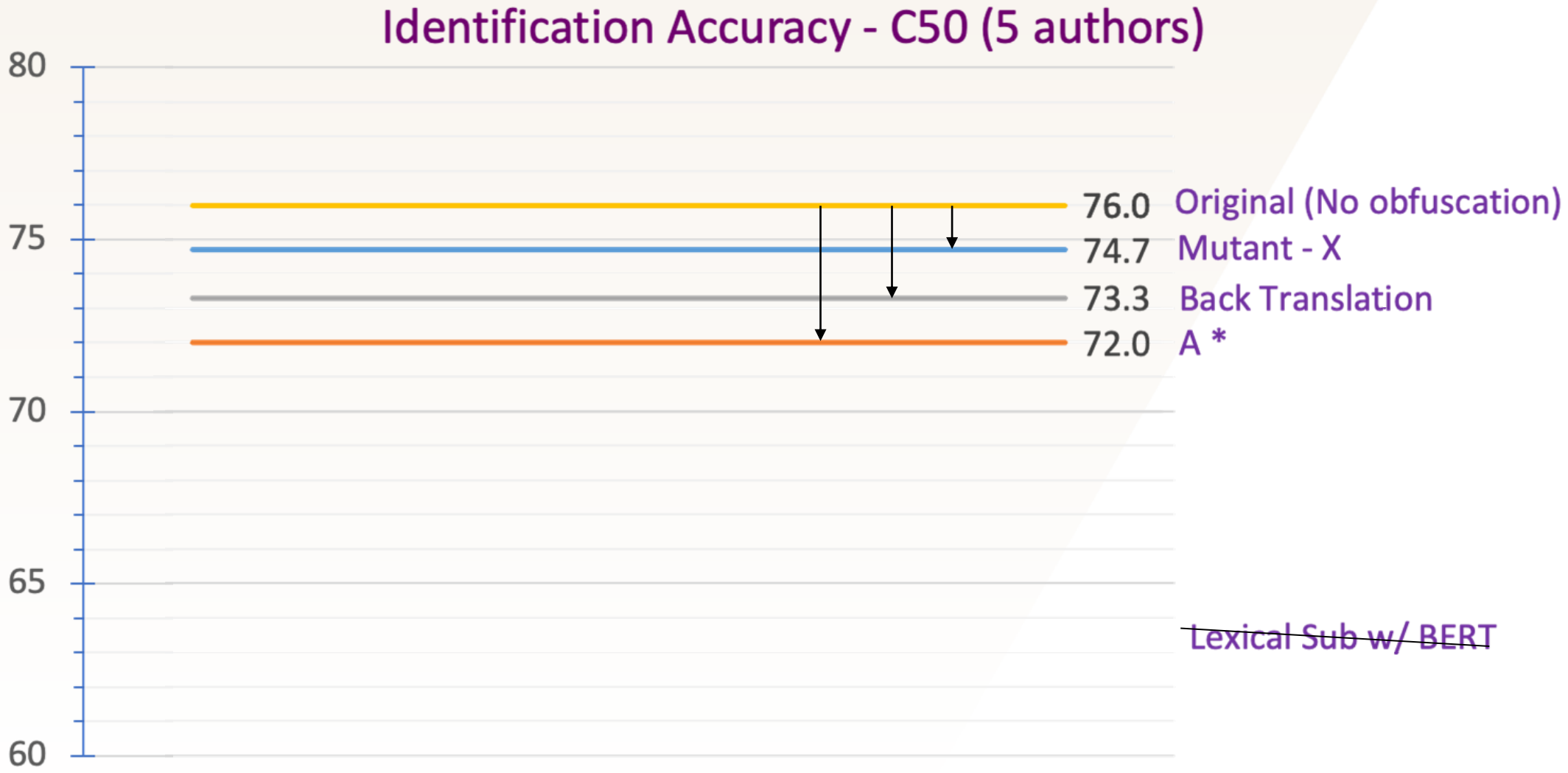
- **Obfuscation-specific tools**
  - Mutant-X (Mahmood et al., 2019)
  - Heuristic Obfuscation Search (A*) (Bevendorff et al., 2019)

Evasion (results)



☞ **Identification:** Masking (Stamatatos, E. 2018) followed by character n-grams as features, and a linearSVM classifier.

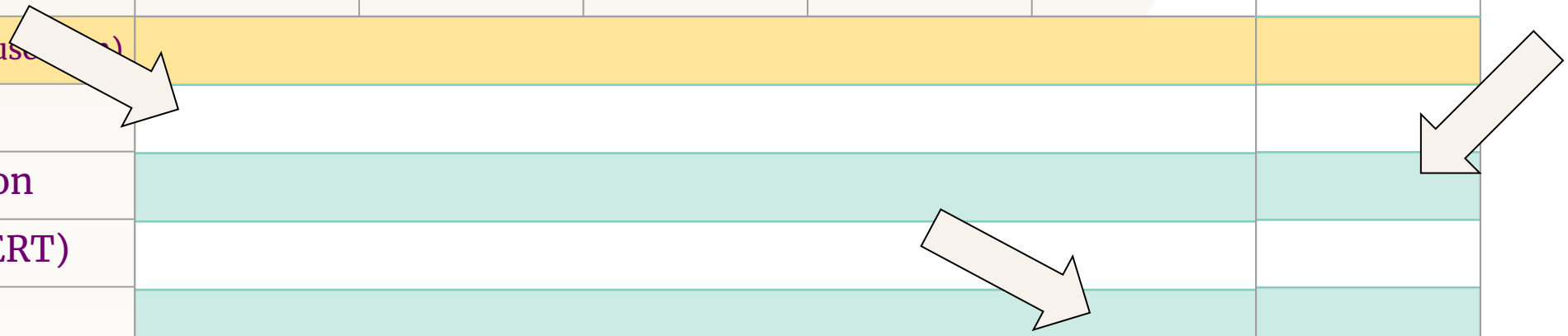
Evasion (results)



☞ **Identification:** Masking (Stamatatos, E. 2018) followed by character n-grams as features, and a linearSVM classifier.

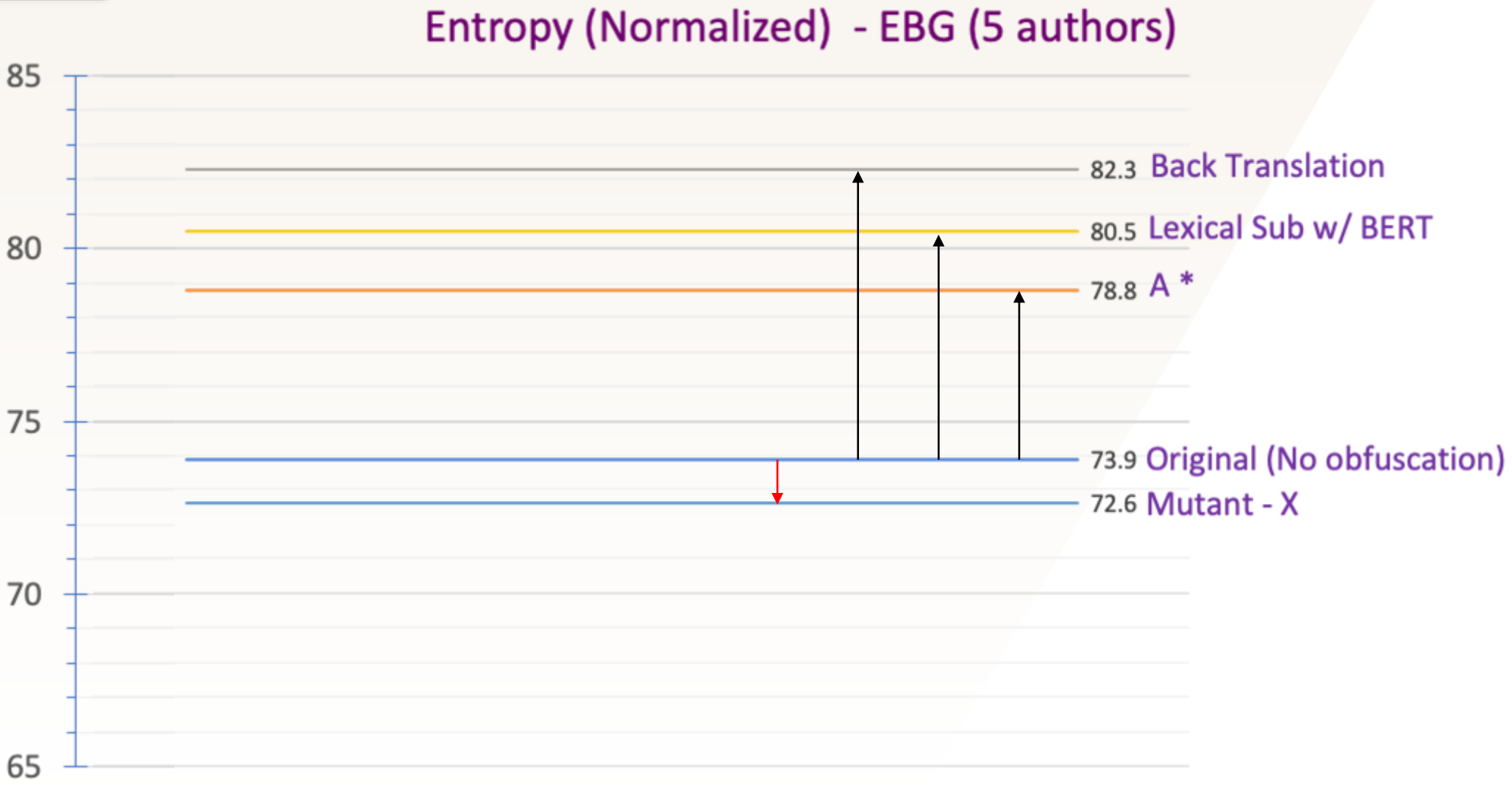
Content Preservation (results)

	Rouge-1	Rouge-2	Rouge-L	BLEU	METEOR	QuestEval
Original (no obfus[illegible])						
A*						
Back Translation						
Lexical Sub (BERT)						
Mutant-X						

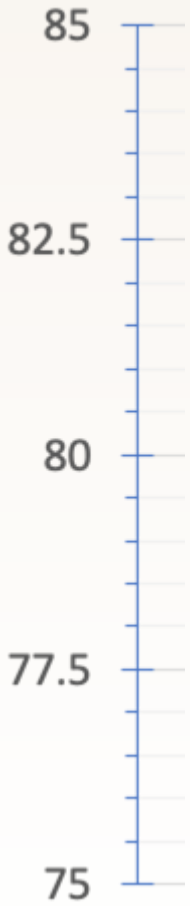


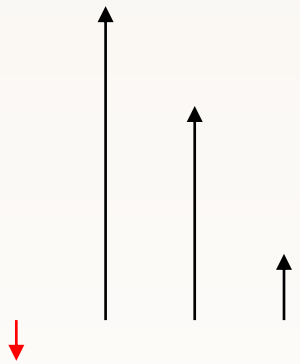
* Average score over 100, randomly sampled sentences from the EGB dataset.

Misattribution



Misattribution





In conclusion

- We need to use SOTA NLP evaluation techniques which are changing rapidly.
- Current evaluation metric revealed new results.
- Contrast to common belief, Back translation is very competitive with the SOTA.
    - It has low misattribution *arguably* because it has been trained on various writing styles.
- Finally, ... a huge room for improvement!

Thank you very much!

- I just defended my thesis!
- Looking for a research scientist role in North America or the Gulf region.
- I work on NLP/Privacy (Style Analysis)

https://malikaltakrori.github.io/

Empirical Setup

- Datasets:
  - Extended Brennan–Greenstadt Corpus (**EBG**), Reuters Corpus Volume 1(**C50**)
  - Two configurations: **5** authors, **10** authors.
- Identification method:
  - Masking (Stamatatos, E. 2018) character n-grams as features, and a linearSVM classifier.
- Content preservation:
  - QuestEval (Scialom, 2021)

Empirical Setup

	C50				EBG			
Authors	**5**		**10**		**5**		**10**	
Training set								
Docs	75		150		55		110	
Docs / authors:	15	(0.0)	15	(0.0)	11	(0.0)	11	(0.0)
Avg. doc Len (W)	478	(46.4)	452	(60.8)	496	(6.1)	494	(4.8)
Avg. doc Len (C)	3007	(273.1)	2861	(366.9)	3157	(24.0)	3120	(41.8)
Testing set								
Docs	75		150		55		110	
Docs / authors:	15	(0.0)	15	(0.0)	7	(4.0)	6	(3.2)
Avg. doc Len (W)	480	(86.2)	479	(77.6)	496	(14.1)	497	(12.5)
Avg. doc Len (C)	3032	(567.2)	3036	(473.9)	3068	(102.7)	3046	(130.8)
Total docs	150		300		90		169	

Table 8: Corpora statistics. (Mean and SD)