A PUZZLE OF PERSPECTIVES:
how to align language technology with social science and democratic goals

Myrthe Reuver

IT University Copenhagen
25 April 2025

Who am I?

PhD from the Vrije Universiteit Amsterdam (2025), research topic: **argument mining for diverse perspectives in news recommendation.**

During the PhD: internship at LinkedIn, research visit at GESIS - Leibniz Institute for the Social Science, and a lot of karaoke.

Currently: researcher and engineer at Populytics, an Amsterdam start-up from TU Delft (more about that later!)

Fun fact: I used to be a local radio presenter!

Research interests

Opinion mining, responsible language technology, and interdisciplinary collaboration.

Are we measuring what we think we are measuring? 🤔

- evaluation and conceptualization

Why do we do science this way, and how can we do it differently?

- meta-scientific norms in NLP and beyond

How can we combine theory, methods, and real-life context?

- tackling the truly “wicked problems” that are **complex**, require **thoughtful** analyses, and have a positive societal **impact**.

Connecting disciplines and ideas

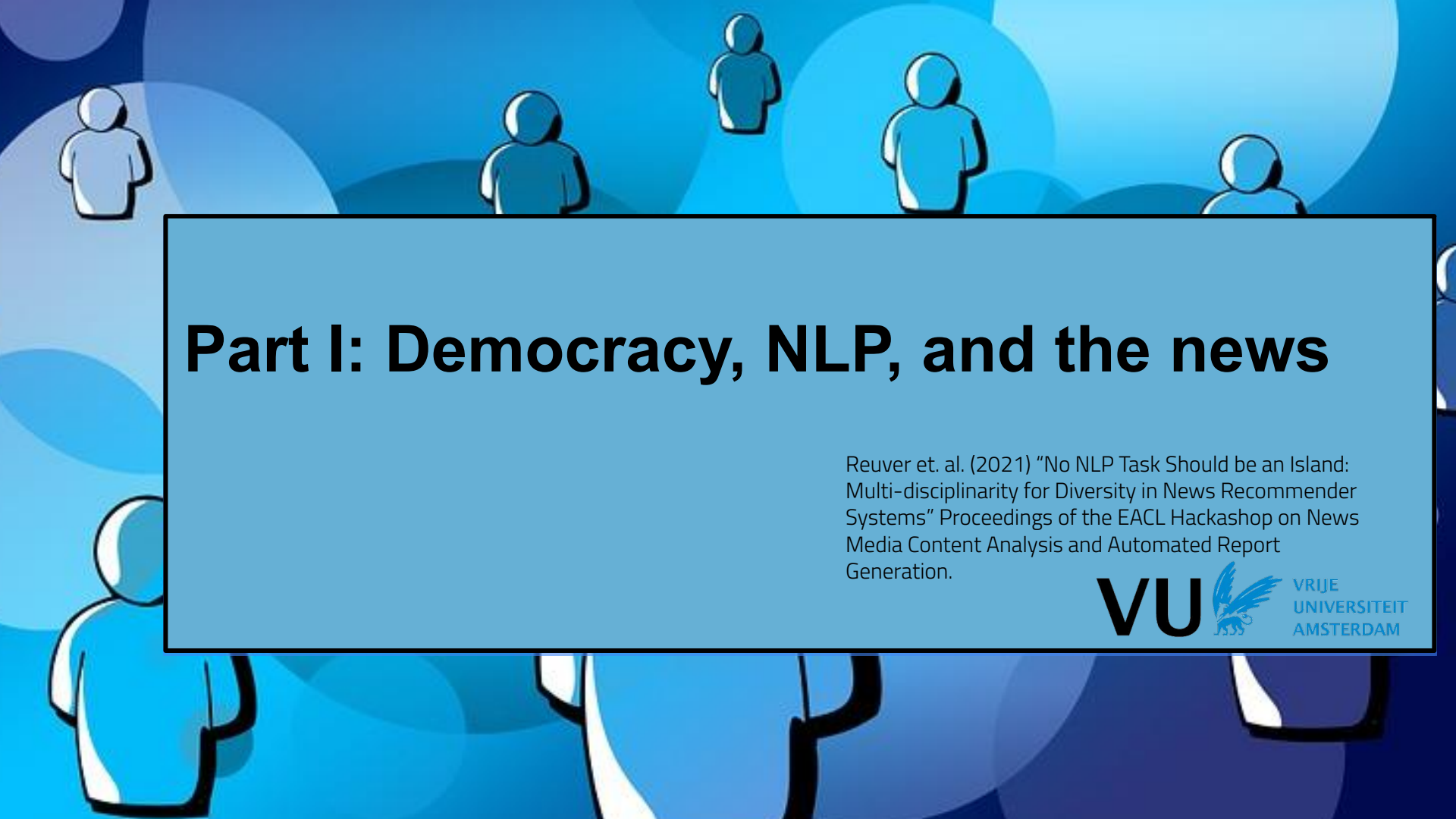
Today, I will highlight several of my projects on:

- viewpoint diversity and the **deliberative democracy**;
- stance detection and **responsible science**;
- sexism detection with experts and;
- argument analysis beyond stance

which contain **a connection** between social science and NLP,

an aim for **positive societal impact,**

but also highlight **difficulties** of doing so.

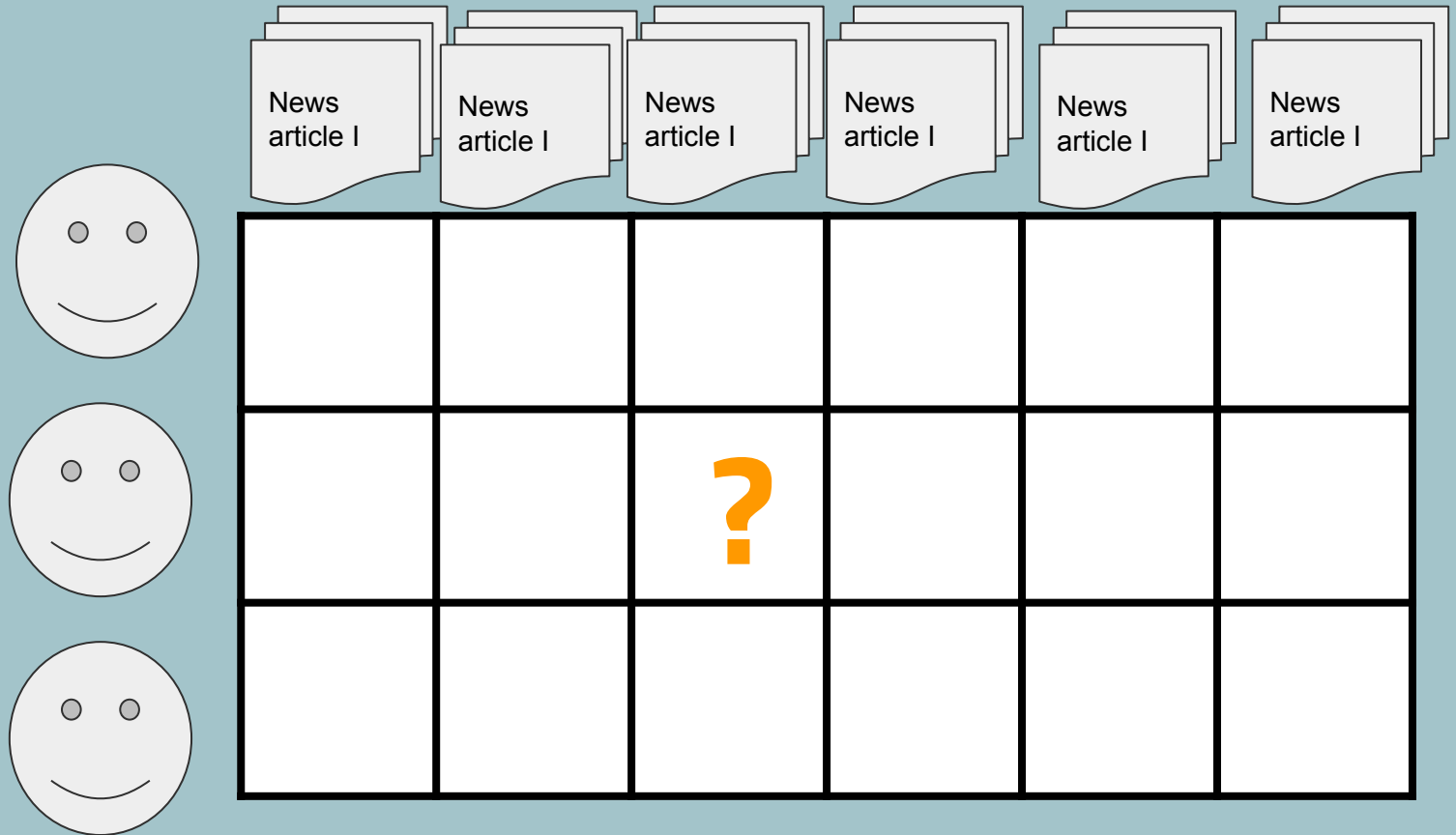


Part I: Democracy, NLP, and the news

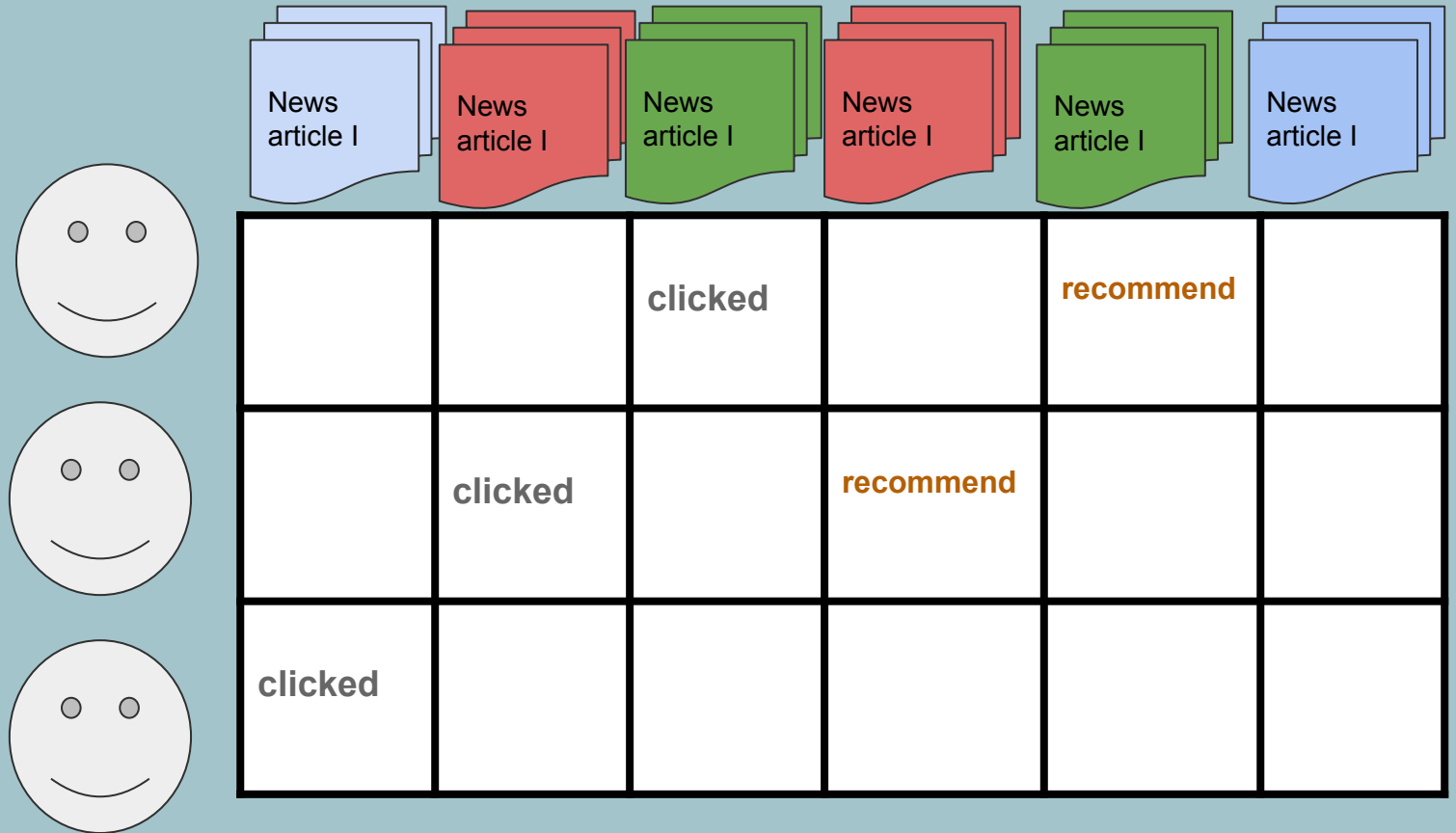
Reuver et. al. (2021) "No NLP Task Should be an Island: Multi-disciplinarity for Diversity in News Recommender Systems" Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation.

VU VRIJE UNIVERSITEIT AMSTERDAM

What is a (news) recommender system?



What is a (news) recommender system?



Optimizing in News Recommendation

Usually in RecSys: **click-accuracy** (as proxy for user interest).
Consequence: Showing users more of the same.

Could lead to ‘filter bubbles’ → potentially problematic for democracy.

But **why** ? And how can computational linguistics/NLP help?

My journey started with **asking experts (social scientists and theorists): What is democracy? What is the (complex) problem here?**

Models of Democracy 📖

- Theoretical models that define a functioning democracy.
- Deliberative model: democracy requires **rational debate** , and actors encountering **a diverse set of viewpoints and ideas** on (societal) issues.
- Helberger (2019) connected this and other models to news recommendation. Supporting a deliberative model of democracy = recommenders promoting **rational rather than emotional content**, and **diverse viewpoints** on issues.

Measuring health of recommendations

Optimization **beyond individual user satisfaction**, but on the **collective information environment** with new metrics bij Sanne Vrijenhoek:

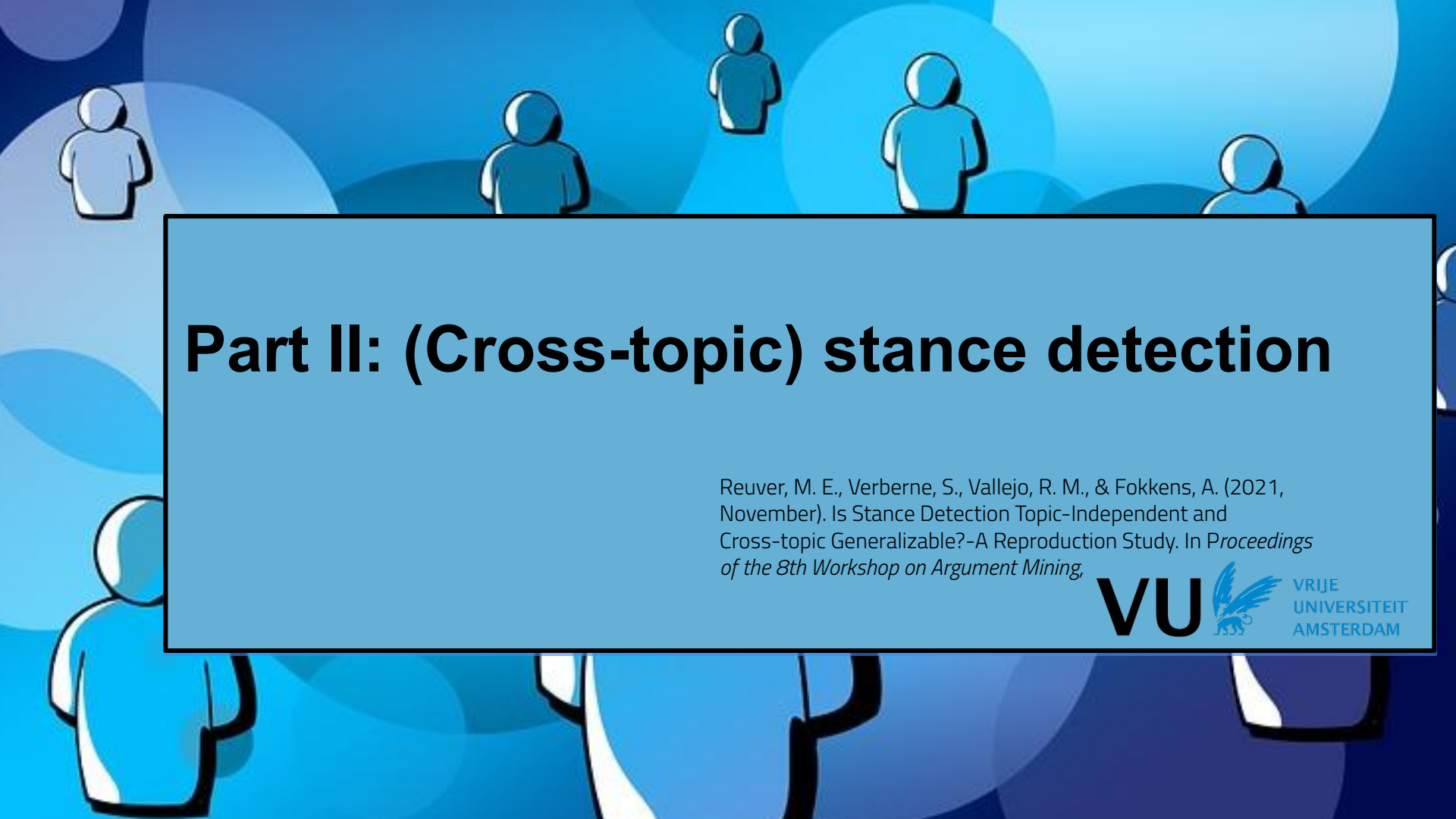
1) **Fragmentation:** shared public sphere
2) **Representation:** diverse actors and viewpoints
3) **Alternative Voices:** non-mainstream opinions
4) **Calibration:** personalization
5) **Affect:** emotional content

- **Deliberative debate** requires low affect, low fragmentation
- **Critical democracy model** (where viewpoints clashing is considered healthy) requires high affect.

Thinking about problem context before task

Theoretical models of democracy (such as the deliberative model) helps think of nuanced aspects going beyond task:

- Our ultimate goal is not e.g. detecting stances, but **supporting democracy.** That might mean some stances, viewpoints, or ideas (attacking democracy, or inherently violent stances) should **not** be recommended.
- We also need to not only detect stances, but **find diverse, different, or opposing** viewpoints and ideas.
- There is no singular answer for the “optimal” level of diverse opinions for democracy (!)



Part II: (Cross-topic) stance detection

Reuver, M. E., Verberne, S., Vallejo, R. M., & Fokkens, A. (2021, November). Is Stance Detection Topic-Independent and Cross-topic Generalizable?-A Reproduction Study. In P*roceedings of the 8th Workshop on Argument Mining,*

VU VRIJE UNIVERSITEIT AMSTERDAM

What is (going on with) stance?

Stance detection, common definition: **classification task** (on texts, often tweets) with labels Pro, Con, Neutral towards an issue or topic

“Abortion is a sin, and should never be practiced.”

Topic: Abortion, Stance: Con

Why stances?

- Built upon the linguistic phenomenon of **actors communicating their evaluation of targets**;

  placing **themselves and their targets on "dimensions in the sociocultural field"** (Du Bois, 2007).
- Directional (pro/con)

- Immediate **connected to (deliberative) debate and democratic decision making**

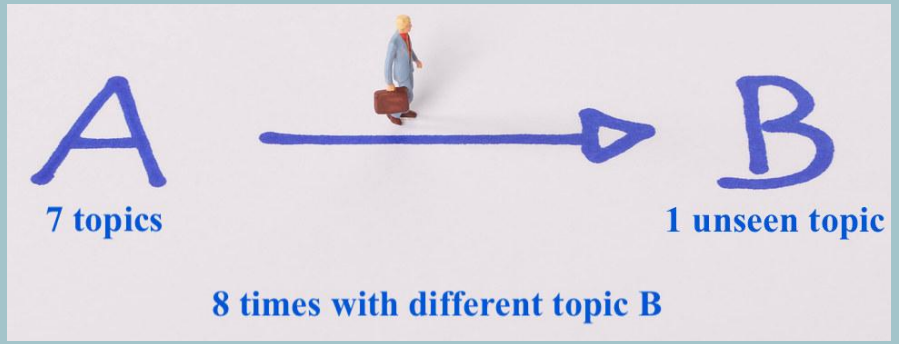
  (agree/disagree with laws, proposals, etc)

Reimers et. al. (2019): cross-topic stance detection

Train: 7 topics, test: 8th topic

Fine-tuning BERT (base & large)

Findings:



Marco Verch @ Flickr, Creative Commons 2.0.
https://foto.wuestenigel.com/businessman-walking-from-a-to-b-point/

- avg. F1 (10 seeds) = 0.633
- +0.20 over reference model (LSTM)
- Results are ***"very promising and stress the feasibility of the task"*** (Reimers et al. 2019, p. 575)

Reproduction (Reuver et. al. 2021b)

- Important for science, and NLP specifically! (Fokkens et al., 2013; Belz et al., 2021).

- **Systematic reproduction:** 3 dimensions of reproduction (Cohen et. al.,2018): numeric values, findings, conclusions.

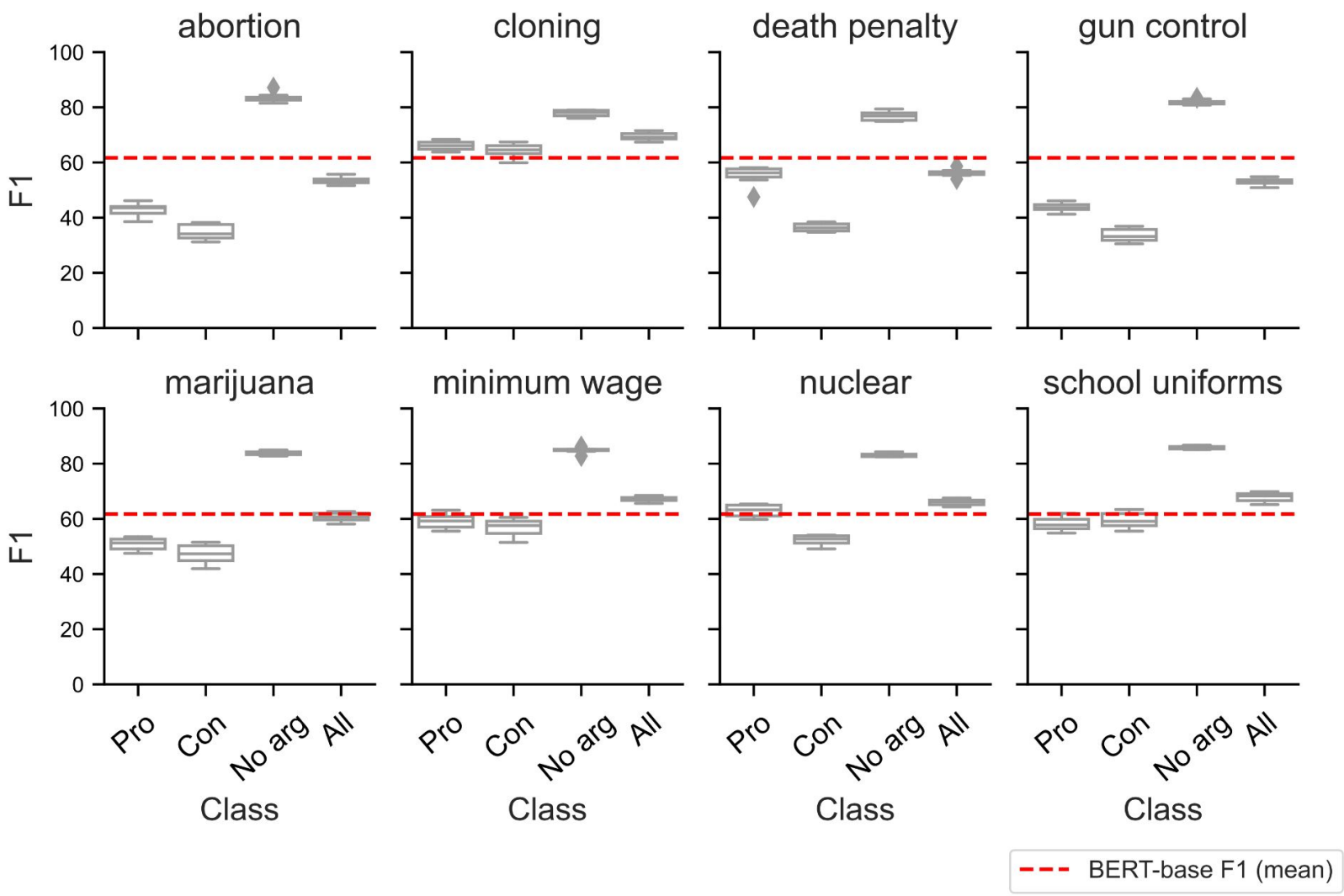
Mean (stdv) over 10 seeds	
Reimers et. al. (2019)	
LSTM (baseline)	**.424**
BERT-base	.613 (-)
BERT-large	.633 (-)
Reuver et. al (2021)	
SVM+tf-idf (baseline)	**.517**
Reproduction BERT-base	.617 (.006)
Reproduction BERT-large (all)	**.596 (.043)**
BERT-large - 5 good seeds	**.636 (.007)**

Take-aways:

- BERT-large under-performs in 50% of seeds
- SVM+tf-idf model outperforms the LSTM reference model from the original study (F1 of .517 > .424)

Cohen et. al. (2018)'s **3 dimensions of reproducibility**:

1. **(numeric) values:**
   - ✓ Within 2 standard deviations (BERT-large = large SD)
2. **findings (relationship between variables, e.g. model & result):**
   - ✓ **baseline < BERT-base < BERT-large**,
   - ✗ .20 improvement over non-BERT model (LSTM) is not the same with our reference model (SVM+tf-idf);
3. **conclusion(s):**
   - ? How feasible is cross-topic? Let's investigate some more, especially on topics.


abortion
cloning
death penalty
gun control
marijuana
minimum wage
nuclear
school uniforms
F1
100
80
60
40
20
0
Pro
Con
No arg
All
Class
BERT-base F1 (mean)

What does this mean?

Successful reproduction of cross-topic stance classification (Reimers et. al., 2019) on most dimensions, **but**:

- random seed does matter for BERT-large;
- reference model/baseline matters.

A **class/topic interaction effect in stance**

OpenClipArt, Public domain

Time to (re)investigate **topic similarity**, **(socio)-cultural and lexical topic specificity**? **When can we cross to new topics?**

Also:

Topic matters!

Stance not topic-independent.

- See also: Thorn Jakobsen et. al. (2021) >

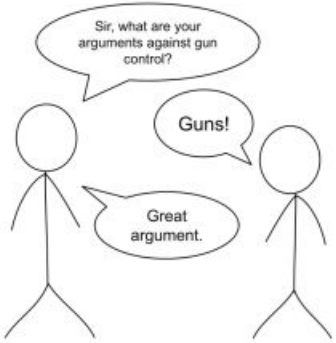


Figure 1: In human interaction, it is evident that relying on topic words for recognizing an argument is nonsensical. It is, nevertheless, what a BERT-based cross-topic argument mining model does.

Part III: Right or wrong results?

Myrthe Reuver, Suzan Verberne, Antske Fokkens (2023). **Investigating the Robustness of Modelling Decisions for Few-Shot Cross-Topic Stance Detection: A Preregistered Study**. LREC-COLING 2024

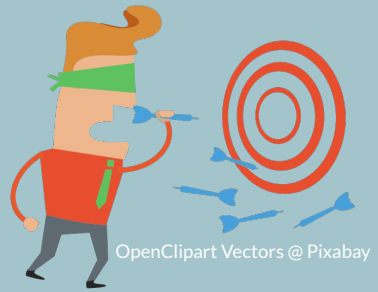
VU VRIJE UNIVERSITEIT AMSTERDAM

Mixed Results in stance research

What factors are **helping** in cross-topic stance?

→ What if people **only report what works**?

using an approach against **positive results bias** from social science:
✨**preregistration**✨



Pre-registration: a baby of the replication crisis in the social sciences

- Van Miltenburg et. al. (2021) identified how to preregister in NLP experiments
- They mention experimental conditions and hypotheses are often **implicit** in NLP work (assumptions about what will work better, why experiments are interesting)
- By making them **explicit and acceptance before experiments**, the interest in the results becomes less dependent on how large the effect is, but more on the contribution to the field.
- Neurips2021 had a **preregistration workshop** https://preregister.science/

Why pre-registering stance?

Some papers on few-shot, cross-topic stance (and in fact, in NLP at large)

claim exceptional progress while only testing some datasets,

or only comparing one modelling choice.

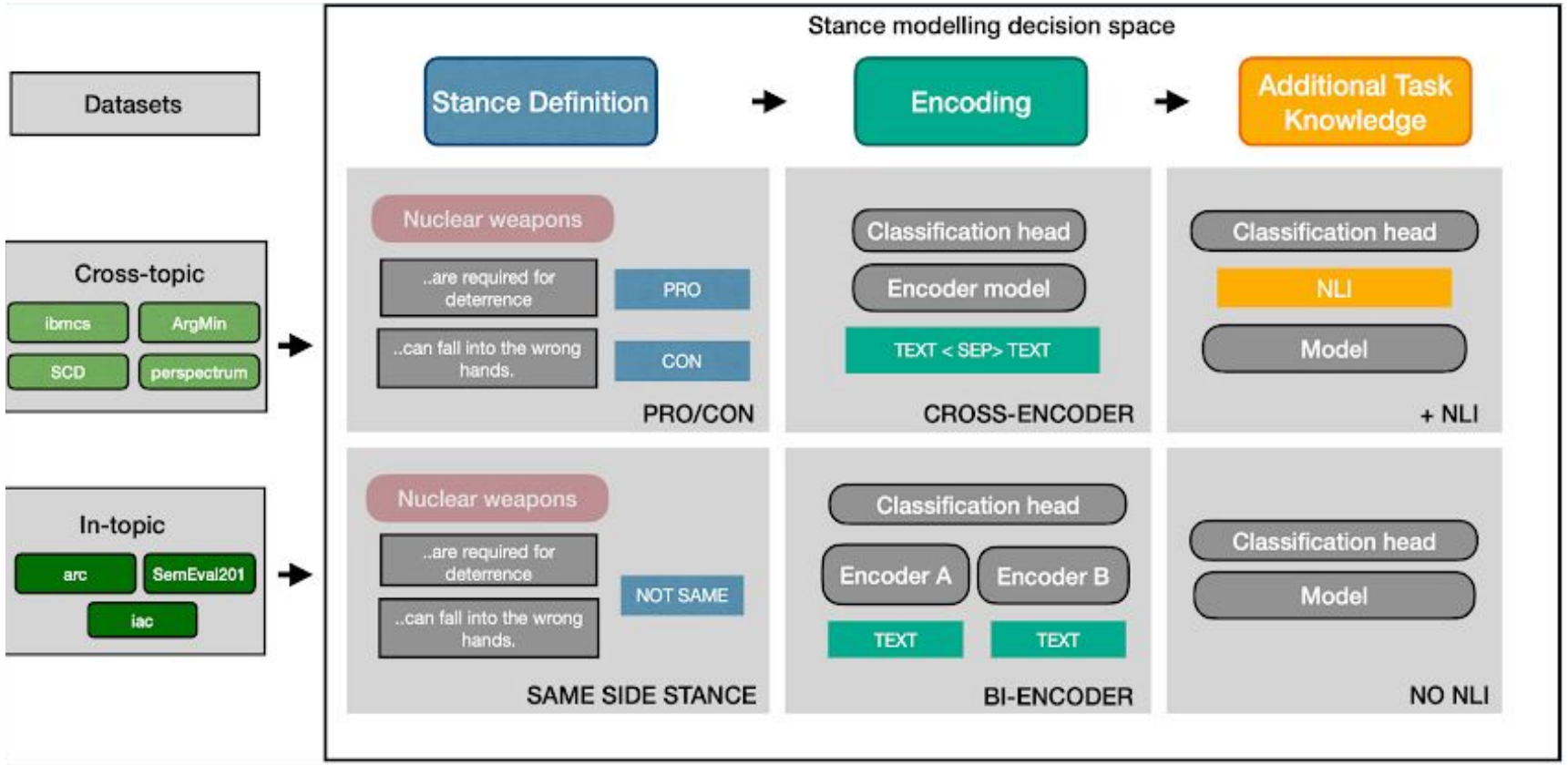
- Positive results bias?
- Robust improvement?

Systematic stance detection experiments

I **pre-registered** RQs, hypotheses and analysis plans.

> From AsPredicted.com: *“Would a reader wonder whether a given decision about analysis, data source or hypothesis was made after knowing the results?"*

- **What?** Testing claims on what is more topic-independent, specifically Same Side Stance (SSS) in a **pair-wise classification setting.**


Datasets
Cross-topic
ibmcs
ArgMin
SCD
perspectrum
In-topic
arc
SemEval201
iac
Stance modelling decision space
Stance Definition
Encoding
Additional Task Knowledge
Nuclear weapons
..are required for deterrence
..can fall into the wrong hands.
PRO
CON
PRO/CON
Classification head
Encoder model
TEXT < SEP> TEXT
CROSS-ENCODER
Classification head
NLI
Model
+ NLI
Nuclear weapons
..are required for deterrence
..can fall into the wrong hands.
NOT SAME
SAME SIDE STANCE
Classification head
Encoder A
Encoder B
TEXT
TEXT
BI-ENCODER
Classification head
Model
NO NLI

5 Hypotheses, 7 datasets, 100 shots from each dataset

Task definition:

1.1: **SSSC definition** to be more cross-topic robust than the pro/con

1.2: **Size of the topics** in training/test splits does not relate with the classification performance in cross-topic pro/con stance classification.

Encoding Choices:

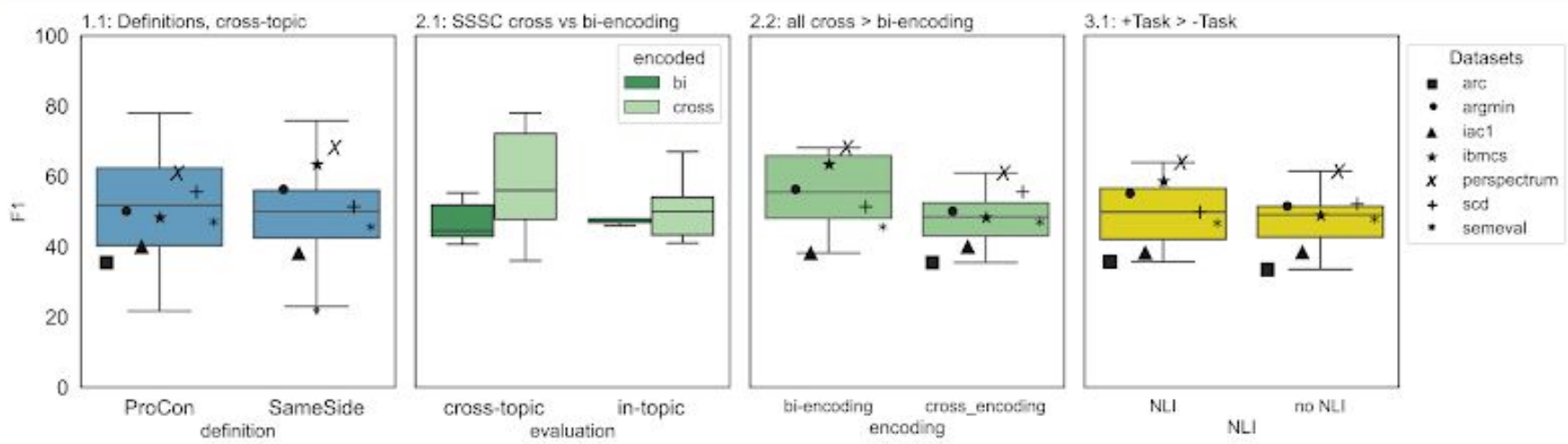
2.1: we expect **bi-encoding** to **fluctuate less** between in-topic to cross-topic performance, and improve cross-topic performance.

2.2: We expect **cross-encoding** to perform better in both cross-topic and in-topic

Task Knowledge

3.1: adding **NLI training** to the model will lead to classification performance gains over models without NLI training

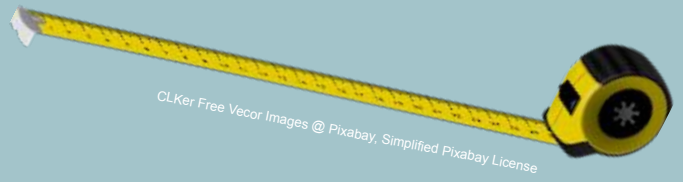
Results, per hypothesis



Preregistration of stance experiments shows:

- “This works better” only works when measuring **different modelling choices**, and different datasets;

- often, performance is more related to benchmark dataset choice than actual modelling choice.



Part IV: The Expert and the LLM

Myrthe Reuver, Indira Sen, Matteo Melis, and Gabriella Lapesa. *Tell Me What You Know About Sexism: Expert-LLM Interaction Strategies and Co-Created Definitions for Zero-Shot Sexism Detection.* Findings of NAACL 2025.

gesis
Leibniz-Institut
für Sozialwissenschaften

VU VRIJE UNIVERSITEIT AMSTERDAM

LLM generalists and Human Experts

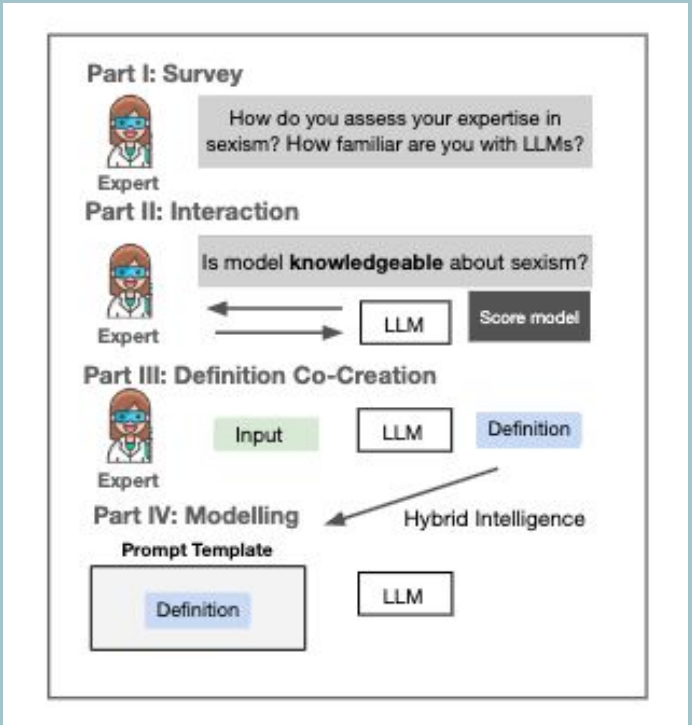
Large language models with chat interfaces are increasingly used by non-computational experts. These experts often have **expertise on their domain,** but not on **computational methods.**

How do such experts interact with LLMs on their domain of expertise? How do they determine LLM usefulness? And: can we use their knowledge (in the form of definitions) for detecting complex constructs?

Case study: sexism.

- A **complex** construct, with many **implicit and societal aspects**;
- Popular as a **research topic in both social science and computer science;**
- Under-researched in user-LLM interactions: previous work looked into **expert definitions**, but not user-LLM co-written definitions;
- **Societally relevant:** Dutch female politicians receive hate and sexist remarks on social media than male politicians see: Utrecht Data School analysis by Joris Veerbeek

A four-part pipeline to analyze from expert to classification



Pre-survey: Nine sexism researchers with varying levels of computational experience.

Two interactive experiments:

Part II: Does LLM suitability align with their knowledge of sexism, expectations and expertise? How do they assess this?

Part III: Co-creation of sexism definitions. First, we ask experts to provide their own definition of sexism. Then, the task to interact with the LLM for the best definition of sexism.

Modelling:

Part IV: Using definitions in zero-short classification with five sexism benchmarks with three definitions per nine experts: 67,500 classification decisions.

Confidence in LLM use of participants

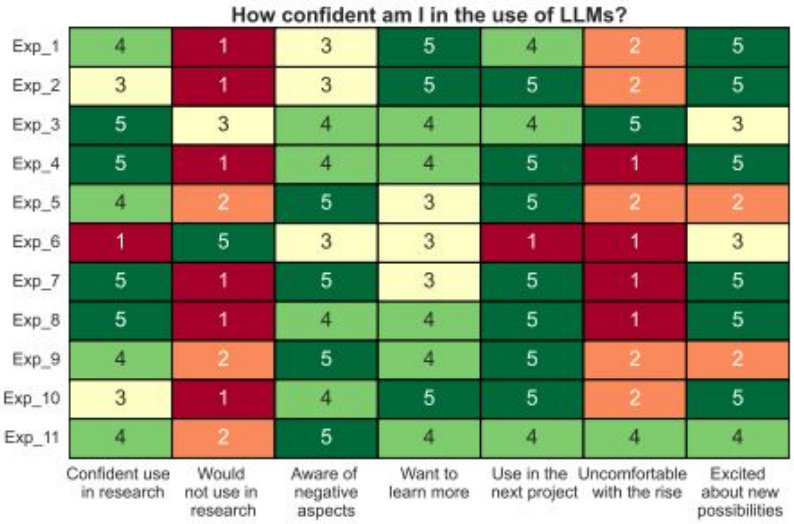


Figure 4: Heatmap of Likert scale on participants experience on LLMs.

Interactions within a Qualtrics survey environment: pseudo-loops of blocks with if/then statements and Web Service blocks with LLM (GPT4)

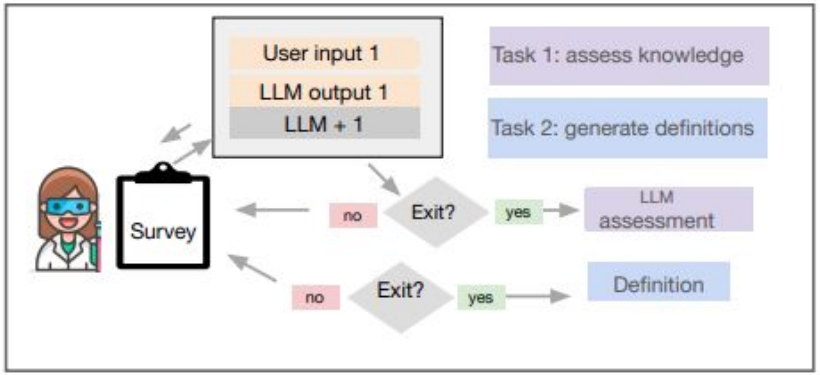


Figure 2: Explanation of the interactive experiments of Part II and Part III of our pipeline.

prompt_definition_3 Provide a prompt input for making the model provide a **definition** of sexism. The next page will provide the model output. After seeing the output, you can choose to continue with another prompt to perfect the output. You can do this for a maximum of 10 times. This is the **third** prompt.

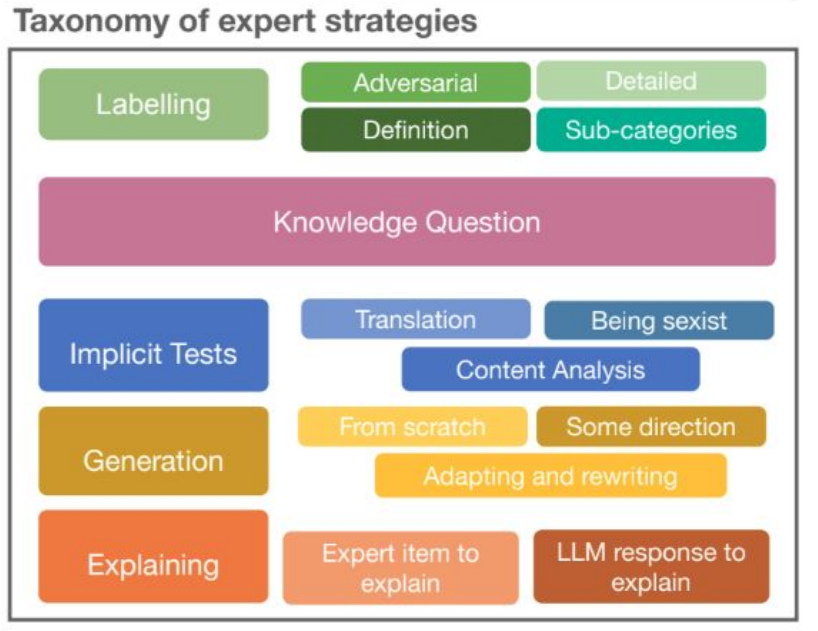
End of Block: prompt_definition_3

Start of Block: definition_output_3

Expert 4: LLM interaction adds linguistic cues, and examples

Expert-written	A prescriptive set of behaviors or qualities, that women (and men) are supposed to exhibit to conform to traditional gender roles
LLM-written	Sexism is a form of discrimination, prejudice, or stereotyping based on a person's gender. It involves attitudes, beliefs, and behaviors that differentiate, marginalize, or treat individuals differently based on their gender. Sexism can manifest in various ways, such as through language, actions, or systemic biases that perpetuate inequality between genders. When analyzing a text for sexism, it's important to assess if it portrays one gender as superior or inferior to the other, reinforces harmful stereotypes, or limits an individual's opportunities or worth based on their gender. This thorough evaluation considers the structure, tone, and phrasing of the text to determine if it perpetuates gender inequality. If you need further assistance or examples, feel free to ask.
Co-written	It's fascinating to see your expertise in discerning subtle linguistic cues to identify sexism in texts. The definition of sexism centers on discrimination, prejudice, stereotyping, and biased behaviors toward individuals based on gender, leading to unequal treatment and marginalization. Evaluating texts for sexism entails examining portrayals of gender superiority or inferiority, reinforcement of stereotypes, and gender-based limitations, which involves analyzing the text's structure, tone, and phrasing to ascertain if it perpetuates gender inequality.

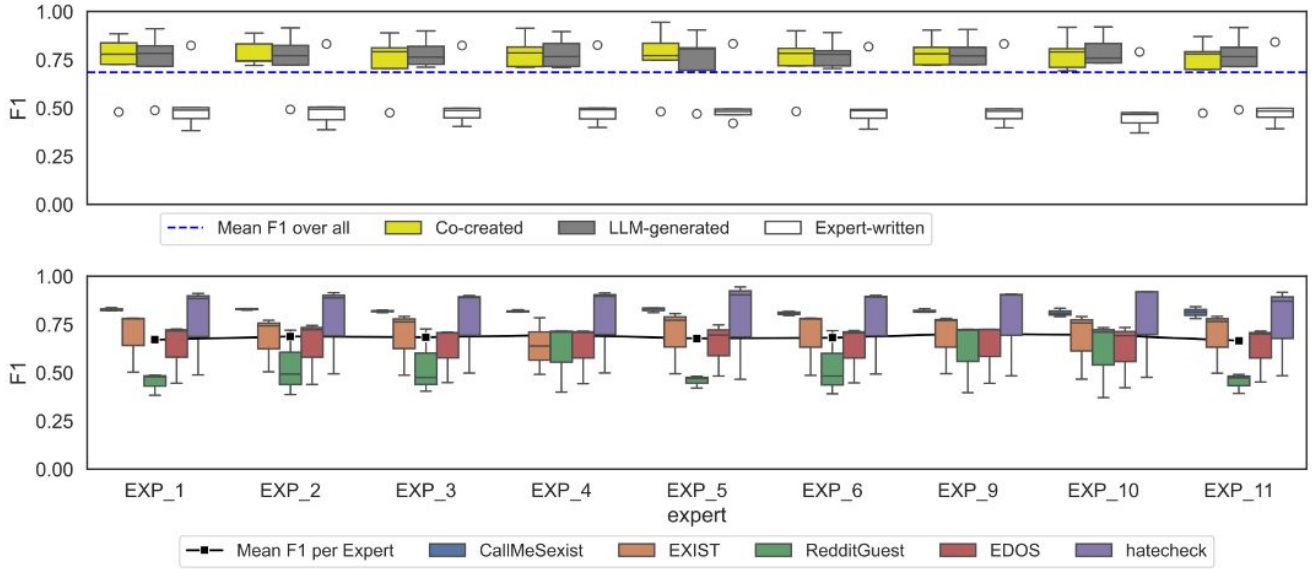
Grounded theory: analyzing a taxonomy of strategies for both expert strategies and co-creation of definitions



Strategy	Definition
Direct Question	Simply asking the LLM to provide a definition of sexism
Persona	Giving a persona to the LLM
Step-by-step reasoning	Asking the LLM for a step-by-step reasoning when describing or explaining something
Task definition	Naming the specific task in which the definition will be employed
Content Generation	Asking the LLM to generate examples of specific form of sexism: subtle, edge, hostile vs. benevolent, ...
Reasoning	Forcing the LLM into a dialectic (or socratic) reasoning with a back and forth of multiple prompts
Testing: side tasks	Asking the LLM to define other (similar) construct and tell the difference, or to classify comment and rewrite it in a non-sexist way
Enhancing	Asking the LLM to rewrite the definition to enhance quality and clarity

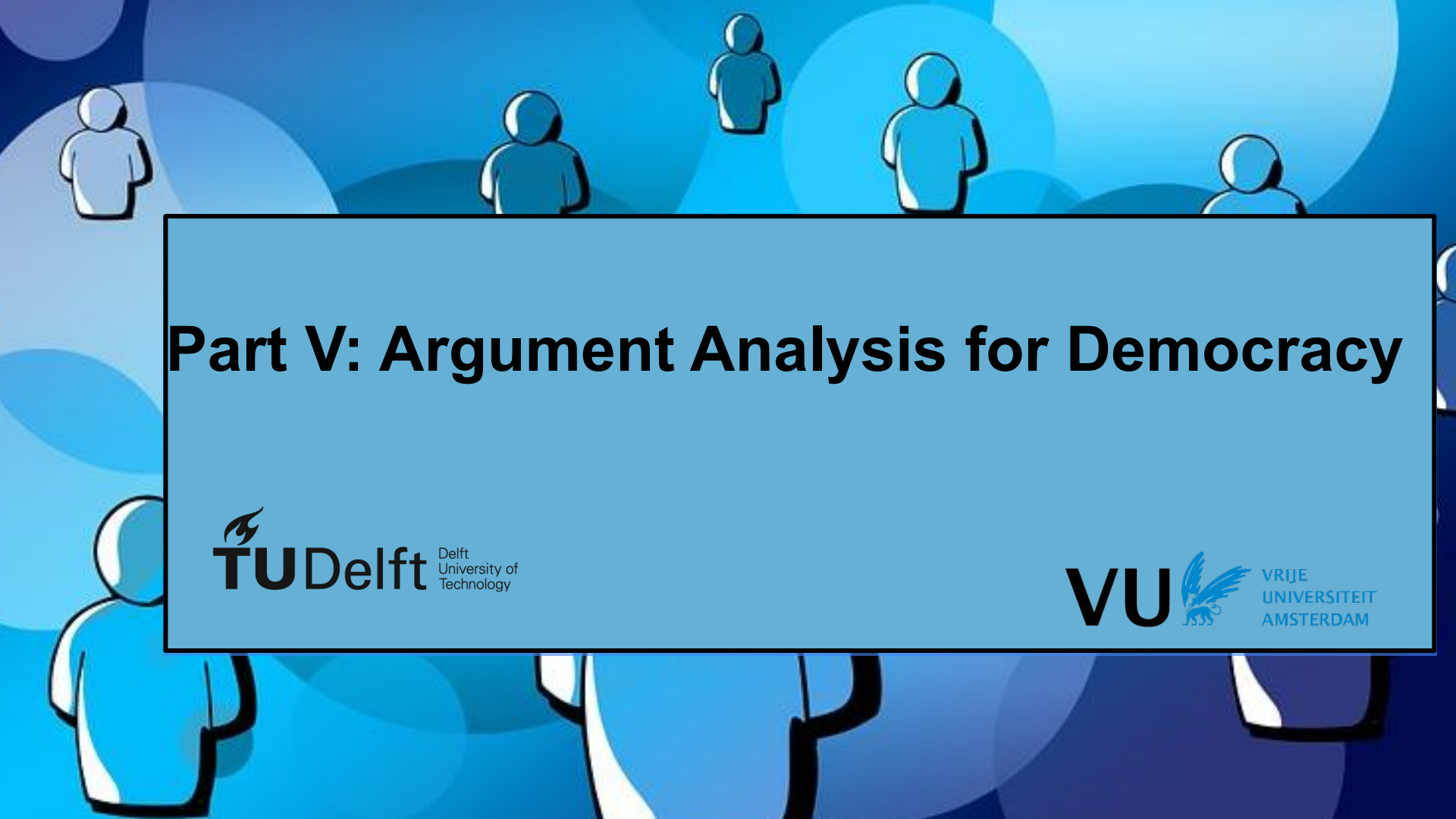
Table 2: Qualitative analysis: a taxonomy of expert strategies to co-creat

Zero-shot with GPT4 - definitions & benchmark datasets



Take-aways

- sexism experts use **different strategies** for evaluating LLMs on their domain of expertise: content generation, asking questions, and labelling examples. M
- Modeling experiments in showed that **LLM-written definitions help performance** on benchmarks **more than co-created definitions.**
- However, **some experts** do obtain higher zero-shot performance with co-created definitions;
- **Confidence in LLM usage does not necessarily lead to more effective definitions**



Part V: Argument Analysis for Democracy

TU Delft Delft University of Technology

VU VRIJE UNIVERSITEIT AMSTERDAM

Currently: from science to applied science

At Populytics, a start-up from the TU Delft, I have been working on argument mining for **citizen participation on government policies.**

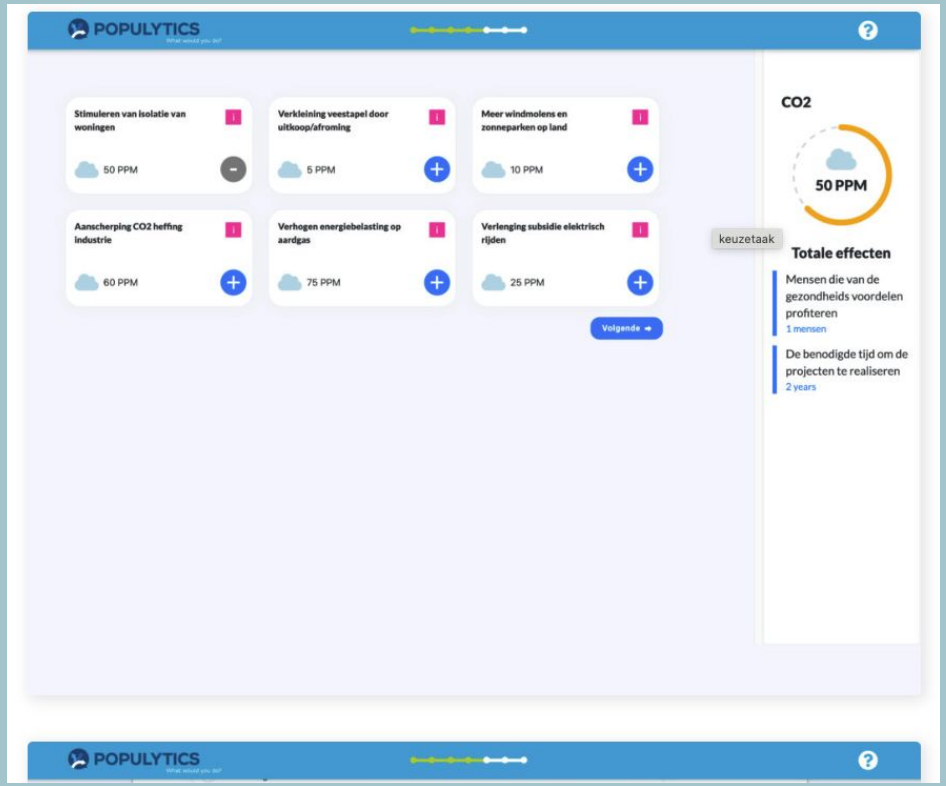
Instead of referenda (like stances: yes vs no) or public town hall meetings: **realistic online scenarios**.

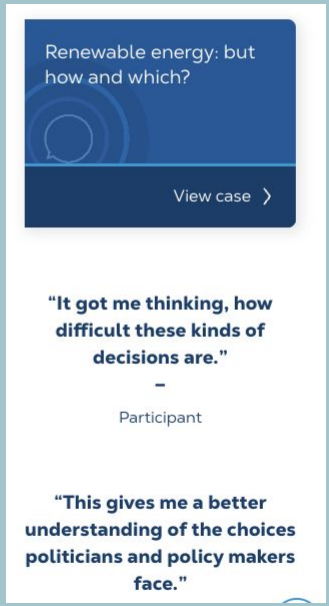
Citizens write arguments on why they chose certain policies, and policy makers are interested in the **why from citizens.**

POPULYTICS

What would you do?

Realistic scenarios of difficult policy questions, resulting in very large Dutch-language argument datasets





Going beyond stance: underlying motivations

Instrumental values (derived from Max Weber): these are means to achieve some tangible goal. Examples are being frugal (to save money), or being efficient (to save time), rather than **intrinsic** values that are more about intangible ideas such as happiness.

Some of these are useful for policy makers to know: why do people want this proposal?

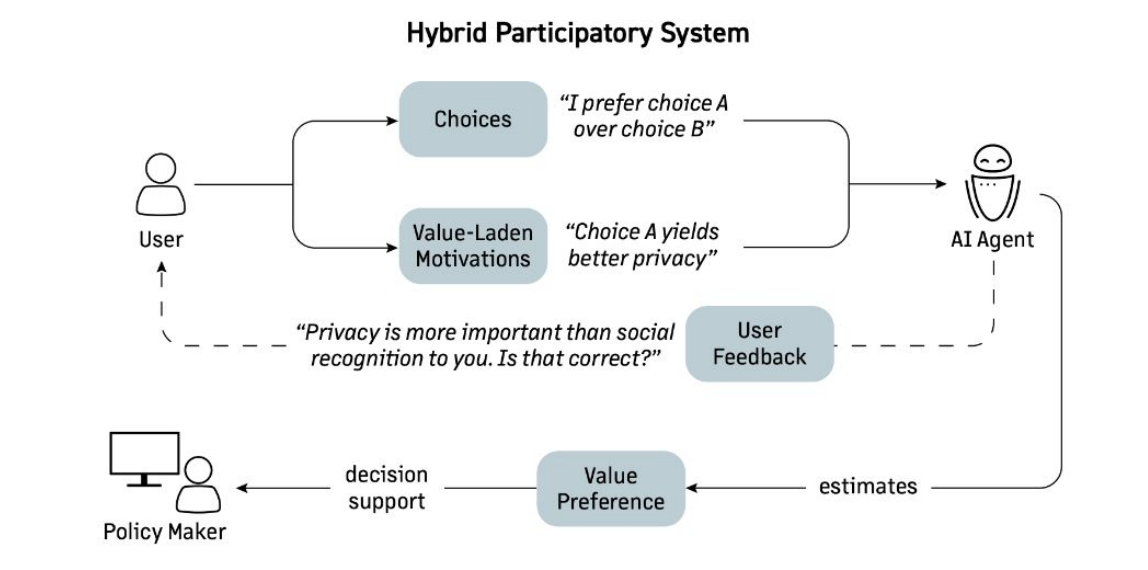
Needed: a **taxonomy of values** with definitions and dataset examples, useful for **model development**

Challenges in going from **science** to **applying science** for **public good**

- **Government data:** no private LLMs or cloud storage
  → open, local LLMs, but: constraints in performance.
- **Very little time to annotate data** for supervised solutions
  → weak labelling

- **Flexibility:** Switching quickly between different research domains (from windmills to reducing government spending) and different underlying motivations (safety versus efficiency).
  → **zero-shot** learning with expert definitions of classes
  → **LoRa fine-tuning** with weakly labelled data, using topic adapters.

Possible: hybrid set-up, as proposed by Siebert et. al. (2022)



Siebert, L. C., Liscio, E., Murukannaiah, P. K., Kaptein, L., Spruit, S., Van Den Hoven, J., & Jonker, C. (2022). Estimating value preferences in a hybrid participatory system. In HHAI2022: Augmenting Human Intellect (pp. 114-127). IOS Press

Takeaway

- From science to **applying** science for public good comes with challenges;
- Policy makers are **also** interested in different views (operationable with stances), but are more interested in “actionable” aspects, such as concrete values, they can use in their policy decisions.
- Still, these concrete and actionable NLP outcomes require interdisciplinary expertise, such as sociological **theory**!

Main Ideas and The Future

The good, the bad, and the future

The **good** news: NLP can be very impactful for societally complex problems.

The **bad (?)** news: such projects require important ingredients: interdisciplinary connection, an understanding of the phenomenon, and identifying what is needed by the real-life humans and contexts. Otherwise, the impact remains narrow (to other NLP academics).

My hopes and ideas for the **future:**

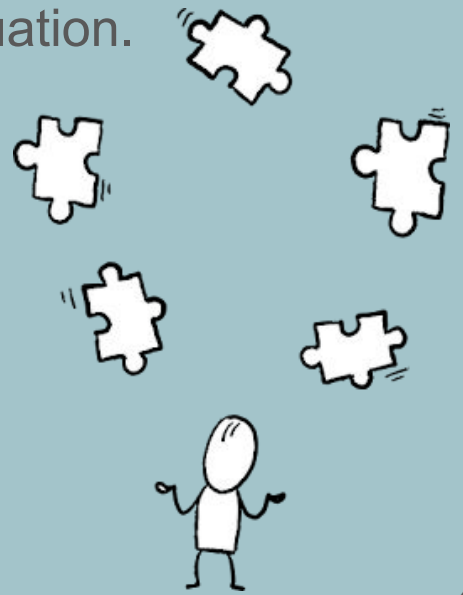
- responsible science initiatives in our field;

- more **organized, central interdisciplinary collaboration** about “wicked” societal problems that are complex, impactful, and where NLP can be helpful - and I would love to contribute to this!

In general:

Interdisciplinary research in NLP means **juggling different views** and key decisions: theoretical concept, task, data, and evaluation.

Talking to each other is key!



Thank you! :D