# TEXT EMBEDDING INVERSION SECURITY FOR MULTILINGUAL LANGUAGE MODELS (ACL 2024)

Yiyi Chen, Heather Lent, Johannes Bjerva

Base Model [1]

black-box

encoder

encoder-

decode

Text

Text

Ad-hoc Translation (AdTrans)

Hello World

Hello World

**5 Cross-lingual Inversion Attacks** 

ME5-English

German

ME5-German

**Embeddings** 

2 Inversion Models + Ad-hoc Translation

**Corrector Model [1]** 

decode

black-box

encoder

**Machine Translation** 

BLEU

BLEU

52,5

35

17,5

0,3

0,2

0,1

English

Hallo Welt

ME5-French

German

ME5-Spanish

French

Inserting

Gaussian Noise

10<sup>-2</sup>

**Noise Level** 

**CLIRMatrix** 

(Multilingual)

10<sup>-1</sup>

ME5\_MULTI

52,5

35

17,5

x: random integer

ME5-German

Spanish

decoder

encoder

**ME5-German:** 

dataset [2]

Spanish

German

**Inversion Model trained on ME5** 

embeddings with German MTG

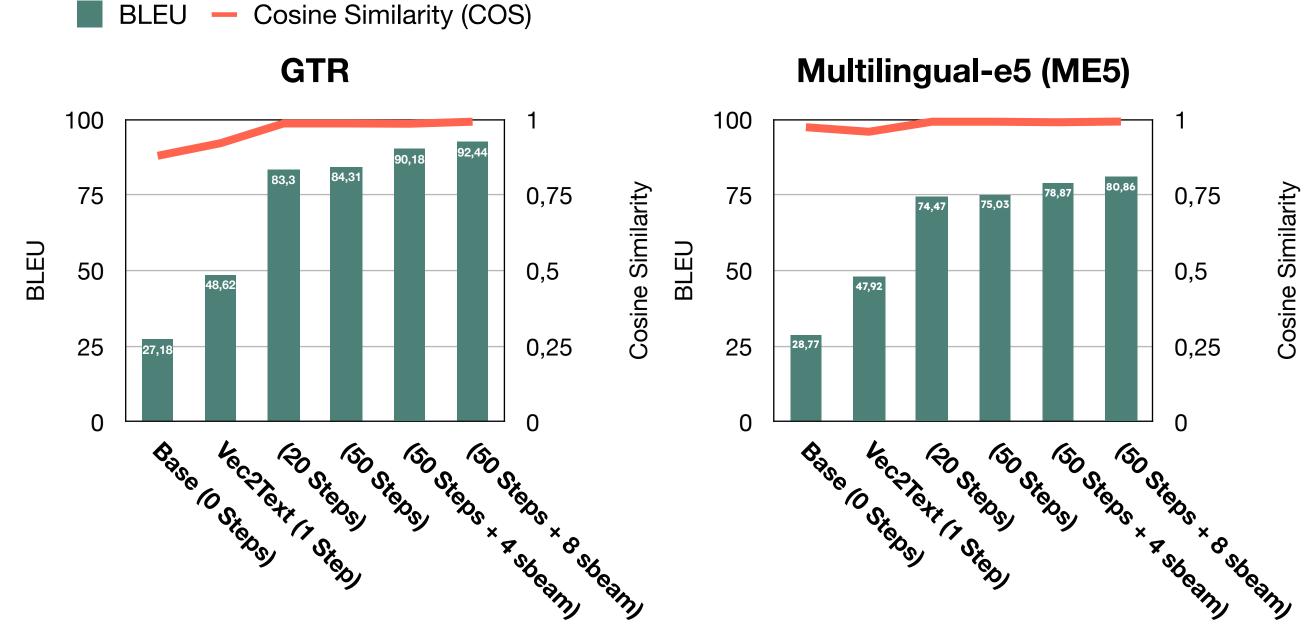
Base

Vec2Text

AdTrans

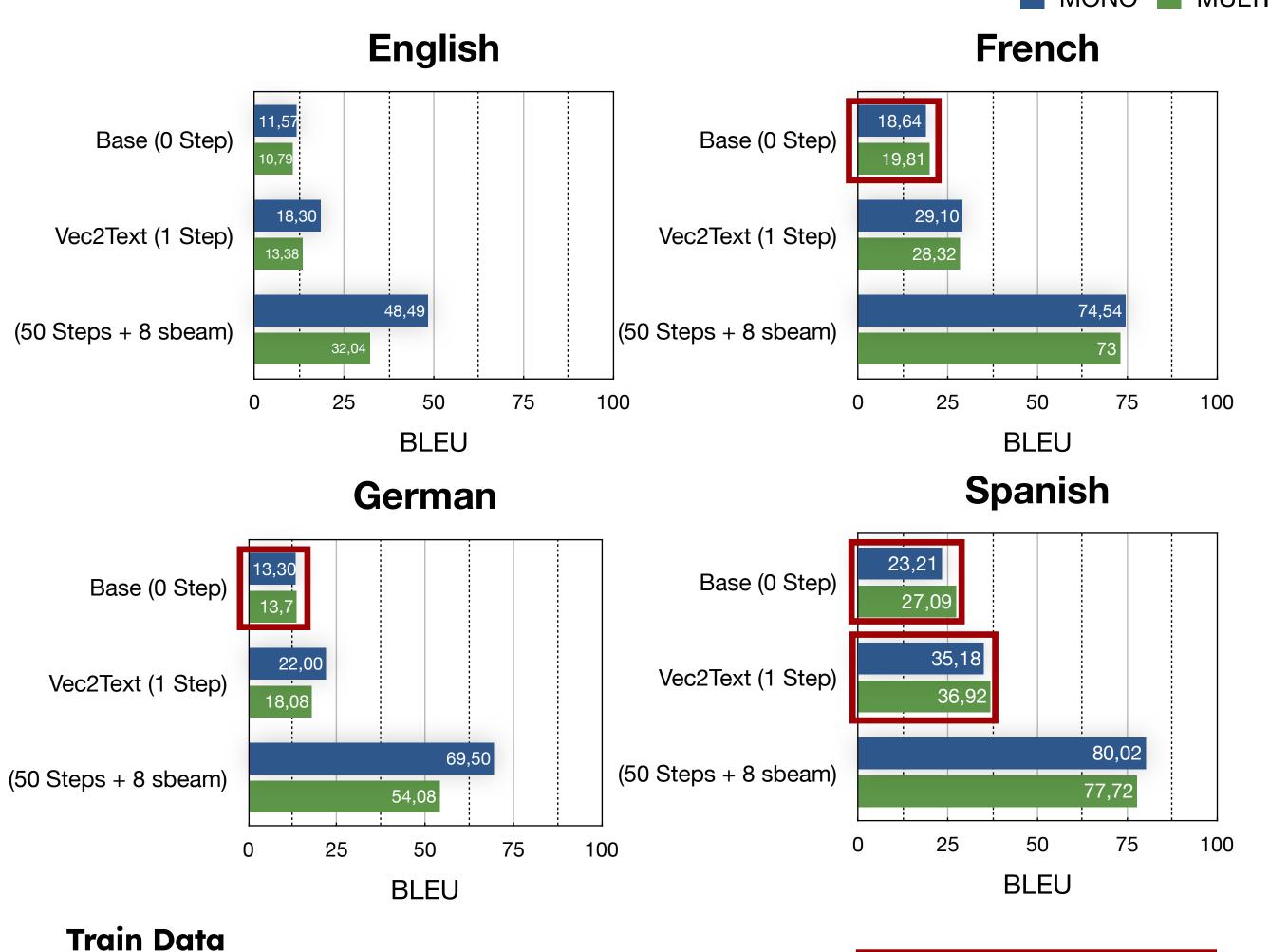
#### 1 Inversion Attack Schema Name: Emily Johnson, Date of Birth: May 12, 1989 Address: 123 Maple Street, Anytown, USA Phone Number: (555) 5551234 EaaS **USER** Name: Emily Johnson, Date of Birth: May 12, 1989 Address: 123 Maple Street Eavesdropping Reconstructing Anytown, USA Phone Number: (555) 5551234 **Embeddings Texts**

# 3 Attack Mono- and Multi-lingual Embeddings in English



- GTR and ME5 trained and evaluated on 5M samples from Natural Questions[3];
- For GTR, BLEU correlates positively with COS;
- COS is higher from the base model for ME5;
- GTR consistently outperforms ME5 in BLEU except for Base Model.

# 4 Attack Multilingual Encoders



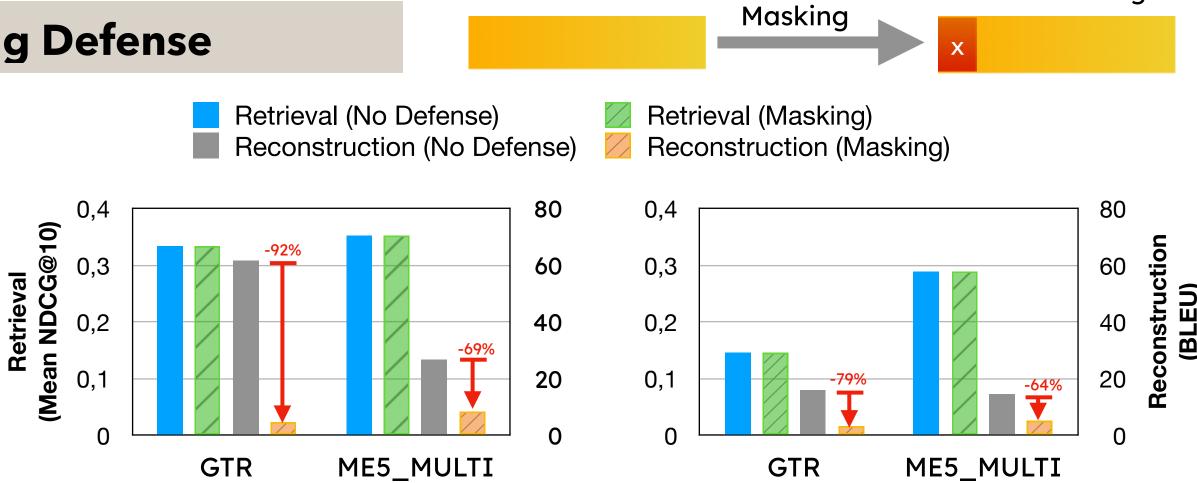
# Spanish **English** French MONO MULTI **6 Noise Insertion Defense** Retrieval(GTR) Retrieval (ME5) Retrieval (Mean NDCG@10) 10-2 10-1 **Noise Level BEIR (English)** Fail to defend ME5, while successfully defending GTR.

7 Masking Defense

GitHub: <a href="https://github.com/siebeniris/MultiVec2Text/">https://github.com/siebeniris/MultiVec2Text/</a>

(See the list of trained inversion models on GitHub)

HuggingFace: <a href="https://huggingface.co/yiyic">https://huggingface.co/yiyic</a>



Successfully defending both monolingual and multilingual language embeddings.

# Inverting English and German Parallel Texts using ME5\_MULTI

MULTI: 1.25 M in each of {LANG} from MTG, total 5M

Input	ford urged to recall 1.3 million suvs over exhaust fumes	ford wird aufgefordert 1,3 millionen suvs wegen abgasen zurückzurufen
Round 1	ford urged to recall fumes from 1.3 million suvs	ford ist auf 1,3 millionen suvs zurückgefordertgas abgerufen
Round 2	ford urged to recall 1.3 million suvs from oversowing fumes	ford ist auf 1,3 millionen suvs in abgas zurückgefordert
Round 3	ford urged to recall 1.3 million suvs omitted fumes	ford ist von 1,3 millionen suvs wegen abgas zurückgerufen
Round 4	ford urged to recall 1.3 million suvs overfuming fumes	ford ist angerufen, dass 1,3 millionen suvs wegen abgas zurückgerufen werden
Round 5	ford urged to recall 1.3 million suvs over of exhaust fumes	ford wird aufgefordert, 1,3 millionen suvs aufgrund von abgas zurückzurufen
Round 6	ford urged to recall 1.3 million suvs over exhaust fumes	ford wird aufgefordert 1,3 millionen suvs wegen abgas zurückzurufen
Round 7		ford wird aufgefordert 1,3 millionen suvs wegen abgasen zurückgerufen

# 8 Conclusion

- First work on multilingual and cross-lingual embedding inversion
- Multilingual models can be more vulnerable than monolingual models
- Traditional defense **only** works for monolingual models
- Novel defense effective for **both** mono- and multi-lingual models
- Advocate for a multilingual approach to LLM and NLP security as an entirety

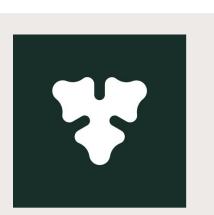
# References

[1] Morris, J. X., Kuleshov, V., Shmatikov, V., & Rush, A. M. (2023). Text embeddings reveal (almost) as much as text. arXiv preprint arXiv:2310.06816.

MULTI outperforms MONO

[2] Kwiatkowski, Tom, et al. "Natural questions: a benchmark for question answering research." Transactions of the Association for Computational Linguistics 7 (2019): 453-466.

[3] Chen, Y., Song, Z., Wu, X., Wang, D., Xu, J., Chen, J., ... & Li, L. (2021). MTG: A benchmark suite for multilingual text generation. arXiv preprint arXiv:2108.07140.





MONO: 5M MTG in {LANG}





