LLMs for Social Science and the Humanities

Prof. Dr. Steffen Eger

steffen.eger@utn.de
https://nl2g.github.io/

NLLG

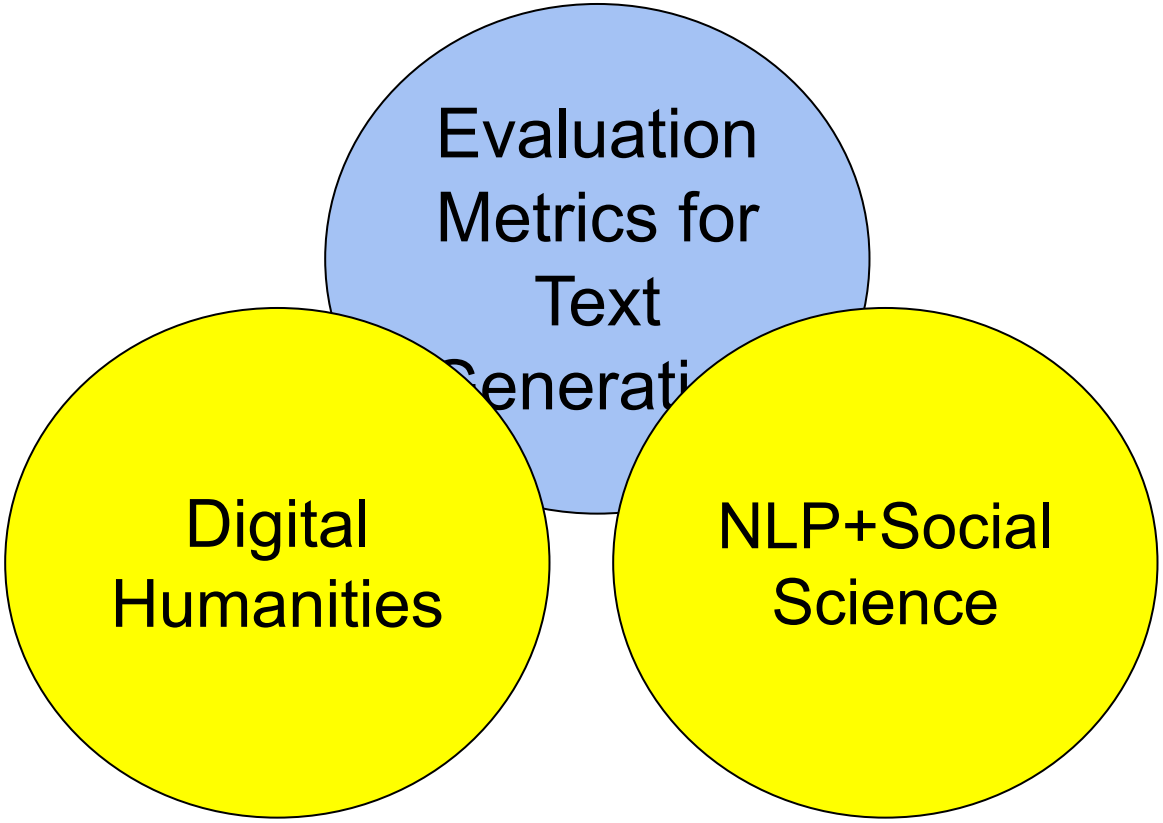
NLLG

Vita

- PhD in Economics (2014)
- PostDoc in Frankfurt and TU Darmstadt (2014-2018)
- Junior Group Leader (2018-2022)
- Interim Professor Bielefeld (2022)
- Heisenberg Group Leader Bielefeld, Mannheim 2022-2024
- Full Professor at UTN, 2024-

Picture source

Research interests



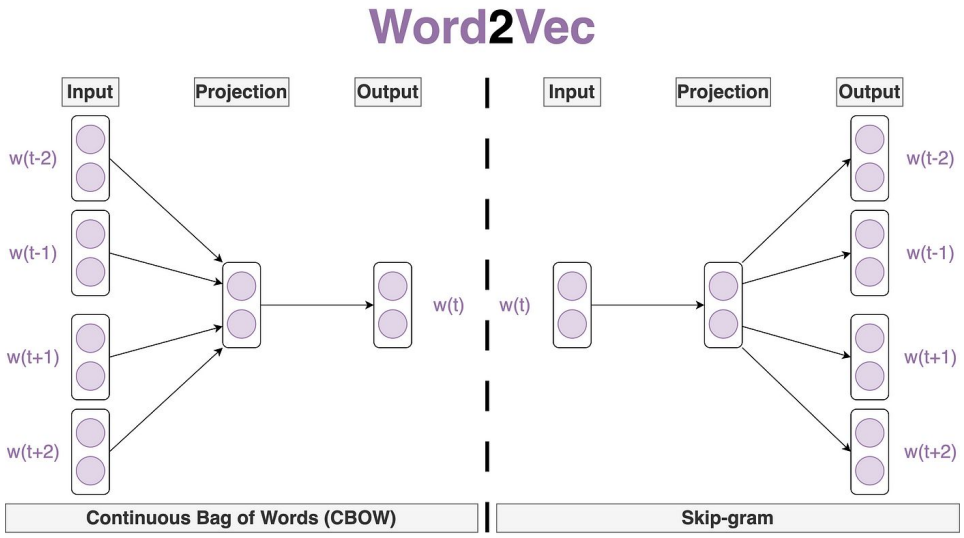
Overview

- 1) A bit on the AI/LLM revolution
- 2) LLMs for detection of social solidarity
- 3) LLMs for literature translation + Evaluation

The AI ®evolution

The AI ®evolution

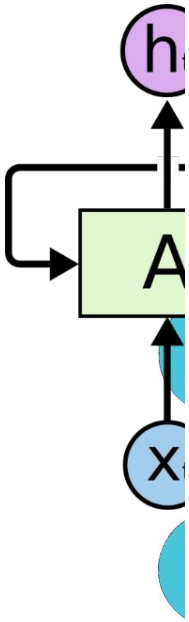
- starting ~2010s: Deep Learning Revolution
  - 2013-2014: “Word2Vec”
  - 2018: BERT
  - 2022: ChatGPT + LLMs



Google

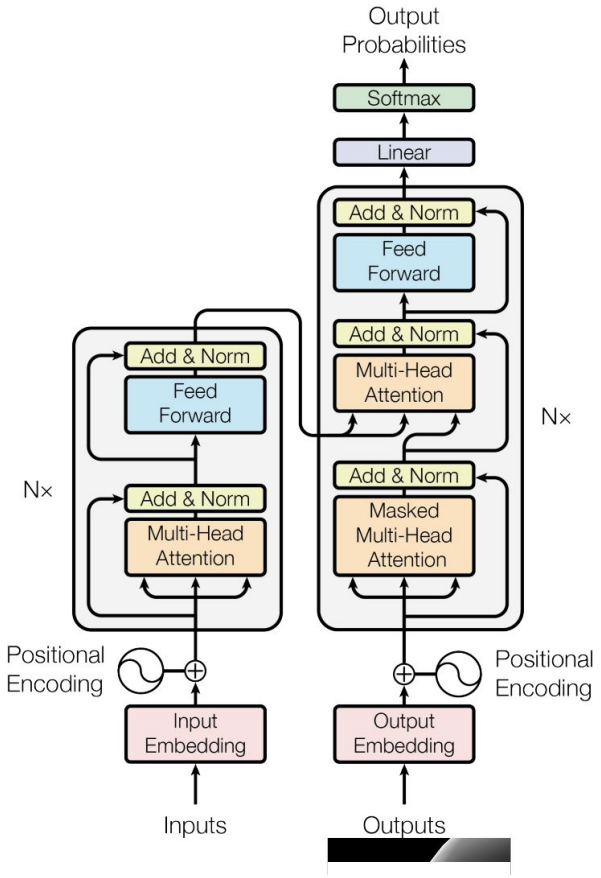
Deep Learning & LLMs

Deep Learning a.k.a. Neural Networks



BERT

Encoder

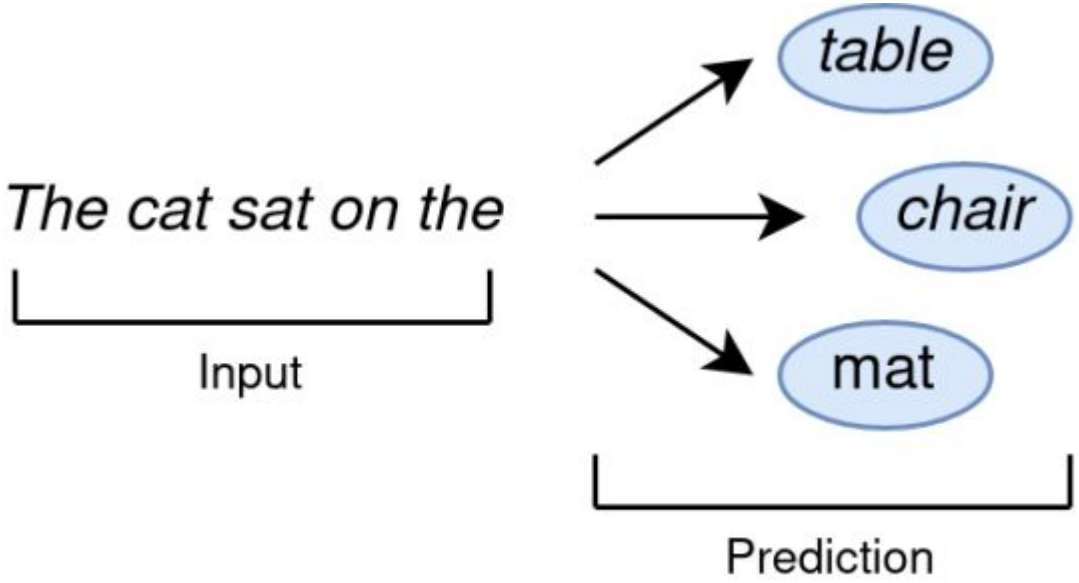


GPT

Decoder

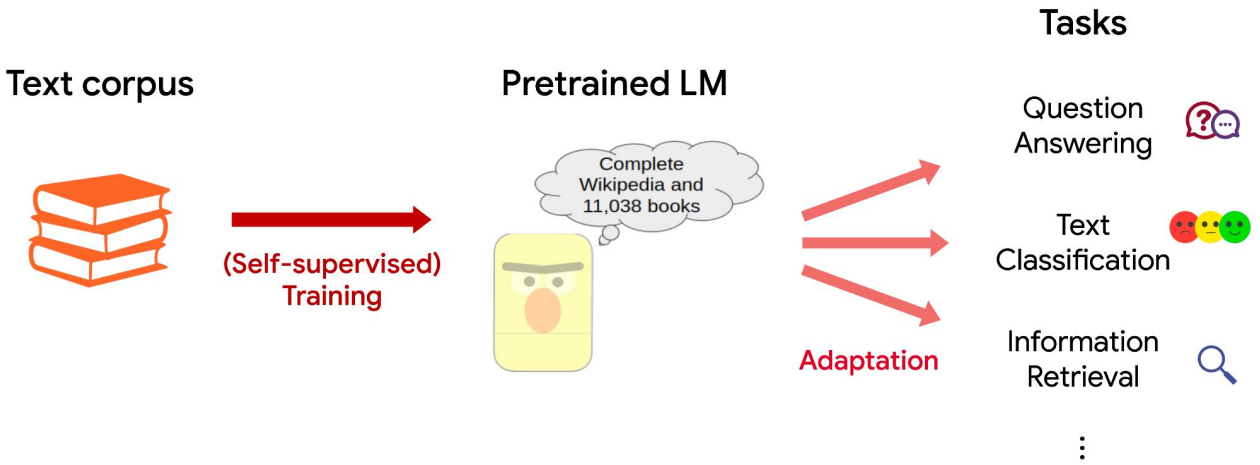
How LLMs work

- (Nowadays) Transformer Architecture
- next token prediction



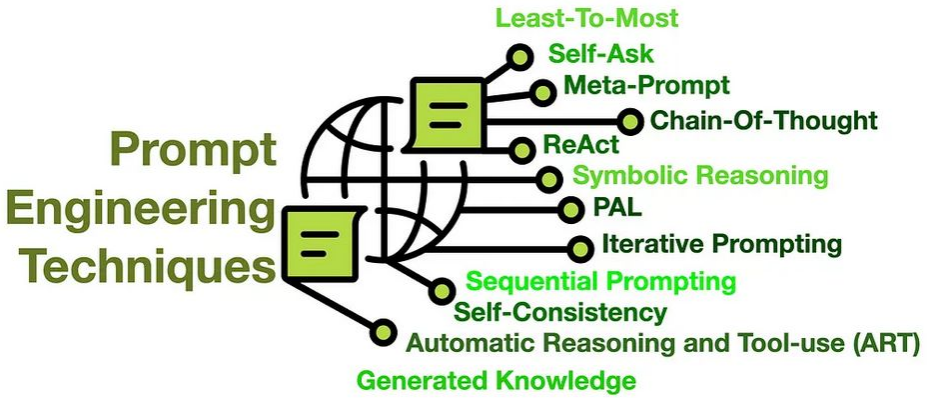
How LLMs work

- (Nowadays) Transformer Architecture
- next token prediction
- pre-training & fine-tuning



How LLMs work

- (Nowadays) Transformer Architecture
- next token prediction
- pre-training & fine-tuning
- prompting & prompt engineering



How LLMs work

- instruction fine-tuning
- human alignment

Prompt *Explain the moon landing to a 6 year old in a few sentences.*

pletion **GPT-3**

```
Explain the theory of gravity to a 6 year old.

Explain the theory of relativity to a 6 year old in a few sentences.

Explain the big bang theory to a 6 year old.

Explain evolution to a 6 year old.
```

InstructGPT

```
People went to the moon, and they took pictures of what they saw, and sent them back to the earth so we could all see them.
```

How LLMs work

Outcomes:

- LLMs learn all kinds of things indirectly, e.g., machine translation

Properties:

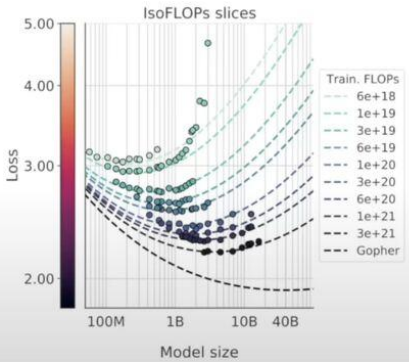
- Size matters; scaling laws

LLM Scaling Laws

Performance of LLMs is a smooth, well-behaved, predictable function of:
- **N**, the number of parameters in the network
- **D**, the amount of text we train on

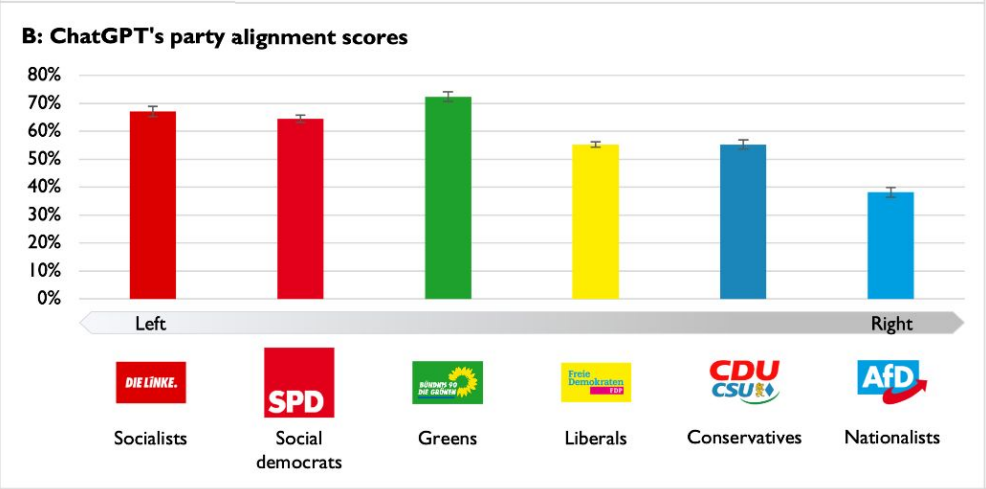
And the trends do not show signs of "topping out"

=> We can expect more intelligence "for free" by scaling



LLM pitfalls

- Biases
  - https://arxiv.org/abs/2301.01768
  - https://aclanthology.org/2023.acl-long.656/



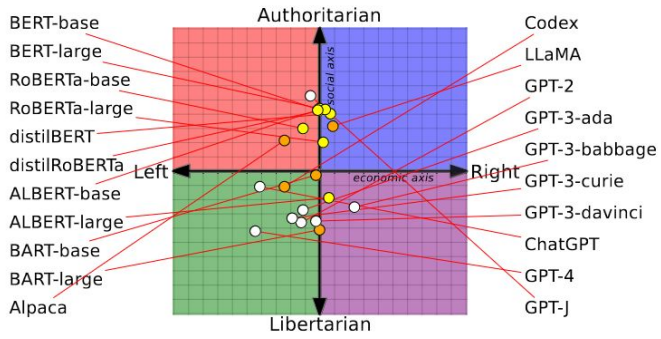
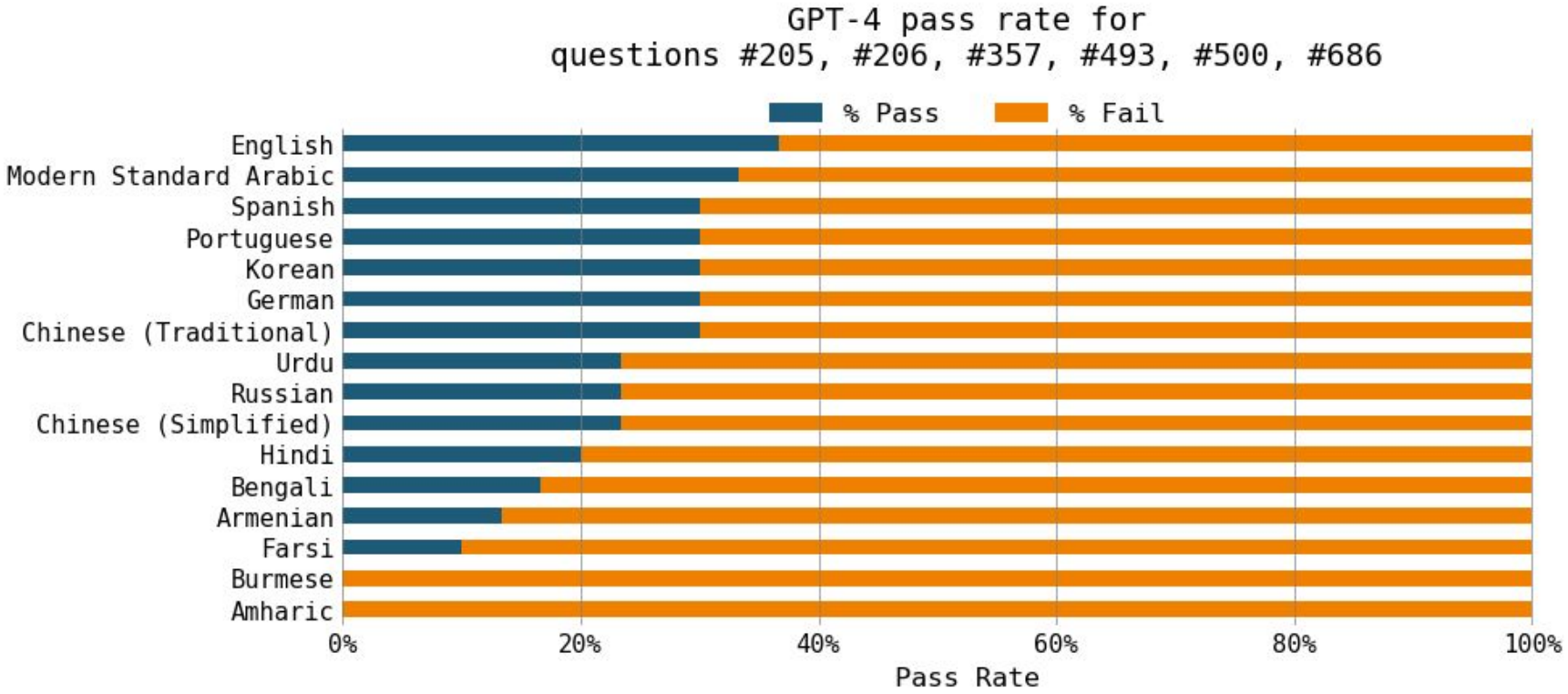


Figure 1: Measuring the political leaning of various pretrained LMs. BERT and its variants are more socially conservative compared to the GPT series. Node color denotes different model families.

LLM pitfalls

- Other languages



LLM pitfalls

- Other Limitations:
  - Hallucination
  - Reasoning
  - …

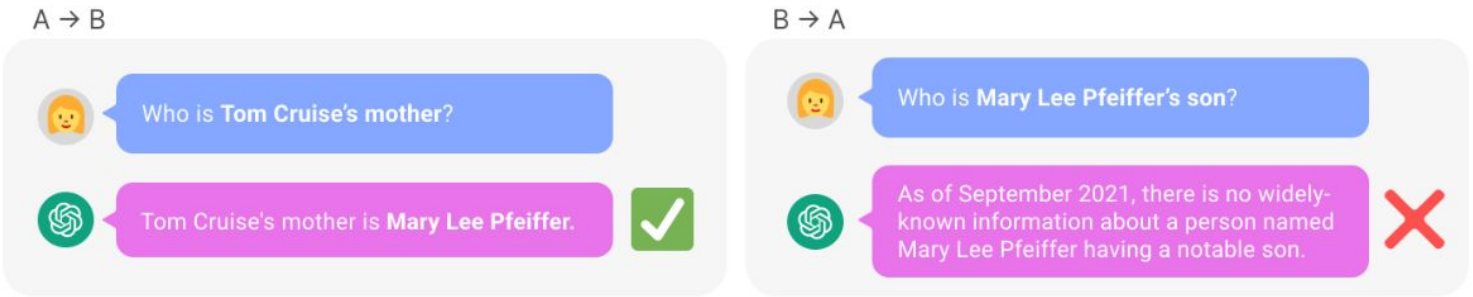
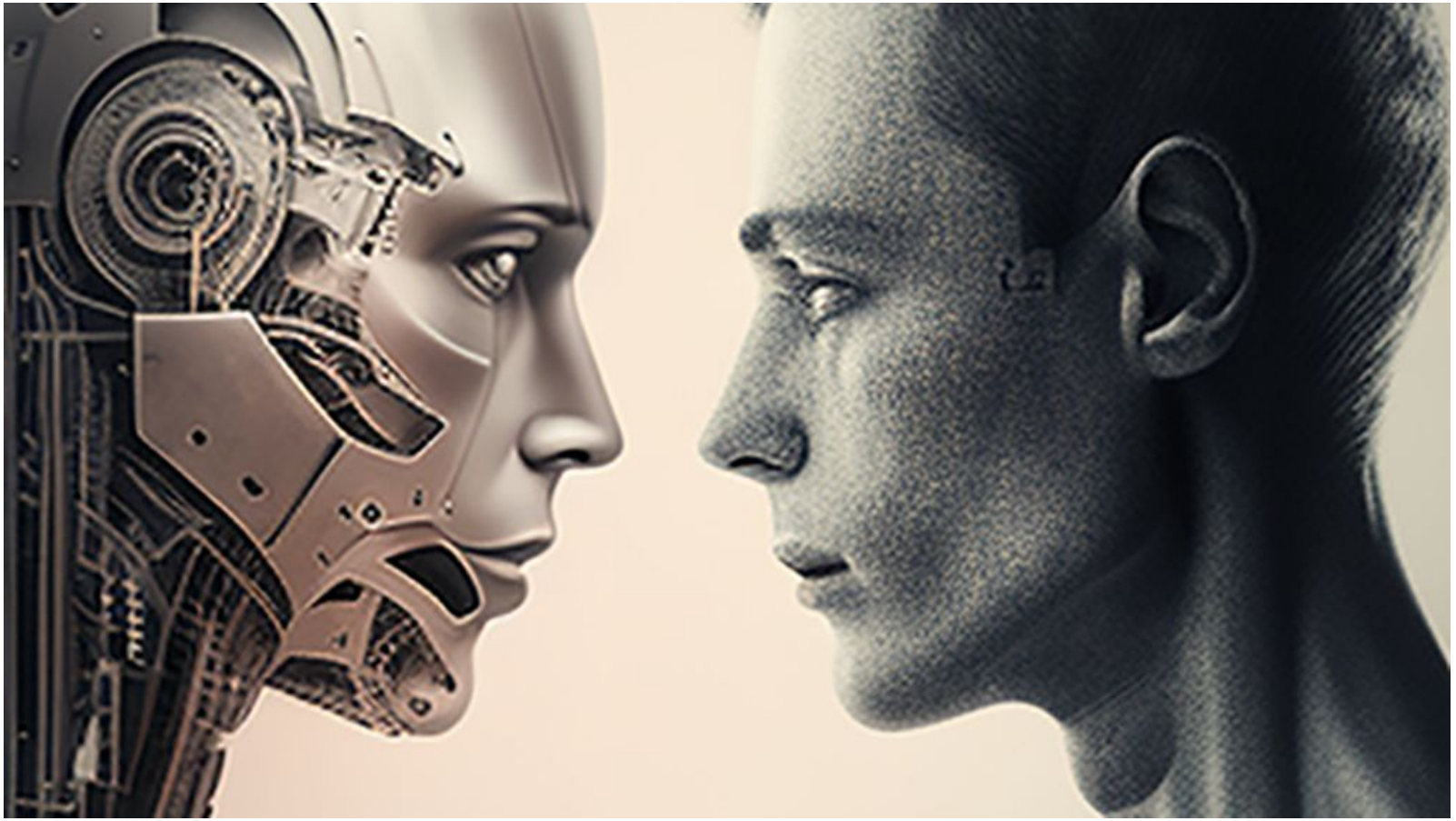


Figure 1: **Inconsistent knowledge in GPT-4.** GPT-4 correctly gives the name of Tom Cruise's mother (left). Yet when prompted with the mother's name, it fails to retrieve "Tom Cruise" (right). We hypothesize this ordering effect is due to the Reversal Curse. Models trained on "*A* is *B*" (e.g. "Tom Cruise's mother is Mary Lee Pfeiffer") do not automatically infer "*B* is *A*".

Man vs. Machine



Fine-Grained Detection of Social Solidarity using LLMs

Why solidarity?

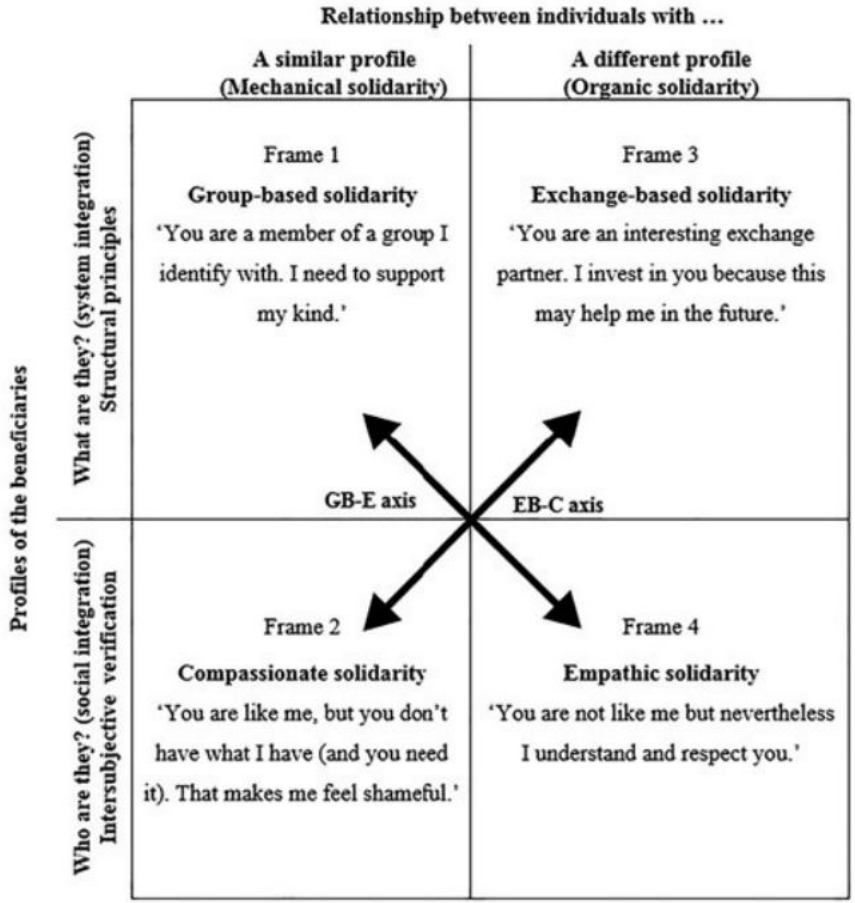
- Solidarity as a cohesive bond that keeps societies connected
- Lack of solidarity → risk that societies break apart
- Solidarity as a key concept in the social sciences

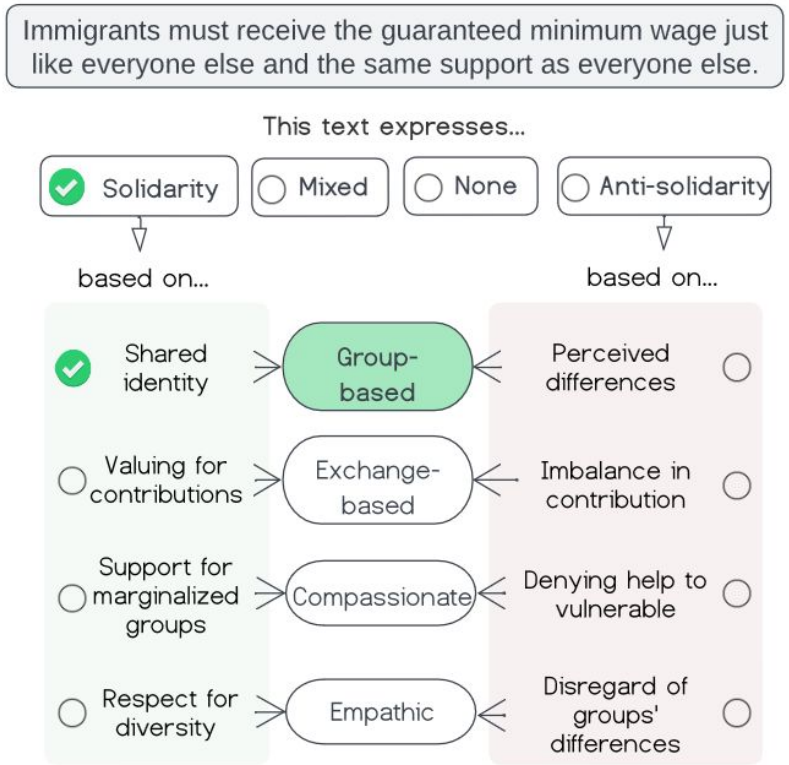
- Tracing solidarity over time can yield insights into where societies are heading towards

Social Solidarity and Solidarity Frames

- “Willingness to share resources, be that directly or indirectly” (Lahusen and Grasso 2018)
- More fine-grained scheme based on Thijssen 2012:
  - group-based (anti-)solidarity
  - exchange-based (anti-)solidarity
  - compassionate (anti-)solidarity
  - empathic (anti-)solidarity
- Underlying theory: Durkheim (mechanical vs. organic solidarity) and Honneth

Social Solidarity and Solidarity Frames





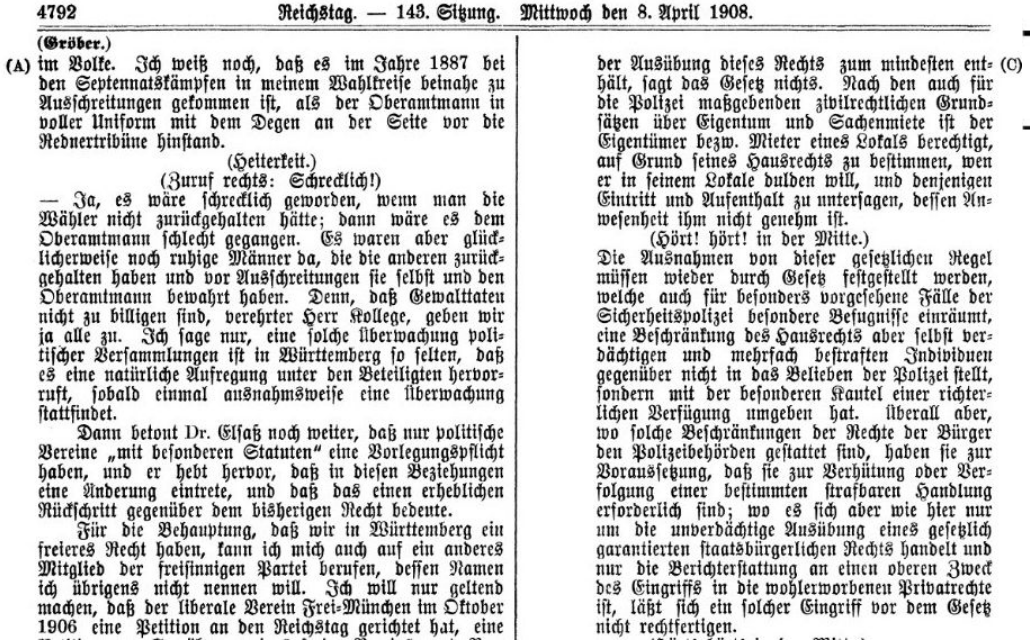
Data I

- DeuParl, German Parliamentary Proceedings
- 1867-2020
- Available online
  - https://www.reichstagsprotokolle.de/
  - https://www.bundestag.de/protokolle
- Preprocessed in our previous work:
  - Walter et al., https://ieeexplore.ieee.org/document/9651887

Data II

Gröber, Abgeordneter: Meine Herren, der Gedanke des Herrn Abgeordneten Gothetn, der dem Beschluß deS hohen Hauses in der letzten Sitzung zu Grunde lag, ist meines Erachtens zweifellos richtig. Es ist nicht er-

Gröber, Abgeordneter: Meine Herren, der Gedanke des Herrn Abgeordneten Gothein, der dem Beschluß des hohen Hauses in der letzten Sitzung zu Grunde lag, ist meines Erachtens zweifellos richtig. Es ist nicht er-



4792 Reichstag. — 143. Sitzung. Mittwoch den 8. April 1908.

(Gröber.)

(A) im Volke. Ich weiß noch, daß es im Jahre 1887 bei den Septennatskämpfen in meinem Wahlkreise beinahe zu Ausschreitungen gekommen ist, als der Oberamtmann in voller Uniform mit dem Degen an der Seite vor die Rednertribüne hinstand.

(Heiterkeit.)

(Zuruf rechts: Schrecklich!)

— Ja, es wäre schrecklich geworden, wenn man die Wähler nicht zurückgehalten hätte; dann wäre es dem Oberamtmann schlecht gegangen. Es waren aber glücklicherweise noch ruhige Männer da, die die anderen zurückgehalten haben und vor Ausschreitungen sie selbst und den Oberamtmann bewahrt haben. Denn, daß Gewalttaten nicht zu billigen sind, verehrter Herr Kollege, geben wir ja alle zu. Ich sage nur, eine solche Überwachung politischer Versammlungen ist in Württemberg so selten, daß es eine natürliche Aufregung unter den Beteiligten hervorruft, sobald einmal ausnahmsweise eine Überwachung stattfindet.

Dann betont Dr. Elsaß noch weiter, daß nur politische Vereine „mit besonderen Statuten" eine Vorlegungspflicht haben, und er hebt hervor, daß in diesen Beziehungen eine Änderung eintrete, und daß das einen erheblichen Rückschritt gegenüber dem bisherigen Recht bedeute.

Für die Behauptung, daß wir in Württemberg ein freieres Recht haben, kann ich mich auch auf ein anderes Mitglied der freisinnigen Partei berufen, dessen Namen ich übrigens nicht nennen will. Ich will nur geltend machen, daß der liberale Verein Frei-München im Oktober 1906 eine Petition an den Reichstag gerichtet hat, eine

der Ausübung dieses Rechts zum mindesten enthält, sagt das Gesetz nichts. Nach den auch für die Polizei maßgebenden zivilrechtlichen Grundsätzen über Eigentum und Sachenmiete ist der Eigentümer bezw. Mieter eines Lokals berechtigt, auf Grund seines Hausrechts zu bestimmen, wen er in seinem Lokale dulden will, und denjenigen Eintritt und Aufenthalt zu untersagen, dessen Anwesenheit ihm nicht genehm ist. (C)

(Hört! hört! in der Mitte.)

Die Ausnahmen von dieser gesetzlichen Regel müssen wieder durch Gesetz festgestellt werden, welche auch für besonders vorgesehene Fälle der Sicherheitspolizei besondere Befugnisse einräumt, eine Beschränkung des Hausrechts aber selbst verdächtigen und mehrfach bestraften Individuen gegenüber nicht in das Belieben der Polizei stellt, sondern mit der besonderen Kautel einer richterlichen Verfügung umgeben hat. Überall aber, wo solche Beschränkungen der Rechte der Bürger den Polizeibehörden gestattet sind, haben sie zur Voraussetzung, daß sie zur Verhütung oder Verfolgung einer bestimmten strafbaren Handlung erforderlich sind; wo es sich aber wie hier nur um die unverdächtige Ausübung eines gesetzlich garantierten staatsbürgerlichen Rechts handelt und nur die Berichterstattung an einen oberen Zweck des Eingriffs in die wohlerworbenen Privatrechte ist, läßt sich ein solcher Eingriff vor dem Gesetz nicht rechtfertigen.

Period	Years	Tokens
KR1	1867-1890	40,585,912
KR2	1890-1918	77,175,976
WR	1918-1933	35,838,922
NS	1933-1942	230,018
CDU1	1949-1969	43,337,027
SPD1	1969-1982	35,208,879
CDU2	1982-1998	55,451,433
SPD2	1998-2005	28,614,189
CDU3	2005-2020	66,192,033

Data III

- Our focus is on two vulnerable groups:
  - women and migrants (foreigners)
- We find 32 keywords for “Migrant” and 18 for “Frau”
  - using similarity (=word embedding association)
  - and manual selection
- E.g., Migrant = *“Flüchtlinge, Ausländer, Emigrant, Immigrant, Vetriebene, Aussiedler, …”*
- Frau = *“Frau, Frauen, Mütter, Hausfrauen, …”*

Data IV

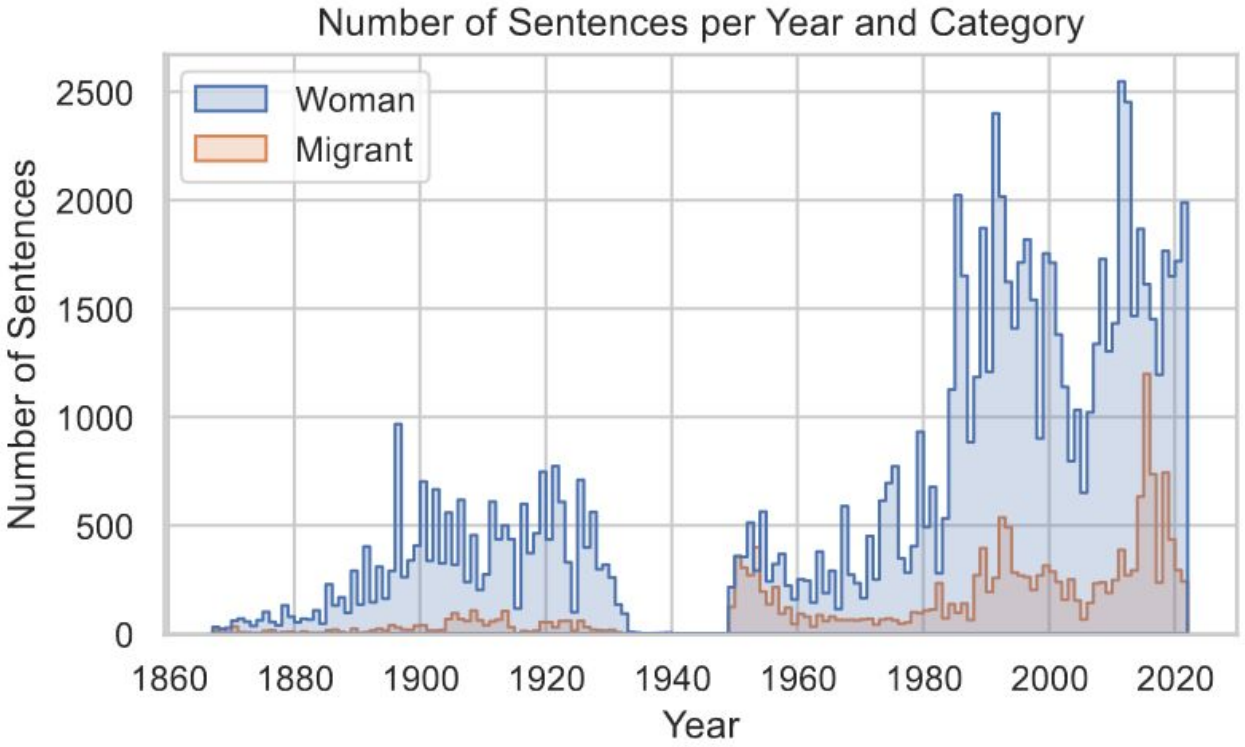
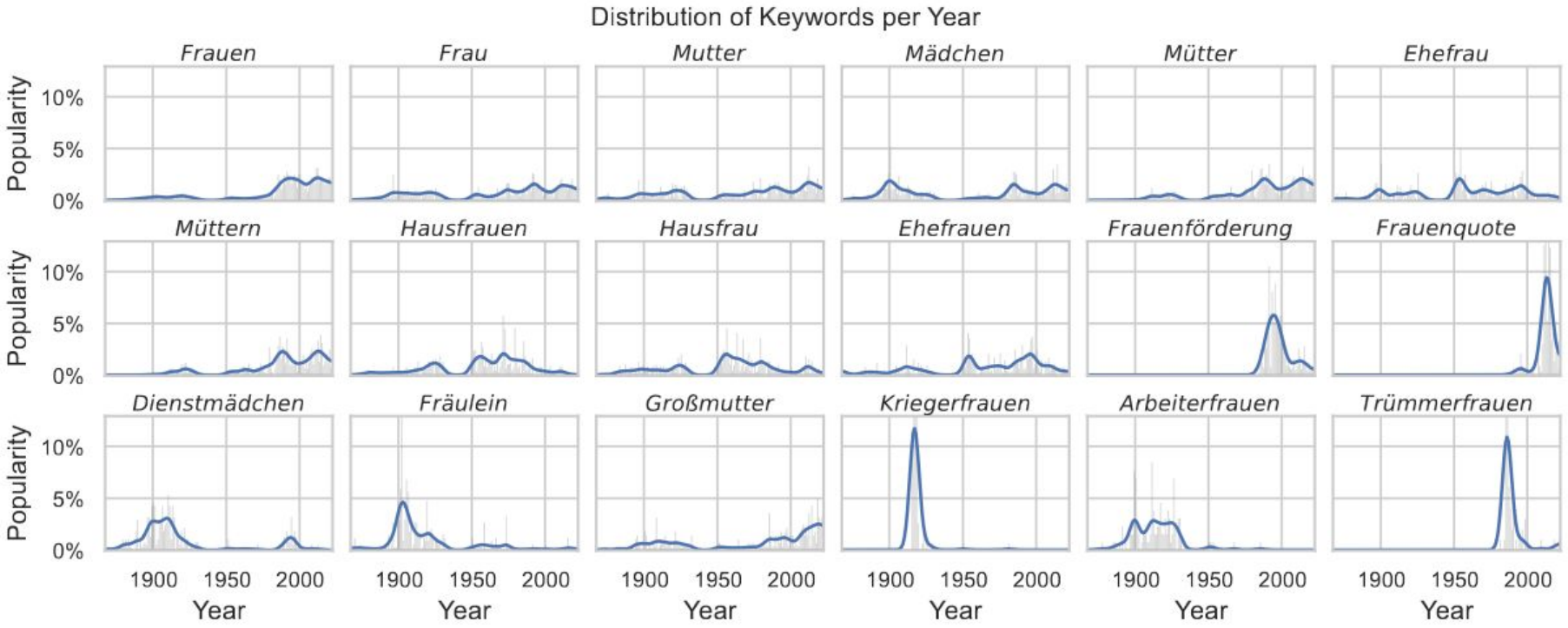


Figure 2: Number of instances in the Woman and Migrant dataset in each year.

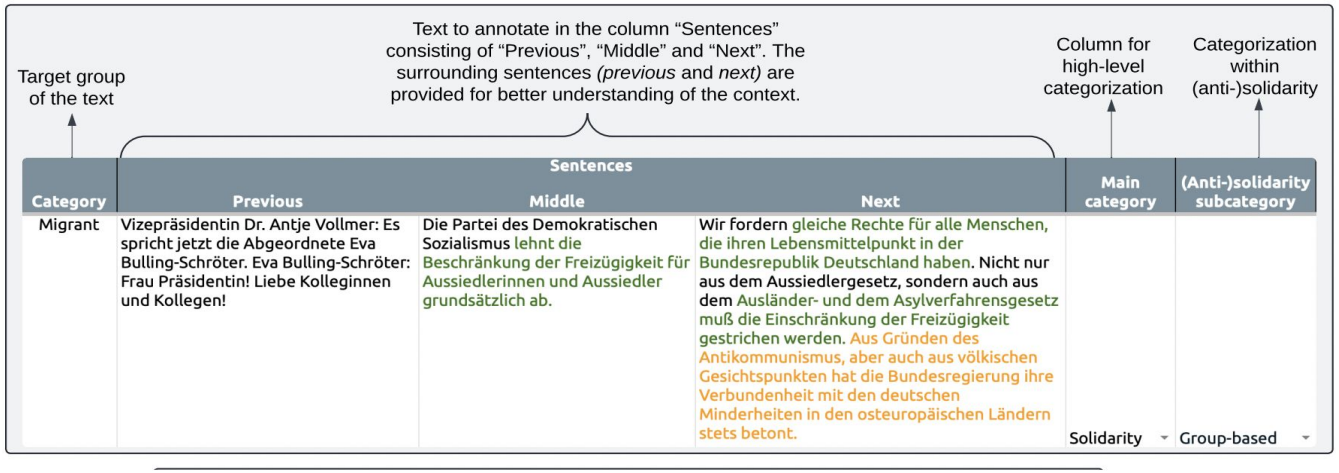
Data V


Distribution of Keywords per Year
Frauen
Frau
Mutter
Mädchen
Mütter
Ehefrau
Müttern
Hausfrauen
Hausfrau
Ehefrauen
Frauenförderung
Frauenquote
Dienstmädchen
Fräulein
Großmutter
Kriegerfrauen
Arbeiterfrauen
Trümmerfrauen
Popularity
10%
5%
0%
1900
1950
2000
Year

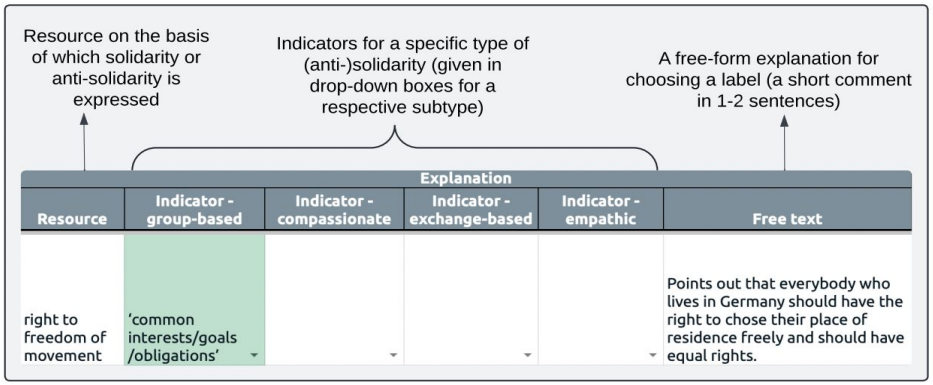
Annotation I

- 5 annotators (Social Science or Computer Science), all students
- ~40 hours per month for 9 month (~ 27k Euro)
  - <3000 instances
- High-level and fine-grained
- further includes highlighting and free-text
- Agreements:
  - 0.62 on high level
  - 0.42 fine-grained
  - hardly disagreement between solidarity and anti-solidarity

Annotation II



(a) Columns for high-level and (anti-)solidarity categorizations.



(b) Columns for providing explanations.

Annotation III

	Women	Migrant	Total per label
Group-based solidarity	112 (3.9%)	188 (6.6%)	300 (10.5%)
Exchange-based solidarity	54 (1.9%)	56 (2%)	110 (3.8%)
Empathic solidarity	125 (4.4%)	21 (0.7%)	146 (5.1%)
Compassionate solidarity	732 (25.6%)	466 (16.3%)	1198 (41.8%)
Solidarity (no subtype)	41 (1.4%)	53 (1.9%)	94 (3.3%)
Total for solidarity	**1064 (37.2%)**	**784 (27.4%)**	**1848 (64.5%)**
Group-based anti-solidarity	10 (0.3%)	197 (6.9%)	207 (7.2%)
Exchange-based anti-solidarity	0 (0%)	48 (1.7%)	48 (1.7%)
Empathic anti-solidarity	17 (0.6%)	3 (0.1%)	20 (0.7%)
Compassionate anti-solidarity	8 (0.3%)	80 (2.8%)	88 (3.1%)
Anti-solidarity (no subtype)	5 (0.2%)	19 (0.7%)	24 (0.8%)
Total for anti-solidarity	**40 (1.4%)**	**347 (12.1%)**	**387 (13.5%)**
Mixed	60 (2.1%)	101 (3.5%)	161 (5.6%)
None	273 (9.5%)	195 (6.8%)	468 (16.3%)
Instances in total	**1437 (50.2%)**	**1427 (49.8%)**	**2864**

(a) Distribution of labels by target group.

Annotation IV

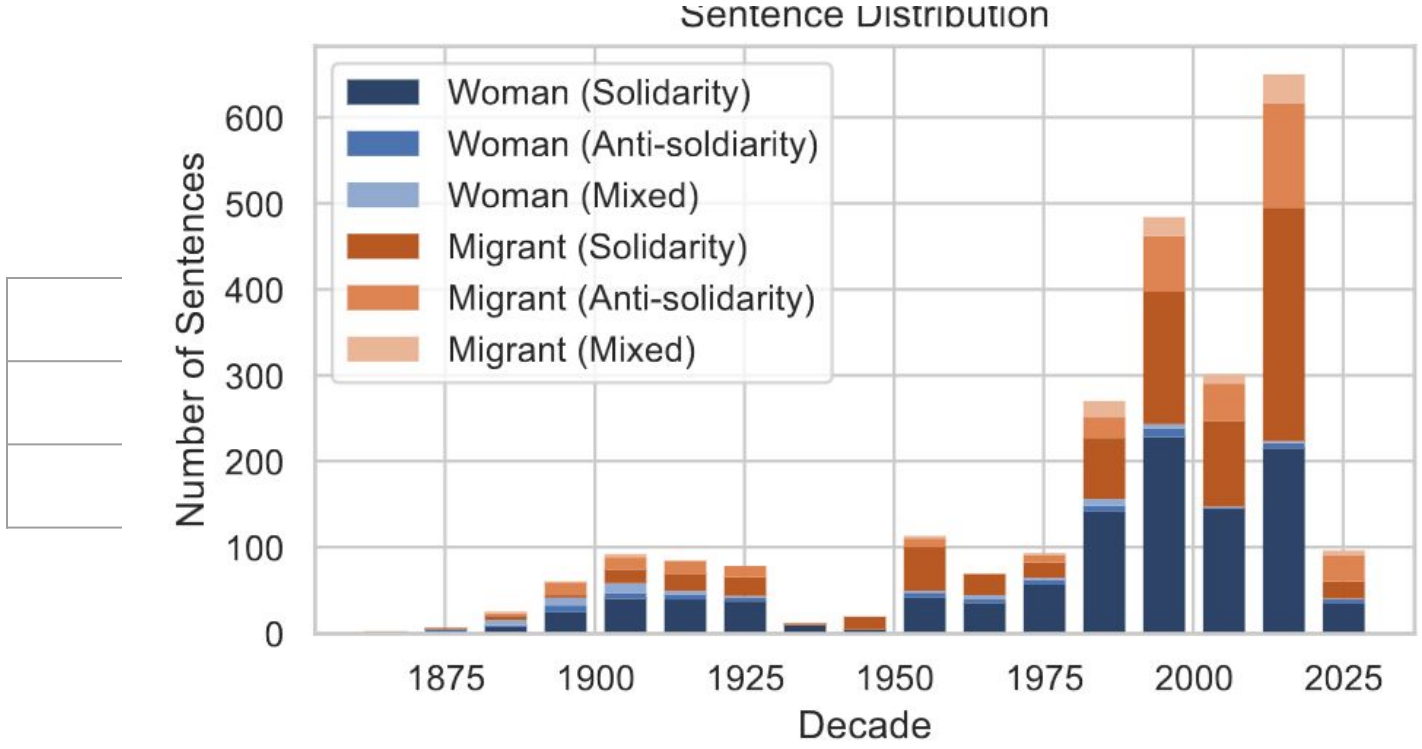


Figure 4: Distribution of instances in the human annotated dataset across time and target groups.

Annotation V: Examples

Gold Standard	Translation of the Original German Text
(1) **Compassionate solidarity towards women (June 29, 1961)**	"In connection with § 1708 BGB, the Bundestag has set the age of 18 as the limit for the obligation to provide maintenance. In the transitional provisions, this stipulation has been repealed for those who had already reached the age of 16 on January 1, 1962. My faction finds this regulation unfair, as it would exempt significant groups of people from this maintenance obligation. **Especially women who have made great efforts to send their children to higher education, for example, would have to bear these costs alone.** [...]"
(2) **Exchange-based anti-solidarity towards migrants (Apr. 19, 2018)**	"[...] **Let me also add: Migration is not necessarily successful – you always act as if that is great – it can fail, and it fails in particular when the immigrants' qualifications are low.** In 2013, before the so-called refugee wave, 40 percent of immigrants from non-EU countries had no qualifications. Since the wave of refugees, stabbings have increased by 20 percent, and we have imported anti-Semitism in the country. Does this make for an outstandingly successful migration?"
(3) **Mixed stance towards migrants (Feb. 2, 1982)**	"[...] We must accept that in a few years we will again need a higher number of foreign workers in the Federal Republic, as Mr. Urbaniak hinted earlier. **In reality, therefore, we must commit to effective integration, which admittedly requires [...] that there can be no exceptions, no alternative, regarding the recruitment stop and the prevention of illegal immigration**. [...]"
(4) **None case (women) (June 17, 2015)**	"[...] 'We want to be free people!' There is probably no better phrase to open today's debate here in the German Bundestag about the popular uprising of 1953. [...] **We remember women and men who, 62 years ago, showed great courage because they wanted to change the course of their country's development and their own lives, because they wanted to be free people.**"

Table 1: Example sentences from our dataset showing (anti-)solidarity towards women/migrants. Bold text is the main sentence, the other sentences are for context. Original German texts are available in Table 3 in the Appendix.

Models + Evaluation

- BERT, 110m parameters, trained on 1500 instances
- SBERT, based on BERT, also trained on 1500 instances
- GPT-3.5 (“ChatGPT”)
- GPT-3.5 fine-tuned
- GPT4

- Test set: ~430 test instances
- We repeat on 3 different splits of train/dev/test
- Evaluation Metric: Macro-F1

Models + Evaluation

For GPT

Analyze the following German text and classify it into one of the high-level categories regarding migrants (refugees — Flüchtlinge, expellees — Vertriebene, asylum seekers — Asylbewerber, immigrants — Einwanderer, and other migrant categories within Germany): SOLIDARITY, ANTI-SOLIDARITY, MIXED, or NONE. If applicable, further specify by choosing the most appropriate subtype (EMPATHIC, EXCHANGE-BASED, GROUP-BASED, COMPASSIONATE) within SOLIDARITY or ANTI-SOLIDARITY. Begin your response by providing the high-level category and then the subtype, if applicable.

- SOLIDARITY: Involves expressions that promote understanding, support, and unity with different groups or individuals (migrants in our case), often emphasizing shared goals, compassion, mutual assistance, and empathic understanding. Consider cases with even slight expressions of solidarity, regardless of the main topic of the text.
- ANTI-SOLIDARITY: Entails expressions that show opposition, disregard, or exclusion towards certain groups or individuals (migrants in our case). This includes emphasizing differences, denying the need for support or assistance, highlighting unequal exchanges between groups, and disregarding the unique characteristics or needs of certain groups. Even slight expressions of anti-solidarity should be considered, irrespective of the primary focus of the text.
- MIXED: A mixed stance toward migrants is characterized by the presence of both supportive and opposing expressions within the same text. This stance emerges in discussions where acknowledgment of migrants' rights, contributions, or needs is juxtaposed with limitations, conditions, or reservations that counteract or diminish the initial support. Key features of a mixed stance include (but are not limited with): conditional hospitality and selective support; balanced policies (e.g. improve the situation of migrants already within the country, while simultaneously seeking to regulate or limit further influx); expressions of empathy or concern for migrants' hardships, contrasted with discussions on practical constraints, such as societal integration challenges, or national security concerns.
- NONE: Texts which neither express solidarity nor anti-solidarity toward migrants in Germany, reflecting a neutral position or the absence of any specific stance. The absence of overt support or opposition does not automatically lead to a NONE classification; subtle cues or implicit messages may still align with solidarity or anti-solidarity categories.

If the text falls into SOLIDARITY or ANTI-SOLIDARITY, please specify further by choosing the most appropriate subtype from the following, after the initial high-level classification.

For SOLIDARITY: EMPATHIC SOLIDARITY, EXCHANGE-BASED SOLIDARITY, GROUP-BASED SOLIDARITY, COMPASSIONATE SOLIDARITY. Definitions:

- EMPATHIC SOLIDARITY: Is coded when a group is different from others and this should be recognized, supported, valued. In applying empathic solidarity to migrants, this can be expressed by (but not limited with): recognition of diversity and individuality; emphasis on the importance of preserving migrants' identities when integrating them into new communities; advocating for the right to live authentically without fear of persecution or discrimination; challenging stereotypes, prejudices against migrants.
- EXCHANGE-BASED SOLIDARITY: Is coded when a speaker refers to the usefulness of 'exchange partners' in terms of their actual or future contributions (economic, cultural, or social, etc.) or willingness to contribute. In applying exchange-based solidarity to migrants, this can be expressed by (but not limited with): emphasis on the importance of migrants' work, skills, and cultural diversity as essential for the host society; support for migrants which is framed as an investment in individuals who contribute to the community.
- GROUP-BASED SOLIDARITY: Is coded when solidarity is based on the idea of unity and support among members of a group, driven by shared characteristics, goals, interests, values and norms, or common rights and duties. The support might be driven by shared characteristics or challenges aiming at broader societal change; fostering inclusivity, equality, societal cohesion (difference from compassionate solidarity). In applying group-based solidarity to migrants, this can be expressed by (but not limited with): a unified effort to address and advocate for migrants' rights, equality, and representation; advocacy aimed at ensuring migrants' full integration; active stance against discrimination and xenophobia.
- COMPASSIONATE SOLIDARITY: Emphasizes providing support to marginalized, disadvantaged, or vulnerable groups, focusing on aid without expecting anything in return. It involves recognizing vulnerabilities, advocating for assistance to alleviate hardships, and offering support purely based on need. While not all indicators must be present, the core of compassionate solidarity lies in acknowledging and

Results

Model	Method	Fine-grained (high-level) W	Fine-grained (high-level) M
BERT	With context	0.02 (0.23)	0.19 (0.33)
	No context	0.03 (0.34)	0.03 (0.39)
SBERT	With context	0.05 (0.42)	0.20 (0.43)
	No context	0.04 (0.39)	0.18 (0.40)
Human upper bound		0.48 (0.72)	0.56 (0.78)

Results

Model	Method	Fine-grained (high-level) W	Fine-grained (high-level) M
GPT-3.5 base	0-shot	0.15 (0.46)	0.19 (0.48)
	Few-shot	0.12 (0.41)	0.27 (0.50)
	No context	0.15 (0.46)	0.20 (0.42)
BERT	With context	0.02 (0.23)	0.19 (0.33)
	No context	0.03 (0.34)	0.03 (0.39)
SBERT	With context	0.05 (0.42)	0.20 (0.43)
	No context	0.04 (0.39)	0.18 (0.40)
Human upper bound		0.48 (0.72)	0.56 (0.78)

Results

Model	Method	Fine-grained (high-level) W	Fine-grained (high-level) M
GPT-3.5 fine-tuned	0-shot	0.18 (0.45)	0.27 (0.53)
	Few-shot	0.22 (0.47)	0.28 (0.48)
	No context	0.14 (0.33)	0.26 (0.40)
GPT-3.5 base	0-shot	0.15 (0.46)	0.19 (0.48)
	Few-shot	0.12 (0.41)	0.27 (0.50)
	No context	0.15 (0.46)	0.20 (0.42)
BERT	With context	0.02 (0.23)	0.19 (0.33)
	No context	0.03 (0.34)	0.03 (0.39)
SBERT	With context	0.05 (0.42)	0.20 (0.43)
	No context	0.04 (0.39)	0.18 (0.40)
Human upper bound		0.48 (0.72)	0.56 (0.78)

Results

Model	Method	Fine-grained (high-level) W	Fine-grained (high-level) M
GPT-4	0-shot	**0.37 (0.60)**	**0.42 (0.73)**
	Few-shot	**0.37 (0.54)**	**0.43 (0.63)**
	No context	0.22 (0.38)	0.33 (0.42)
GPT-3.5 fine-tuned	0-shot	0.18 (0.45)	0.27 (0.53)
	Few-shot	0.22 (0.47)	0.28 (0.48)
	No context	0.14 (0.33)	0.26 (0.40)
GPT-3.5 base	0-shot	0.15 (0.46)	0.19 (0.48)
	Few-shot	0.12 (0.41)	0.27 (0.50)
	No context	0.15 (0.46)	0.20 (0.42)
BERT	With context	0.02 (0.23)	0.19 (0.33)
	No context	0.03 (0.34)	0.03 (0.39)
SBERT	With context	0.05 (0.42)	0.20 (0.43)
	No context	0.04 (0.39)	0.18 (0.40)
Human upper bound		0.48 (0.72)	0.56 (0.78)

Analysis

We use GPT4 to label our data

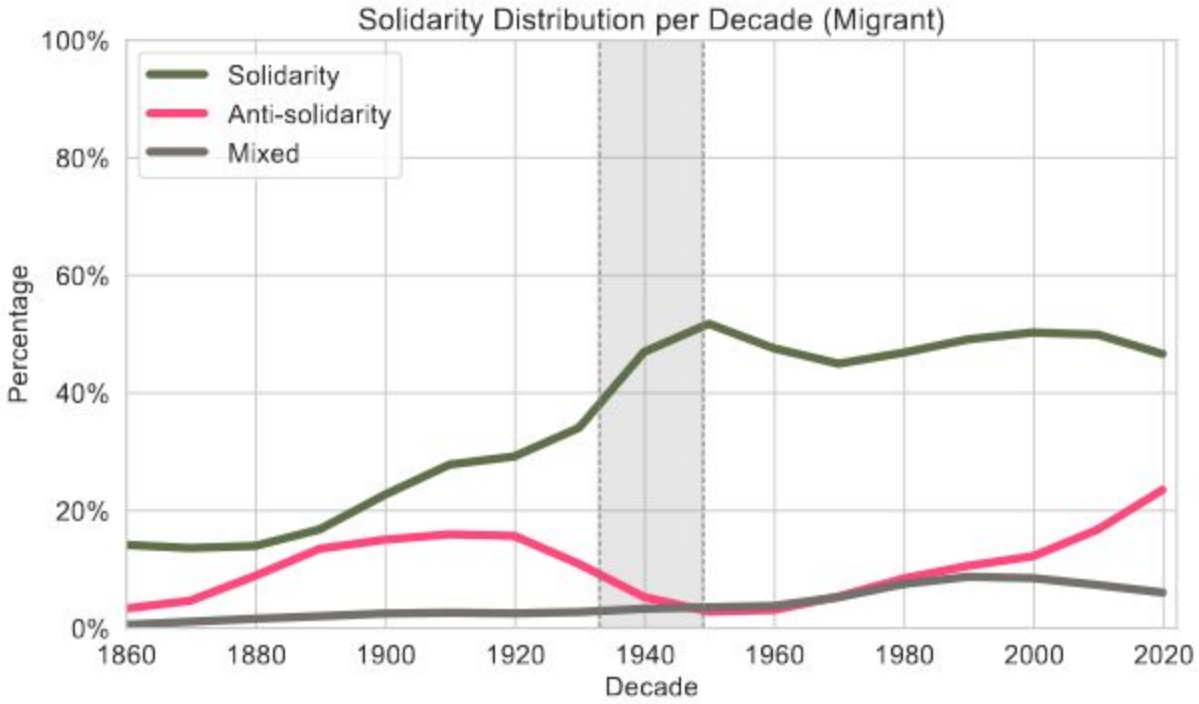
- over time
- large-scale
- focus on migrants

We annotate:

- 18k instances
- costs: ~500 Euro

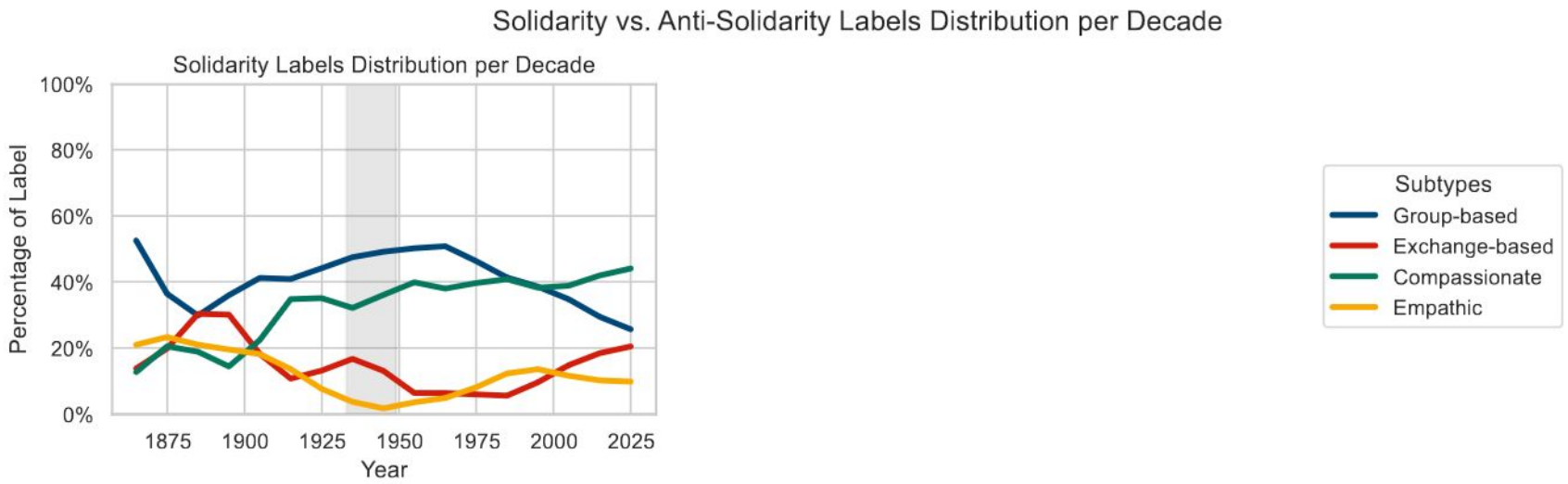
Analysis

How does (anti-)solidarity change over time?



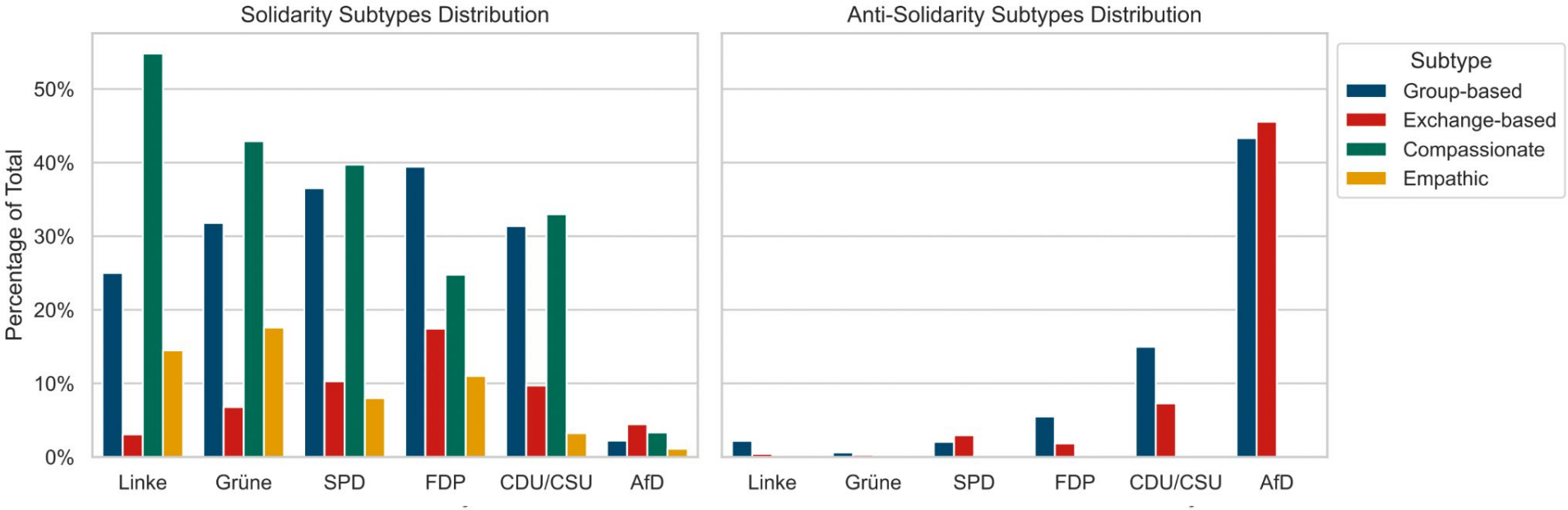
Analysis

How have solidarity and anti-solidarity frames evolved over time?



Analysis

How are solidarity and anti-solidarity frames represented across political parties?



Analysis

What are trends across individual keywords?

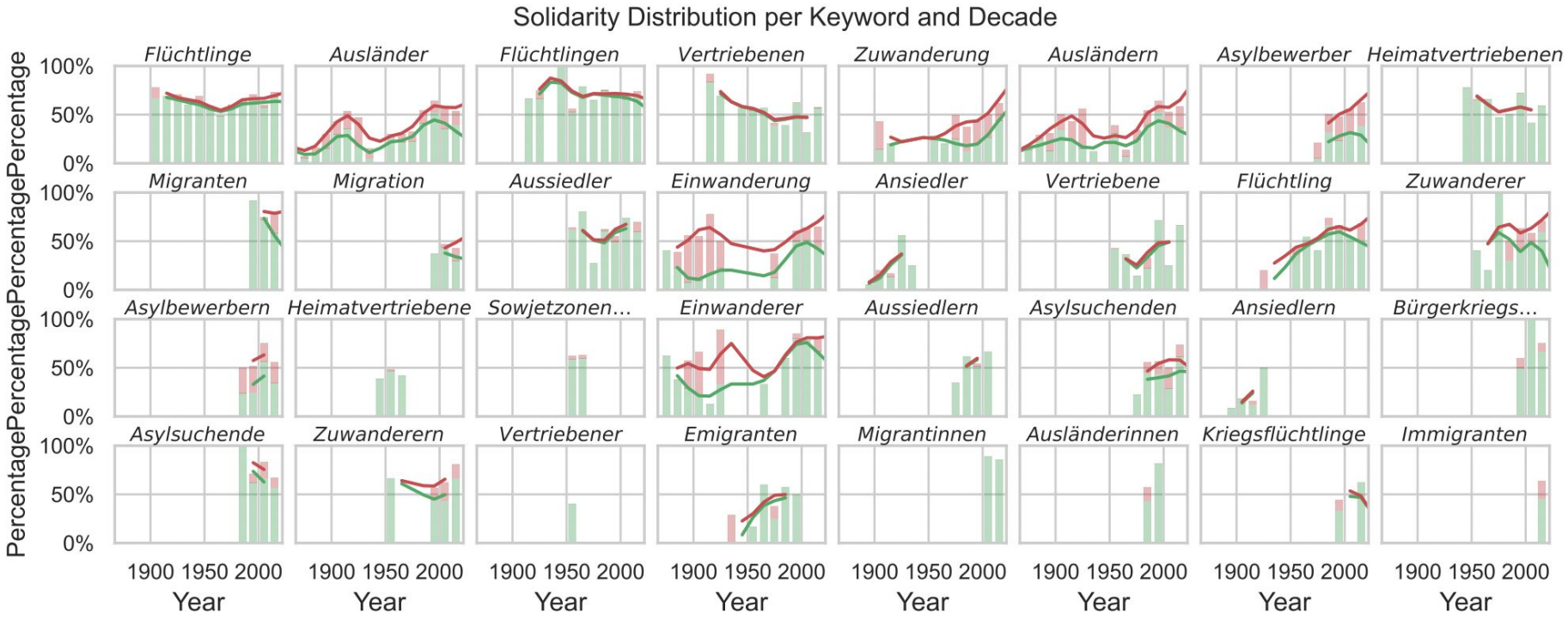


Figure 11: Percentage of sentences showing solidarity/anti-solidarity per decade for all `Migrant` keywords. The keywords are sorted by frequency, which means that the reliability decreases towards the bottom-right.

Summary

- Fine-grained annotation of social solidarity in German parliamentary proceedings
- Hard task, decent but no excellent agreements among annotators
- GPT4 is the only model that can partly compete with human annotators (being much cheaper)
- Model annotations allow us to address questions regarding trends of solidarity in German society across ~155 years

Limitations

- Only German
- Fair comparison of models?
- Annotator biases?
- Is the model reliable?

LLM-Based Literary Translation & Evaluation

Motivation

- **LLMs revolutionize AI**
- **Translation as “holy grail” of NLP**
- **But: can LLMs also tackle literature translation?**
  - **Complex language**
  - **Historical language**
  - **Little training data**
- **Background:**
  - **Scientific interest**
  - **Limits of AI/LMMs**
  - **Help for professional translators?**

Research Questions

- **RQ1: How well do LLMs perform in literary translation (vs. human)?**
- **RQ2: How can we effectively evaluate literary translation?**

Procedure

- **1) We obtain literature data (source texts)**
- **2) We also obtain human references: translations by human experts (professional translators)**
- **3) We obtain different LLMs**
- **4) We use the LLMs to translate the source texts**
- **5) We evaluate / annotate the human and LLM translations**
  - **quality assessment**

Example

Source:

Die Schwester eilte zur Mutter und hielt ihr die Stirn. Der Vater schien durch die Worte der Schwester auf bestimmtere Gedanken gebracht zu sein, hatte sich aufrecht gesetzt, spielte mit seiner Dienermütze zwischen den Tellern, die noch vom Nachtmahl der Zimmerherren her auf dem Tische standen, und sah bisweilen auf den stillen Gregor hin.

Example

Translation A:
The sister rushed to her mother and held her forehead. The father seemed to have been brought to certain thoughts by the sister’s words, had stood up, played with his servant’s mite between the plates that had been on the table since the hosts’ dinner, and sometimes looked at the silent Gregory.

Translation B:
The sister hurried to the mother and held her forehead. The father seemed to have been brought to more definite thoughts by the sister's words, had sat up straight, played with his servant’s cap between the plates that still stood on the table from the roomers' evening meal, and occasionally looked over at the silent Gregor.

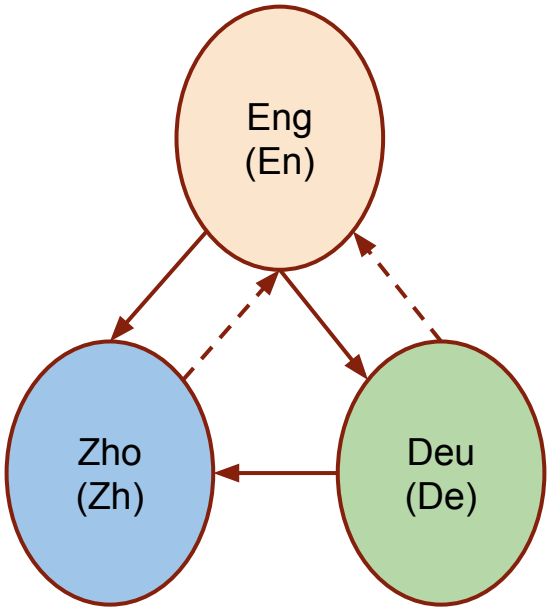
Example

Translation C:
The sister rushed to the mother and cradled her forehead. The father’s thoughts seemed to have cleared in the aftermath of the sister’s words; he sat up straight, played with the cap of his uniform among the dishes that still lay on the table from the boarders’ supper, and from time to time glanced over at Gregor’s inert form.

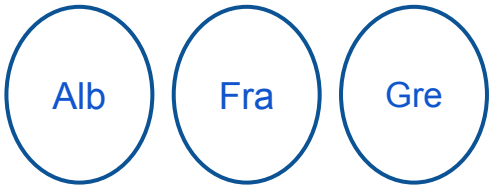
Translation D:
The sister hurried to her mother and held her forehead. The father seemed to have been brought to more definite thoughts by the sister's words, had sat upright, played with his servant's cap between the plates that were still on the table from the chambermaids' supper, and occasionally looked at the quiet Gregor.

Dataset creation: (Beyond) **Paragraph-level** dataset

Paragraph level

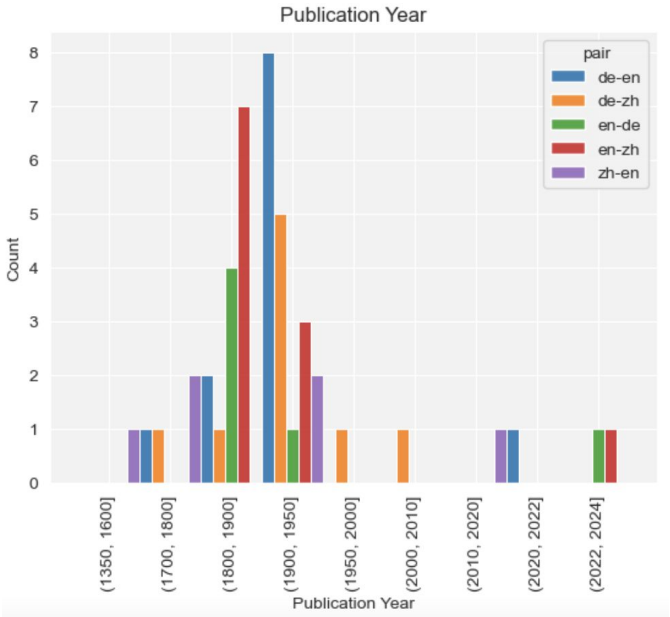


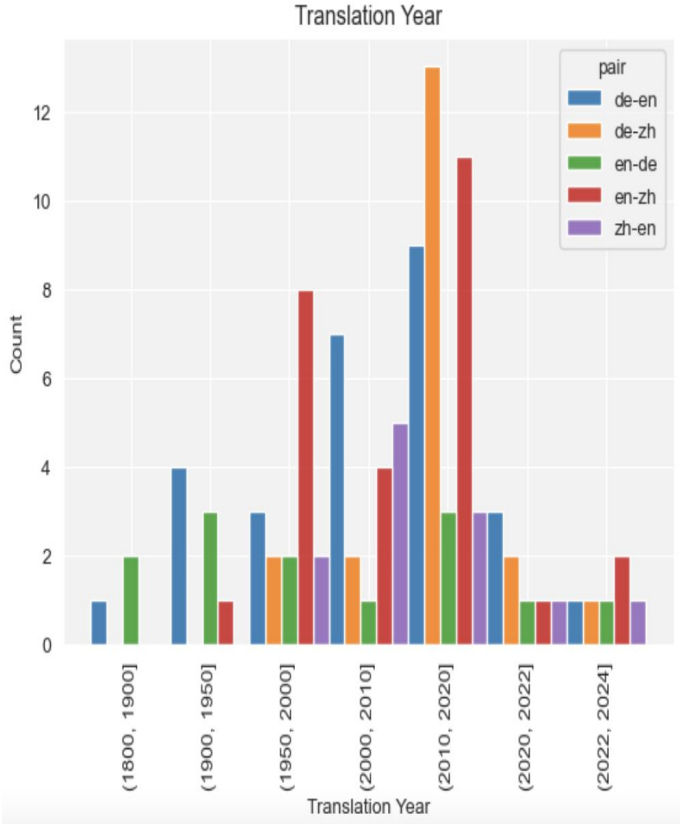
Chapter level



#Language pairs	**5**
Source paragraphs	Published books (open-source/purchase) *(Follow the concept of fair use)*
Target paragraphs	≥ 2 human translations from published classic works and 1 for contemporary works.
Meta-info	**Publication years** of source and target texts
Number of systems	9: 5 LLMs + Google Translate + DeepL + 2*NMT (transformer)
Number of annotated instances	~ 3k: 50 – 60 paragraphs * 10-11 systems (+ 1-2 human translations) * 5 pairs

Distribution of Publication Year & Translation Year (all sources)





Origin of Translations

- **Project Gutenberg: https://www.projekt-gutenberg.org/**
- **Public reading samples (e.g., Amazon …)**
- **Purchase**

Note: we pay more attention to follow the concept of fair use (neglected by other works)

- proper citations of the source paragraph
- only use the dataset for research purpose
- (future) fair use agreement to access our dataset

An example: the source paragraph (Eng) and different human translations (Deu)

Src:
I returned to my book—Bewick's **History of British Birds**: the **letterpress** thereof I cared little for, generally speaking; and yet there were certain introductory pages that, child as I was, I could not pass quite as a blank. [...]

Human Translation 1:
Ich kehrte zu meinem Buche zurück – Bewicks **Geschichte von Englands gefiederten Bewohnern**; im allgemeinen kümmerte ich mich wenig um den **gedruckten Text des Werkes**, und doch waren da einige einleitende Seiten, welche ich, obgleich nur ein Kind, nicht gänzlich übergehen konnte. [...]

Human translation 2:
Ich wandte mich wieder meinem Buch zu – Bewicks **Britischer Vogelkunde**. **Um den Text** kümmerte ich mich im Allgemeinen recht wenig, doch gab es ein paar Seiten, die ich selbst als Kind nicht einfach überspringen konnte. [...]

Book Title

Term

An Example

Src:
[...] and yet there were **certain introductory pages** that, child as I was, I could not **pass quite as a blank**. They were those which treat of the haunts of sea-fowl; of "the solitary rocks and promontories" by them only inhabited, of the coast of Norway, studded with isles from its **southern extremity**, the **Lindeness**, or Naze, to the North Cape— [...]

Human Translation 1:
[...] im allgemeinen kümmerte ich mich wenig um den gedruckten Text des Werkes, und doch waren da **einige einleitende Seiten**, welche ich, obgleich nur ein Kind, **nicht gänzlich übergehen konnte**. Es waren jene, die von den Verstecken der Seevögel handelten, von jenen einsamen Felsen und Klippen, welche nur sie allein bewohnen, von der Küste Norwegens, die von ihrer **äußersten südlichen Spitze**, dem **Lindesnäs** bis zum Nordkap mit Inseln besäet ist. [...]

Human translation 2:
[...] Um den Text kümmerte ich mich im Allgemeinen recht wenig, doch gab es **ein paar Seiten**, die ich selbst als Kind **nicht einfach überspringen konnte**. **Es handelte sich um die Einleitung**, in der von den Schlupfwinkeln der Seevögel die Rede war; von den »einsamen Felsen und Klippen«, die einzig und allein von diesen bevölkert werden; von der Küste Norwegens, die von ihrem **südlichsten Punkt**, dem Kap **Lindesnes** oder Naze, bis zum Nordkap mit Inseln übersät ist und [...]

w/o Sentence splitting

Location

Src:
I returned to my book—Bewick's History of British Birds: the letterpress thereof I cared little for, generally speaking; and yet there were **certain introductory pages** that, child as I was, I could not **pass quite as a blank**. [...]

Human Translation 1:
Ich kehrte zu meinem Buche zurück – Bewicks Geschichte von Englands gefiederten Bewohnern; im allgemeinen kümmerte ich mich wenig um den gedruckten Text des Werkes, und doch waren da **einige einleitende Seiten**, welche ich, obgleich nur ein Kind, **nicht gänzlich übergehen konnte**. [...]

Human translation 2:
Ich wandte mich wieder meinem Buch zu – Bewicks Britischer Vogelkunde. Um den Text kümmerte ich mich im Allgemeinen recht wenig, doch gab es **ein paar Seiten**, die ich selbst als Kind **nicht einfach überspringen konnte**. **Es handelte sich um die Einleitung** [...]

Publication year: 1847
Source: Project Gutenberg

Publication year: 1864
Source: Project Gutenberg

Publication year: 2020
Source: Reclam reading sample

Book & system selection

System selection → diversity in translation quality

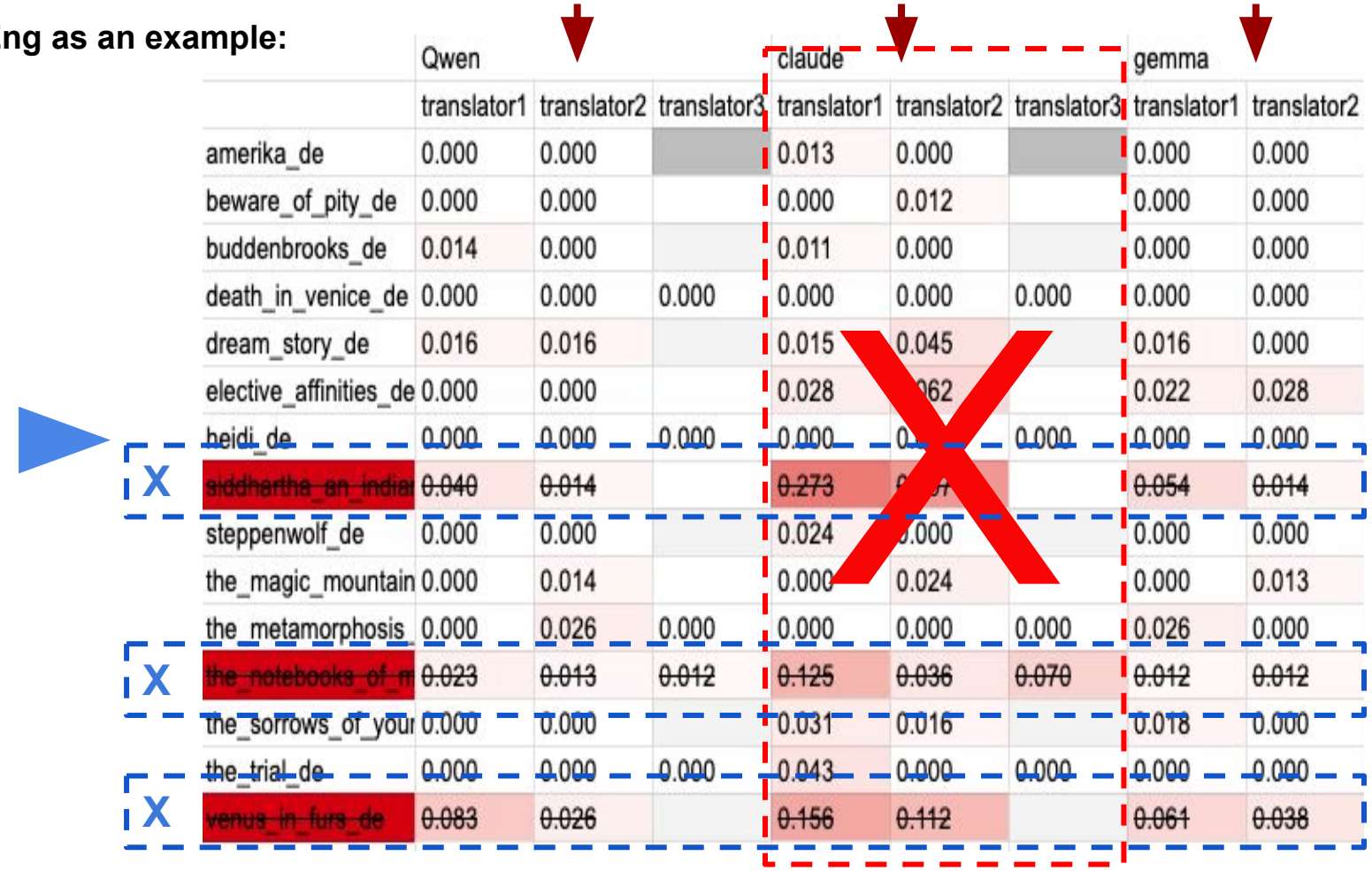
- Commercial: Google Translate, DeepL
- Previous SOTA: NLLB-3.3b, m2m_100_1.2b (facebook)
- LLM candidates:
  - **GPT-4o**
  - Mistral
  - Claude
  - GPT-3.5
  - **Qwen (chinese)**
  - **Unbabel_Tower**
  - **Google_gemma**
  - **Meta-llama 3**
  - …

Book & system selection

Direction Deu-Eng as an example:

Select models

Select books



	Qwen			claude			gemma	
	translator1	translator2	translator3	translator1	translator2	translator3	translator1	translator2
amerika_de	0.000	0.000		0.013	0.000		0.000	0.000
beware_of_pity_de	0.000	0.000		0.000	0.012		0.000	0.000
buddenbrooks_de	0.014	0.000		0.011	0.000		0.000	0.000
death_in_venice_de	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
dream_story_de	0.016	0.016		0.015	0.045		0.016	0.000
elective_affinities_de	0.000	0.000		0.028	062		0.022	0.028
heidi_de	0.000	0.000	0.000	0.000	0.	0.000	0.000	0.000
X siddhartha_an_indian	0.040	0.014		0.273			0.054	0.014
steppenwolf_de	0.000	0.000		0.024	.000		0.000	0.000
the_magic_mountain	0.000	0.014		0.000	0.024		0.000	0.013
the_metamorphosis_	0.000	0.026	0.000	0.000	0.000	0.000	0.026	0.000
X the_notebooks_of_m	0.023	0.013	0.012	0.125	0.036	0.070	0.012	0.012
the_sorrows_of_your	0.000	0.000		0.031	0.016		0.018	0.000
the_trial_de	0.000	0.000	0.000	0.043	0.000	0.000	0.000	0.000
X venus_in_furs_de	0.083	0.026		0.156	0.112		0.061	0.038

Annotation details

- **Annotated Samples**
  1. **Individual paragraph**
  2. **Consecutive paragraphs** from the same chapter → beyond paragraph level

- **Annotators**
  - 5 Students with linguistics and translation studies backgrounds
  - Native speaker in source and/or target languages
  - Expert translators with publication record (small amount due to budget limit)
  - Professional translators from Upwork?

Annotation details

Annotation guideline

MQM (Multidimensional Quality Metrics) [Official site + WMT 2023]

- Highlight error span:
  - Terminology (Mistranslation, Inconsistency)
  - Accuracy (Addition/Omission/Misnomer, Mistranslation-[Overly literal, Temporal effect]),
  - Fluency (Punctuation/Spelling/Grammar, Inconsistency, Coherence)
  - Style (Awkwardness/Unidiomatic, Register, Inconsistent)
  - Non-translation: sentence which is too badly garbled to permit reliable identification of individual errors.
- Select error severity: major, minor

Annotation details

MQM Annotation → Score

MQM (Multidimensional Quality Metrics)

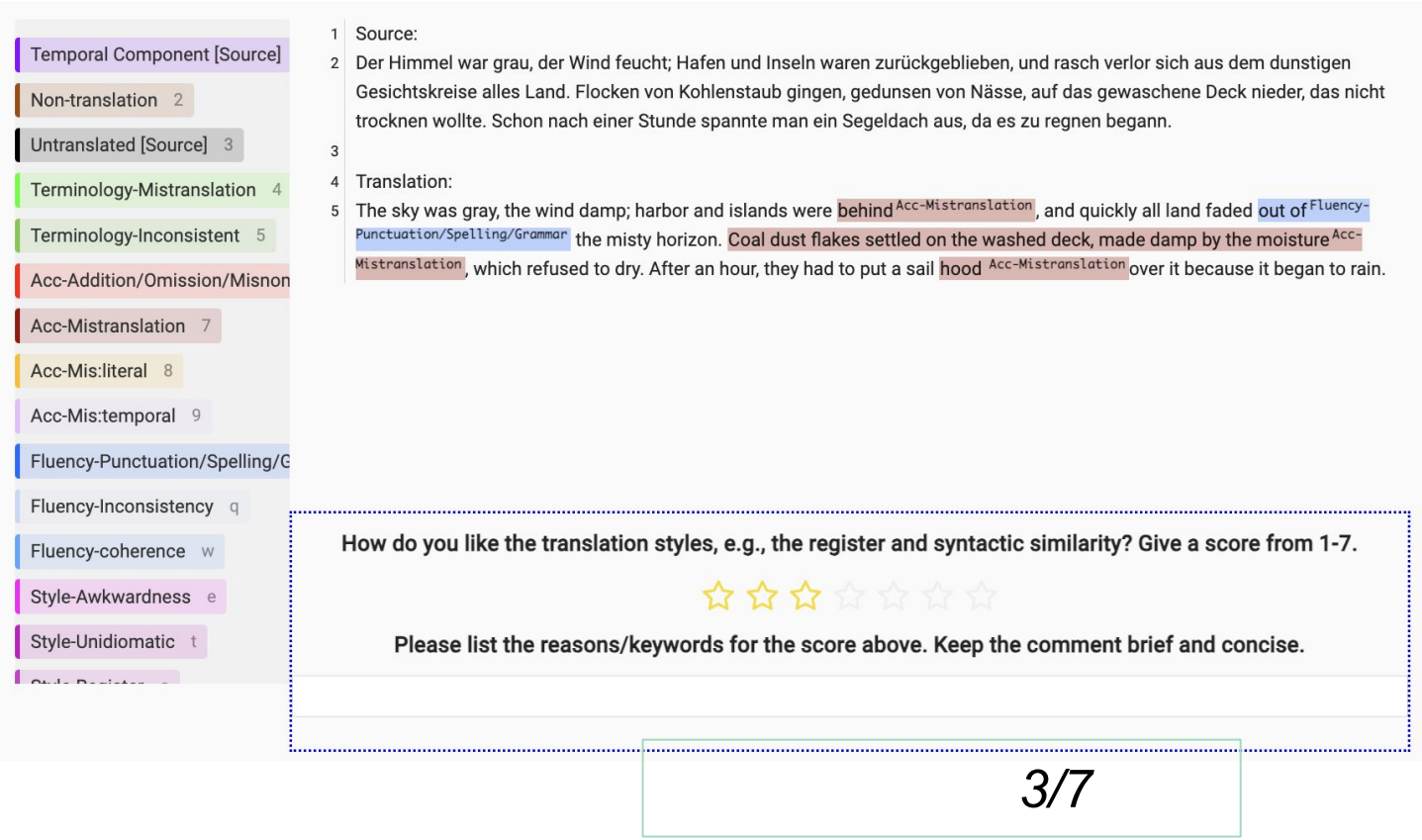
- Highlight error span:
  - **Non-translation**
- Error severity: **major**, **minor**

Severity	Score mapping
minor	-1
major	-5
Non-translation	-25

Annotation details

Reference-less

MQM annotation



Annotation results: Statistics

- **Initial stage:**
  - **Get familiar with the guidelines**
  - **Refine the guidelines based on annotators' feedback**

- **Annotated + revised samples + annotator feedback**

pair	annotated instances	number of tokens (source)	number of sentences (source)	number of tokens (translation)	number of sentences (translation)
en-zh	225	132.3	4.2	190.7	4.9
zh-en	97	178.7	6.6	138.4	8.2
de-en	284	158.3	6.5	202.0	7.1
en-de	258	127.5	4.1	131.2	4.5
de-zh	273	169.9	5.7	216.9	6.0

Annotation results: Agreements

Pairwise annotation agreement between two annotators			
Pair	**MQM score mapping**	**Span-level**	**Instances**
En-Zh	0.771	0.261	22
Zh-En	0.712	0.334	44
De-En	0.565	0.285	24
En-De	0.718	0.316	24
De-Zh	-	-	

Note: The agreement for MQM score mapping is measured by Kendall's Tau; The span-level comparison is measured by character-level Cohen's kappa. We only consider the span label in this calculation.

Annotation results: System Rankings

Rank of systems based on our MQM annotation mapping score

Model	Open source?	Model size	De-En	En-De	De-Zh	En-Zh	Zh-En	Mean	Rank
Human 1,2,3	partial	86 billion neurons	-5.0	-7.3	-3.4	-6.0	-16.0	-7.5	1
DeepL	close	?	-11.1	-6.7	-7.8	-20.8	-49.8	-19.2	4
Google Translate	close	?	-6.8	-5.9	-17.1	-13.8	-25.2	-13.7	3
GPT-4o	close	175 billion	-5.0	-5.4	-7.1	-16.0	-17.0	-10.1	2
google_gemma	open	7 billion	-20.3	-40.4	-42.5	-55.7	-46.1	-41.0	8
meta-llama 3	open	8 billion	-10.0	-28.6	-31.9	-51.7	-36.0	-31.6	5
qwen2	open	7 billion	-17.7	-67.5	-11.4	-23.8	-47.6	-33.6	7
unbabel_tower	open	7 billion	-14.0	-25.8	-30.0	-47.3	-50.5	-33.5	6
nllb	open	3.3 billion	-33.4	-43.7	-38.3	-56.3	-66.8	-47.7	10
m2m	open	1.3 billion	-20.4	-30.7	-50.4	-66.6	-61.8	-46.0	9
Mean			-14.4	-26.2	-24.0	-35.8	-41.7		

Annotation results: System rankings

Rank of systems based on overall preference of translation style

Model	Open source?	Model size	De-En	En-De	De-Zh	En-Zh	Zh-En	Mean	Rank
Human 1,2,3	partial	86 billion neurons	4.8	5.9	6.3	6.9	5.8	5.9	1
DeepL	close	?	4.0	5.0	5.3	5.4	2.8	4.5	4
oogle Translate	close	?	4.3	5.4	4.1	6.0	5.1	5.0	3
GPT-4o	close	175 billion	5.1	5.4	5.7	5.7	5.8	5.5	2
google_gemma	open	7 billion	3.2	1.5	2.1	2.0	3.1	2.4	8
meta-llama 3	open	8 billion	3.6	2.4	3.1	1.8	3.9	3.0	6
qwen2	open	7 billion	3.2	1.2	5.1	5.2	2.8	3.5	5
unbabel_tower	open	7 billion	3.7	2.6	2.8	2.8	2.4	2.9	7
nllb	open	3.3 billion	2.0	2.3	2.0	1.4	1.2	1.8	10
m2m	open	1.3 billion	1.9	1.9	2.3	2.1	1.5	2.0	9

Annotation results

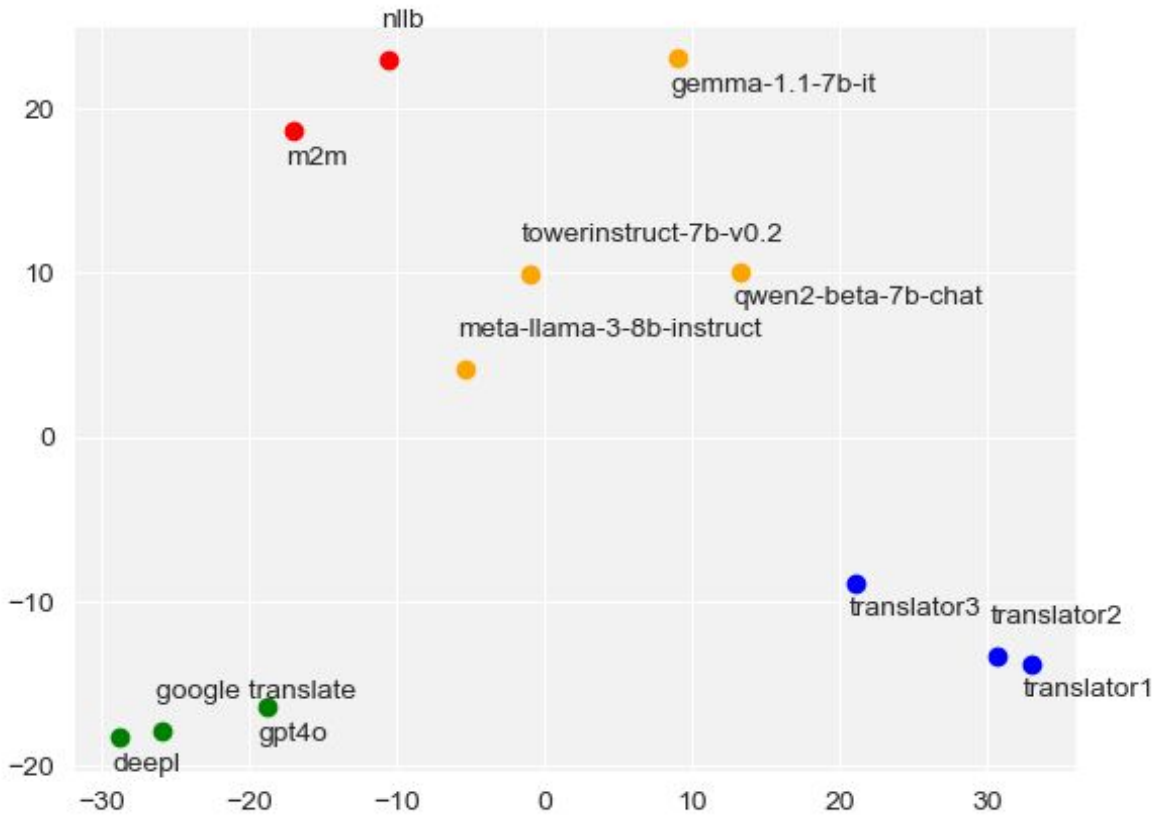
Error distribution over all language pair (% of error types aggregated over top3 systems vs. other LLMs)

	Top-3 outputs: human, gpt-4o, google translate				
Aspects	**de-en**	**en-de**	**de-zh**	**en-zh**	**zh-en**
Accuracy	45.3%	40.6%	66.0%	69.2%	40.2%
Style	36.7%	15.9%	18.6%	8.8%	5.4%
Terminology	1.0%	0.4%	-	11.3%	46.7%
Coherence (Fluency + consistency errors)	18.3%	41.7%	16.5%	11.9%	8.7%

	Other LLMs' outputs				
Aspects	**de-en**	**en-de**	**de-zh**	**en-zh**	**zh-en**
Accuracy	68.7%	56.0%	79.3%	74.1%	53.2%
Style	21.0%	13.7%	7.9%	8.0%	3.8%
Terminology	0.5%	0.4%	2.7%	9.0%	34.0%
Coherence (Fluency + consistency errors)	8.9%	28.1%	10.7%	9.0%	10.0%

Human vs. LLMs

Clusters of systems based on pairwise lexical overlap



Preferences + Automatic Evaluation Metrics

The percentage of instances where human translations are preferred over machine translations

	Human > LLMs			Human > LLMs (excluding GPT-4o, DeepL, Google Translate)		
	human MQM	style rating	Gemba-mqm (metric)	human MQM	stylerating	Gemba-mqm (metric)
de-en	47.6%	42.9%	9.5%	81.0%	71.4%	23.8%
de-zh	56.5%	47.8%	0.0%	87.0%	87.0%	17.4%
en-de	19.0%	42.9%	0.0%	81.0%	100.0%	47.6%
en-zh	81.0%	47.6%	14.3%	100.0%	90.5%	28.6%
zh-en	50.0%	50.0%	25.0%	100.0%	75.0%	25.0%

Summary

- **We introduced a complex annotation scheme for assessing the quality of literary translations**
- **And a dataset of literary paragraphs, together with their translations**
- **Assessed 9 different systems and up to 3 human translations**
- **Human translations are still top, however, GPT-4o rivals them**
- **Human translations are *different***

- **Limitations:**
  - **Data contamination**
  - **Quality of our human annotators**
  - **Ethical aspects:**
    - **Professional literary translators may not like this kind of research**

Discussion

- **Is error annotation appropriate for literature?**

pair	Human > LLMs (GPT-4o, DeepL, Google Translate, Qwen 2)						Human > LLMs (excluding GPT-4o, DeepL, Google Translate)				
	MQM	SQM	MQM-SQM	BWS	GEMBA (Orig)	GEMBA (Lit)	MQM	SQM	MQM+SQM	GEMBA (Orig)	GEMBA (Lit)
De-En	60.0%	60.0%	80.0%	73.3%	6.7%	0.0%	86.7%	80.0%	93.3%	26.7%	33.3%
En-De	35.0%	60.0%	65.0%	86.7%	0.0%	10.0%	95.0%	100.0%	100.0%	70.0%	60.0%
De-Zh	30.0%	25.0%	30.0%	95.0%	15.0%	20.0%	95.0%	90.0%	95.0%	45.0%	50.0%
En-Zh	45.8%	25.0%	50.0%	80.0%	0.0%	8.3%	87.5%	79.2%	87.5%	41.7%	50.0%

Table 3: Percentage of segments where human translations are preferred over machine translations per annotation scheme and GEMBA-MQM versions. GEMBA (Orig) and GEMBA (Lit) represent GEMBA-MQM (Original) and GEMBA-MQM (Literary) respectively.

THÄNK$!

https://nl2g.github.io/