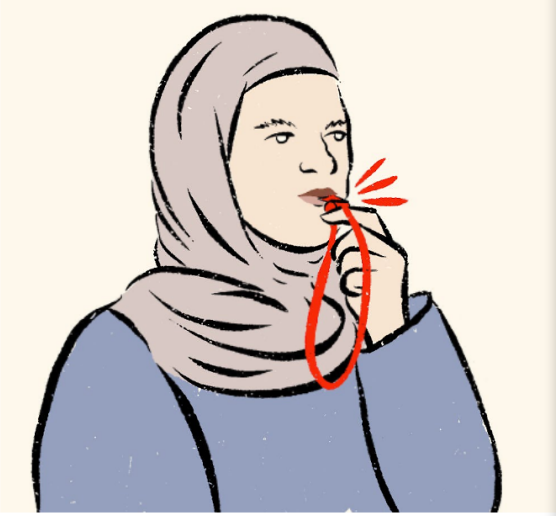
Whistleblowing

Wrongdoing,
misconduct,
illegal activities



Retaliations

- Harassment
- Job loss
- Blacklisting
- Death Threats

Protective Laws

"But am I protected?"

Anonymous reporting

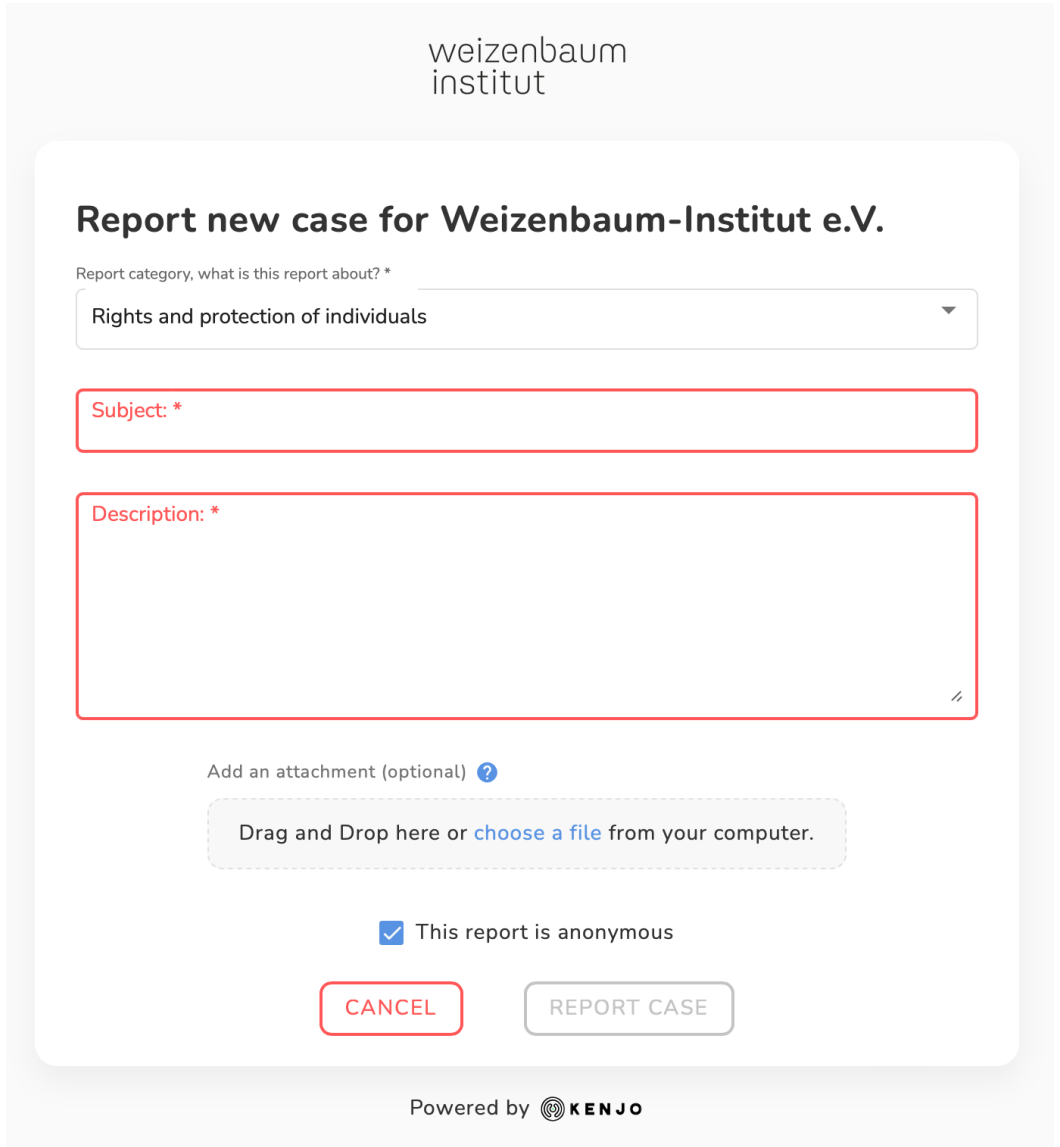
TECHNISCHE UNIVERSITÄT BERLIN

Um eine angemessene Bearbeitung zu ermöglichen, ist es wichtig, dass Hinweise so konkret wie möglich gemeldet werden. Daher bitten wir bei der Meldung folgende Fragen zu erläutern:

- Wer hat den Verstoß begangen?
- Was ist passiert?
- Wo ist es passiert?
- Wann ist es passiert?
- Wie lässt sich der Verstoß belegen?

Beschäftigte und Lehrbeauftragte der TU haben die Möglichkeit, Hinweise anonym oder unter Angabe Ihrer Identität abzugeben. Um einen Hinweis einzureichen, haben Sie mehrere Optionen zur Auswahl:

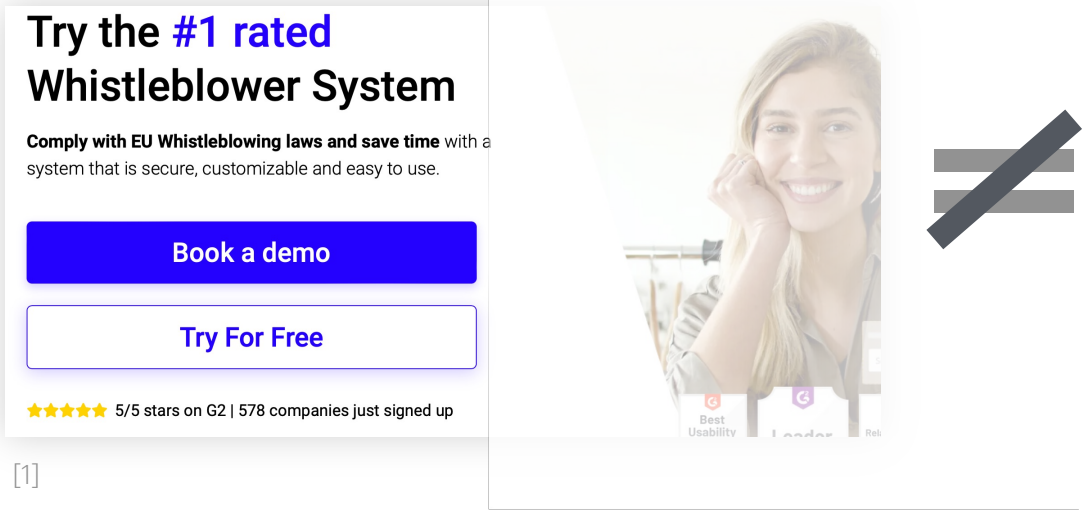
1. Kontakt per E-Mail: Hinweisgeber
2. Telefonisch oder nach Terminvereinbarung: (030) 314 79839
3. Hauspost: IR1



https://www.weizenbaum-institut.de/meldestelle-fuer-whistleblower-am-weizenbaum-institut/

Whistleblowing Software

Anonymous communication



[1]

≠

Anonymous reporting

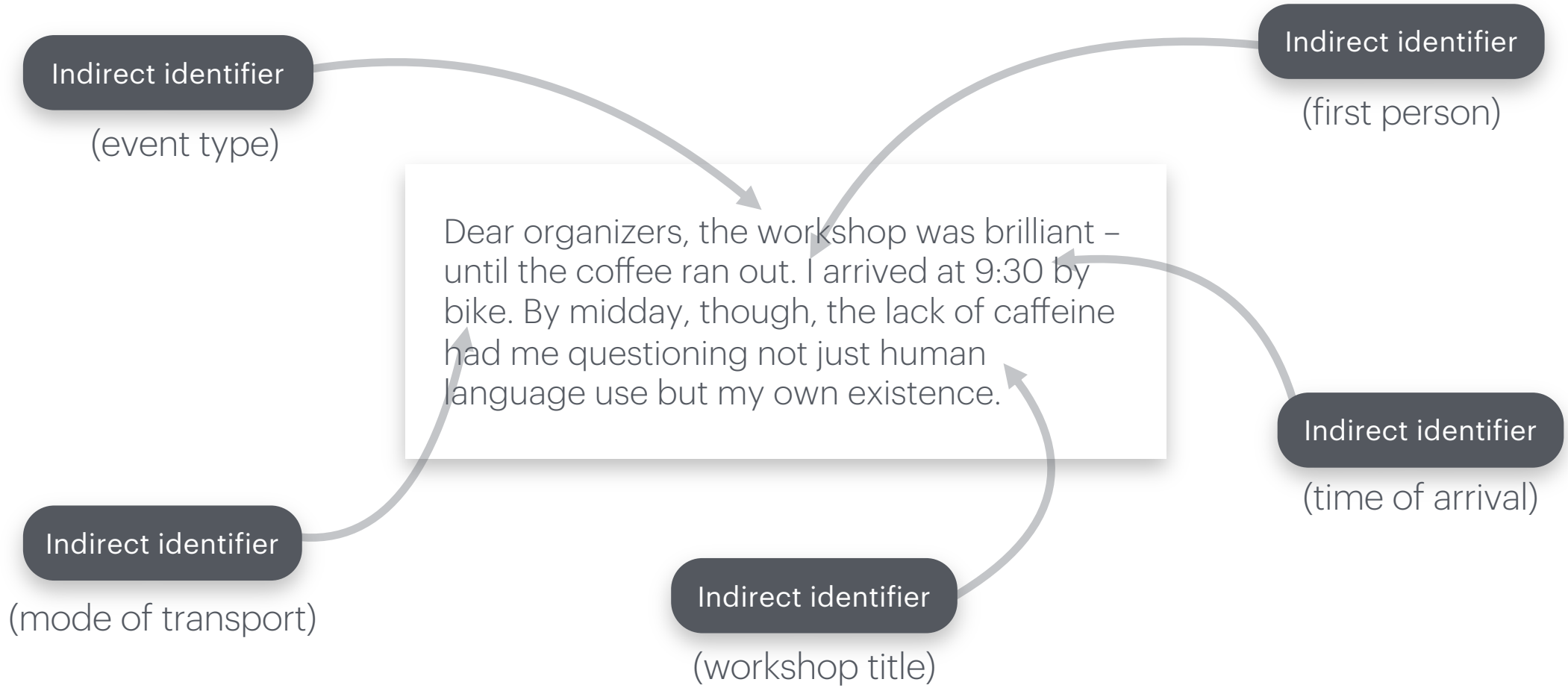
due to...

- Author's unique **writing style**
- Specific **content**

[2]

[1] https://whistleblowersoftware.com/en [Accessed 29-May-2024]
[2] Bettina Berendt and Stefan Schiffner, 2022. The International Review of Information Ethics.

Dear organizers, the workshop was brilliant – until the coffee ran out. I arrived at 9:30 by bike. By midday, though, the lack of caffeine had me questioning not just human language use but my own existence.


Indirect identifier
(event type)
Indirect identifier
(first person)
Dear organizers, the workshop was brilliant – until the coffee ran out. I arrived at 9:30 by bike. By midday, though, the lack of caffeine had me questioning not just human language use but my own existence.
Indirect identifier
(time of arrival)
Indirect identifier
(mode of transport)
Indirect identifier
(workshop title)

Related Work

Text Sanitization

NER [1]

NER + Coref-Resolution[2]

Fine-Tuned BERT model for PIIs [3]

"On ~~24 January 2023~~, ~~John Smith~~ poured..."

Underestimates (inside and outside document) context

Any term (not just named entities) can be identifying [4]

Privacy-preserving data publishing (PPDP)

"On 24 January 2023, John Smith poured..."

Differential Privacy

0.234	0.284	0.612
0.103	0.203	0.203
1.452	0.759	1.222
0.654	0.252	0.344
...	...	...

Noise

0.137	0.334	0.334
0.403	0.103	0.163
1.432	1.852	1.652
0.754	0.554	0.354
...	...	...

e.g. [5]

"On January 2023, J. Smith do..."

K-Anonymity

Grammatical errors

Limited variation

[1] Larbi et al., 2022. A Systematic Study on Clinical Text Processing.
[2] Adams et al., 2019. Linköping University Electronic Press.
[3] Kleinberg et al., 2022. arXiv preprint arXiv:2208.13081 (2022).
[4] Arvind Narayanan and Vitaly Shmatikov., 2010. Commun. ACM 53, 6 (2010).
[5] Mattern et al., 2022. Findings of the Association for Computational Linguistics: NAACL 2022.

Related Work

Text Sanitization

Dear ~~organizers~~, the workshop was brilliant – until the coffee ran out. I arrived at ~~9:30~~ by bike. By midday, though, the lack of caffeine had me questioning not just human language use but my own existence.

Privacy-preserving data publishing (PPDP)

Dear organizers, the workshop was brilliant – until the coffee ran out. I arrived at 9:30 by bike. By midday, though, the lack of caffeine had me questioning not just human language use but my own existence.

Related Work

Text Sanitization

Dear ~~organizers~~, the workshop was brilliant – until the coffee ran out. I arrived at ~~9:30~~ by bike. By midday, though, the lack of caffeine had me questioning not just human language use but my own existence.

Privacy-preserving data publishing (PPDP)

Dear **coordinators**, **a** class was great – until the espresso **running out**. I arrived at **10:30** by car. By day, though, the lack from caffeine had me questioning not just human language use but my own existence.

Related Work

Dear organizers, the workshop was brilliant – until the coffee ran out. I arrived at 9:30 by bike. By midday, though, the lack of caffeine had me questioning not just human language use but my own existence. After trying to refill my blue water bottle twice without success, I realized I couldn't make it through the rest of the day and had to leave early.

Original (annotate mode)

Dear organizers, the workshop was brilliant– until the coffee ran out. I arrived at 9:30 by bike. By midday, though, the lack of caffeine had me questioning not just human language use but my own existence. After trying to refill my blue water bottle twice without success, I realized I could n't make it through the rest of the day and had to leave early.

Sanitized

Press the *Sanitize* button to start.

Sanitize ✨

Original (annotate mode)

Dear organizers, the workshop was brilliant– until the coffee ran out. I arrived at 9:30 by bike. By midday, though, the lack of caffeine had me questioning not just human language use but my own existence. After trying to refill my blue water bottle twice without success, I realized I could not make it through the rest of the day and had to leave early.

Sanitized

Dear organizers, the workshop was a great success, but the lack of coffee caused me to question not only human language use, and my own existence. After attempting to refill my water container two times without success I was forced to leave the event at an unsuitable time.

Sanitize ✨

Broken English With A Heavy German Accent Translator

Translate from Normal Language into Broken English With A Heavy German Accent

Normal Language	→	Broken English With A Heavy German Accent
Dear organizers, the workshop was brilliant – until the coffee ran out. I arrived at 9:30 by bike. By midday, though, the lack of caffeine had me questioning not just human language use but my own existence. After trying to refill my blue water bottle twice without success, I realized I couldn't		Dear organizers, workshop was sehr brilliant – until the coffee is gone, ja? I come by bike at 9:30, very nice! But by midday, oh, mein Gott, no caffeine make me question not only human language, but my own existence, ja! I try to refill my blue Wasserflasche two times, but no success, ach! I realize, I cannot make it through the rest of the day, so I must leave early, ja.

Translate

Original (annotate mode)

Dear organizers, workshop was sehr brilliant– until the coffee is gone, ja? I come by bike at 9:30, very nice! But by midday, oh, mein Gott, no caffeine make me question not only human language, but my own existence, ja! I try to refill my blue Wasserflasche two times, but no success, ach! I realize, I can not make it through the rest of the day, so I must leave early, ja.

Sanitized

The workshop was excellent, but it was interrupted by the lack of coffee, which led me to question my own existence, resulting in a premature departure.

Sanitize ✨

Risk Mitigation

Anonymization Operations

	Nondescript phrase, e.g. “certain place”	Deletion of dependent phrases	Removal	Deletion of dependent phrases	Nondescript phrase, e.g. “somebody”	LLM rephrasing
Risk	**Names of Named Entities**	**Common Nouns**	**Modifiers**	**Proper Nouns**	**Pronouns**	**Stylometric Features**
High	Suppression	Suppression	Suppression	Suppression	Suppression	Perturbation
Medium	Perturbation	Generalization	Perturbation	Generalization	Suppression	Perturbation
	Zero-weight in LLM generation	Hypernym from WordNet	Zero-weight in LLM generation	Broader Wikidata term	Nondescript phrase, e.g. “somebody”	LLM rephrasing

Risk Mitigation

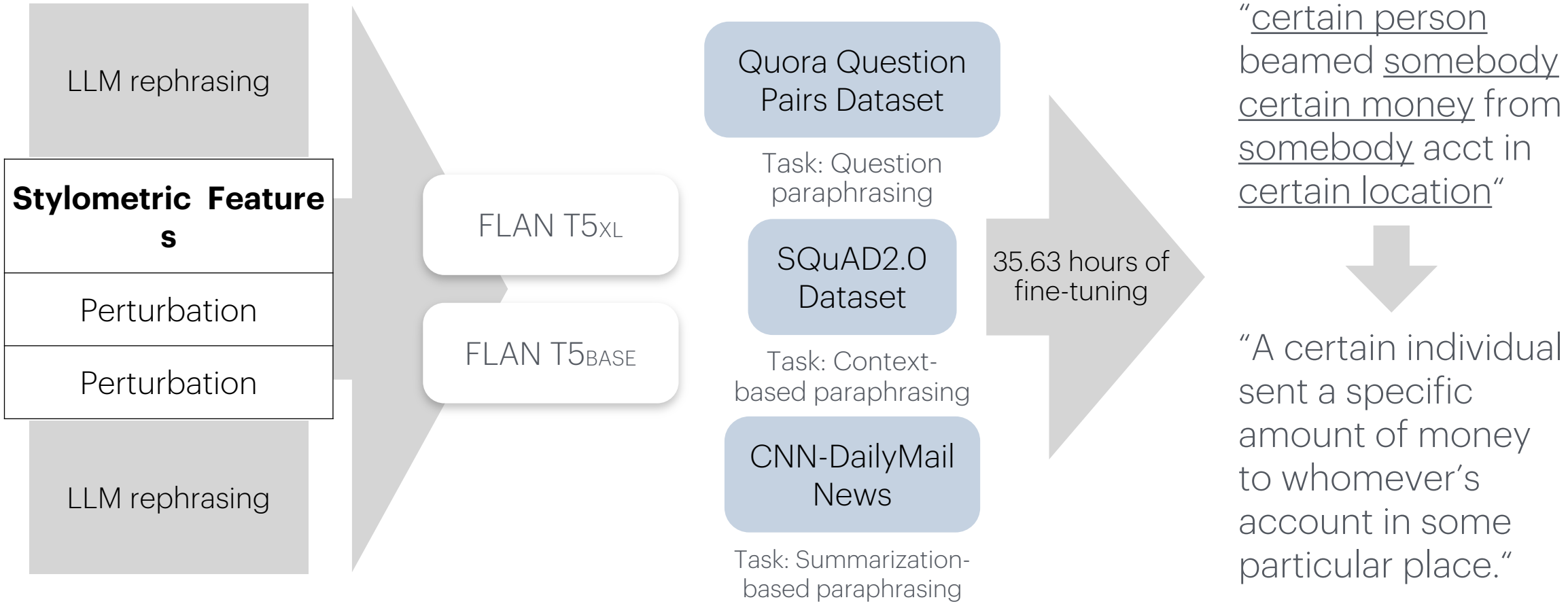
Anonymization Operations

	Nondescript phrase, e.g. “certain time“ for “9:30“	Deletion of dependent phrases	Removal	Deletion of dependent phrases	Nondescript phrase, e.g. “somebody“	LLM rephrasing
Risk	**Names of Named Entities**	Common Nouns	Modifiers	**Proper Nouns**	Pronouns	Stylometric Features
High	Suppression	Suppression	Suppression	Suppression	Suppression	Perturbation
Medium	Perturbation	Generalization	Perturbation	Generalization	Suppression	Perturbation
	Zero-weight in LLM generation	Hypernym from WordNet	Zero-we in LL generat	Broader Wikidata entity, e.g. “water container“ for “water bottle“	ndescript rase, e.g. mebody“	LLM rephrasing

Risk Mitigation

Anonymization Operations

Risk Mitigation
LLM Rephrasing

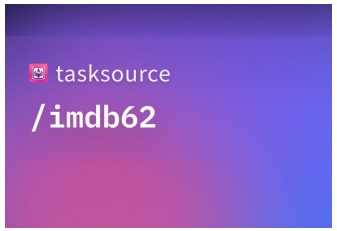


Evaluation Results

EUROPEAN COURT OF HUMAN RIGHTS
COUR EUROPÉENNE DES DROITS DE L'HOMME

Text Anonymization Benchmark (Pilán et al. 2022)

Focus: *Comparative benchmark for privacy protection and utility preservation*



IMDb62 movie reviews dataset

Focus: *Protection against Authorship Attribution Attacks*

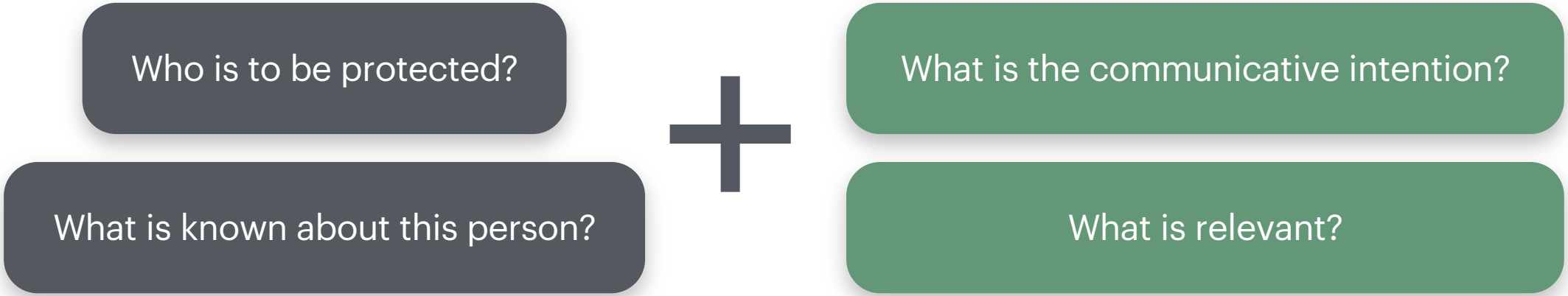
UNITED STATES HOUSE COMMITTEE ON
WAYS & MEANS
CHAIRMAN JASON SMITH

Whistleblower hearing (Hunter Biden tax evasion case, 2023)

Focus: *Qualitative view on rewritings*

[1] Pilán et al., 2002. Computational Linguistics 48, no. 4.
[2] https://huggingface.co/datasets/tasksource/imdb62 [Accessed 29-May-2024]
[3] https://waysandmeans.house.gov/?p=39854458 [Accessed 29-May-2024], "#2"

“Retains overall meaning“



Programmierpraktikum im SoSe 2025

“Re-Identification and Anonymity in Natural Language“

Programmierpraktikum: NLP for Social Good

Titel des Moduls
Programmierpraktikum: NLP for Social Good

Leistungspunkte
6

Modulverantwortliche*r
Berendt, Bettina

Sekretariat
Keine Angabe

Ansprechpartner*in
Staufer, Dimitri

Webseite
https://www.tu.berlin/ias

E-Mail-Adresse
staufer@tu-berlin.de

Lernergebnisse

Die Studierenden:
- sind in der Lage, Large Language Models gezielt für gesellschaftlich relevante Anwendungsbereiche zu optimieren und anzupassen.
- kennen Konzepte und Methoden zur Messung, Evaluierung und Förderung sozialer Aspekte in Daten und Modellen.
- verfügen über ein fundiertes Verständnis der Eigenschaften und Limitationen von Datensätzen, welches ihnen ermöglicht, den sozialen Mehrwert und die Grenzen von Forschungsprojekten realistisch zu bewerten.
- beherrschen Techniken der Datensammlung und -beschaffung und können geeignete Ansätze für ihre sozial relevanten Vorhaben auswählen und anwenden.