Argument Quality Assessment in the Age of Instruction-Following Large Language Models

Henning Wachsmuth, Gabriella Lapesa, Elena Cabrio, Anne Lauscher, Joonsuk Park, Eva Maria Vecchi, Serena Villata, Timon Ziegenbein

h.wachsmuth@ai.uni-hannover.de

Natural Language Processing

Leibniz Universität Hannover

Argument quality assessment

- **Argument**
  - A claim on a controversial issue supported by reasons

  > " *If you wanna hear my view, I think that the EU should allow rescue boats in the* Claim
  > *Mediterranean Sea. Many innocent refugees will die if there are no such boats.*
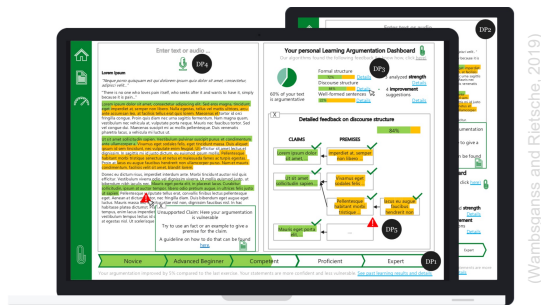 > *Nothing justifies to endanger the life of innocent people.* " Reasons

  sufficient? **relevant?** **effective?**
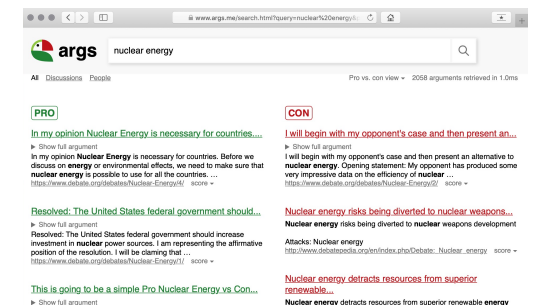
- **Argument quality assessment**
  - Rating an argument or ranking different arguments
  - Critical for various applications
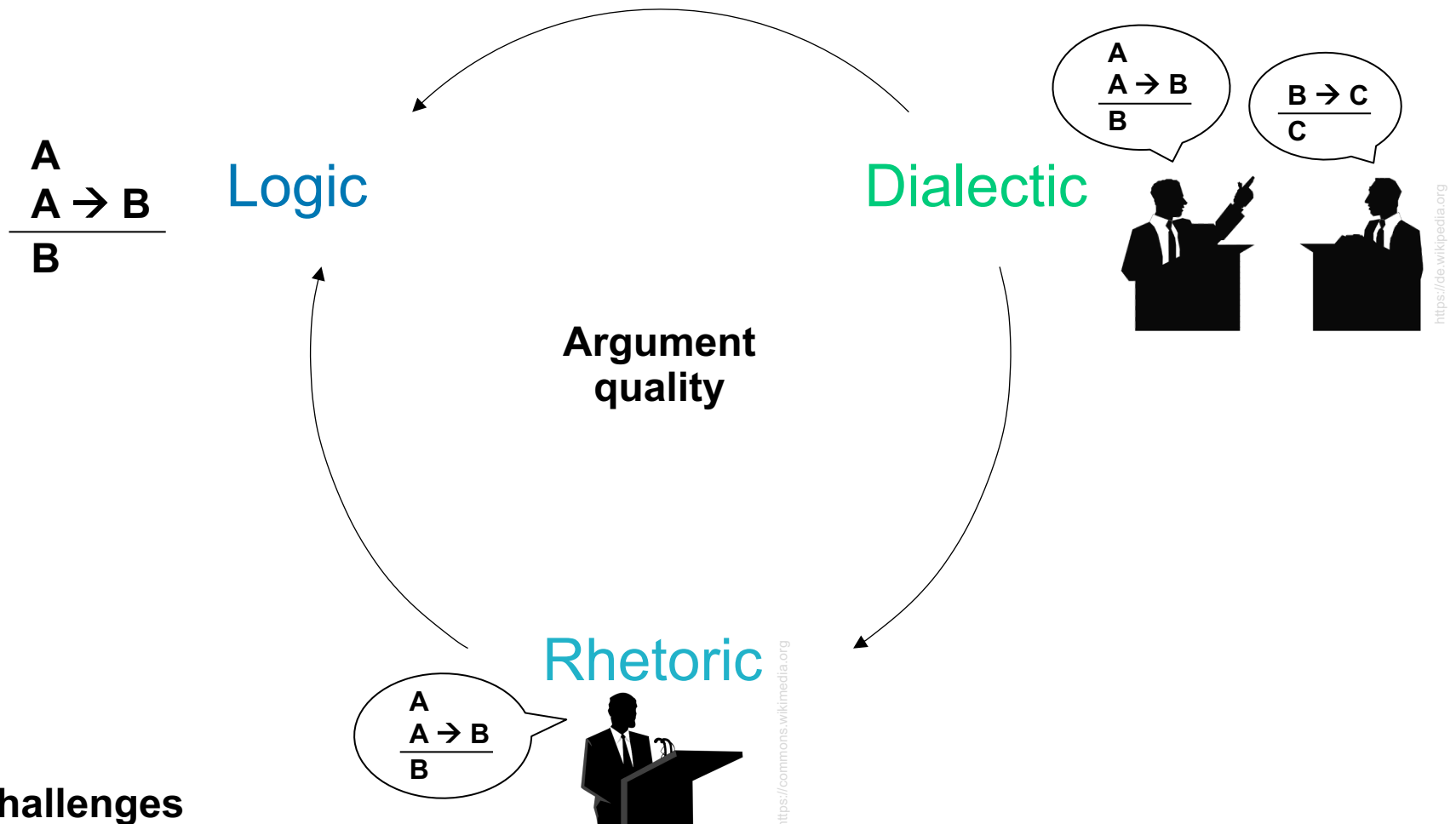

Writing support

(Wambsganss and Rietsche, 2019)


Debating technology

https://flickr.com

Argument search

Challenges of argument quality (Wachsmuth et al., 2017)

Logic

$$\frac{\begin{array}{c} A \\ A \rightarrow B \end{array}}{B}$$

Dialectic

$$\frac{\begin{array}{c} A \\ A \rightarrow B \end{array}}{B} \qquad \frac{B \rightarrow C}{C}$$

Argument quality

Rhetoric

$$\frac{\begin{array}{c} A \\ A \rightarrow B \end{array}}{B}$$

- **Challenges**
  - Diversity of quality. Quality notions vary strongly across contexts
  - Subjectivity of perception. Many quality notions inherently subjective

This position paper
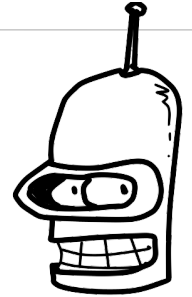
- **Instruction-following large language models (LLMs)**
  - Can leverage knowledge across contexts and tasks
  - Can tackle tasks with little to no fine-tuning

- **Position**
  - Challenges of argument quality still prevail
  - LLMs allow for substantial progress
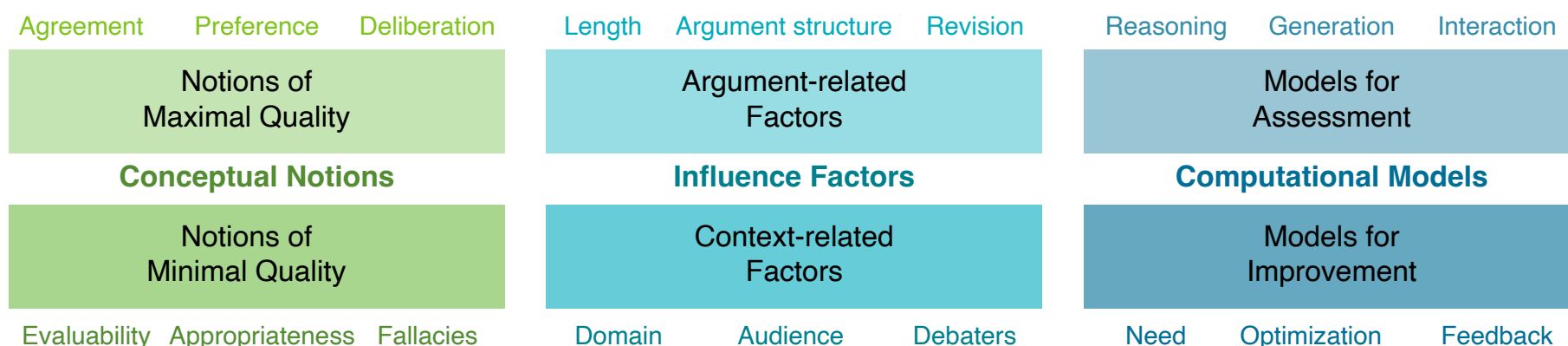 - Need to teach LLMs argument-specific knowledge

  > *" How to drive research on LLM-based argument quality assessment in order to face the prevailing challenges of diverse quality notions and their subjectivity? "*

- **Contributions**
  - Survey of recent research on argument quality
  - Blueprint for argument-specific instruction fine-tuning
  - Discussion of opportunities, limitations, and ethical concerns

Survey of argument quality research

- **Survey of 83 papers**

  - Argument quality in NLP, AI, IR, and CA from 2017 to 2023
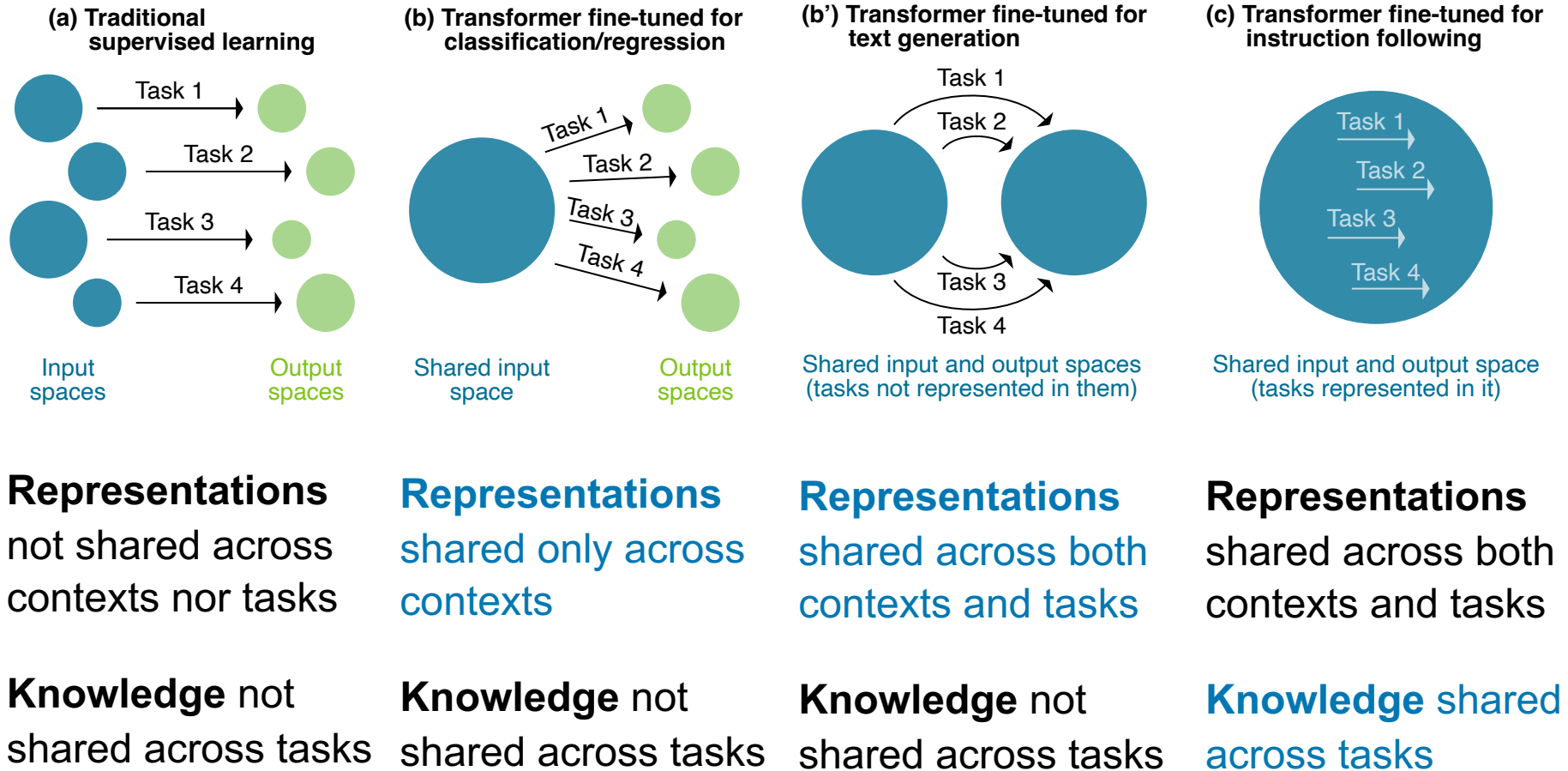 - Organization by primary research directions

Agreement	Preference	Deliberation	Length	Argument structure	Revision	Reasoning	Generation	Interaction
Notions of Maximal Quality			Argument-related Factors			Models for Assessment		
Conceptual Notions			**Influence Factors**			**Computational Models**		
Notions of Minimal Quality			Context-related Factors			Models for Improvement		
Evaluability	Appropriateness	Fallacies	Domain	Audience	Debaters	Need	Optimization	Feedback

- **Primary research directions**

  - Conceptual notions (24 papers). What is quality and how to measure it
  - Influence factors (30 papers). What affects quality perception beyond the text
  - Computational models (21 papers). How to assess and improve quality
    Remaining 8 papers: Applications, data handling, and similar

LLMs for argument quality assessment

(a) Traditional supervised learning

Task 1
Task 2
Task 3
Task 4

Input spaces   Output spaces

(b) Transformer fine-tuned for classification/regression

Task 1
Task 2
Task 3
Task 4

Shared input space   Output spaces

(b') Transformer fine-tuned for text generation

Task 1
Task 2

Task 3
Task 4

Shared input and output spaces (tasks not represented in them)

(c) Transformer fine-tuned for instruction following

Task 1
Task 2
Task 3
Task 4

Shared input and output space (tasks represented in it)

Representations not shared across contexts nor tasks

Knowledge not shared across tasks

Representations shared only across contexts

Knowledge not shared across tasks

Representations shared across both contexts and tasks

Knowledge not shared across tasks

Representations shared across both contexts and tasks

Knowledge shared across tasks

A blueprint for argument-specific instruction fine-tuning

- **Idea**
  - Systematically teach LLMs to follow argument-specific instructions
 - Enable LLMs to handle quality notions, audiences, and more

> " Rate the quality from the perspective of **rhetorical effectiveness,** if presented to a **highly rational person** "

- **Blueprint**
  1. Start from instruction-following LLM. Pretrained transformer suffices, if general instructions added to next step

  2. Acquire a set of argumentation-specific instructions. Derive them from existing guidelines and datasets

  3. Apply instruction fine-tuning on acquired set. Use RLHF, self-instruct, or similar depending on resources

  4. Apply LLM alignment on unseen tasks. Use soft prompting, sociodemographic prompting, or similar
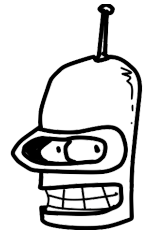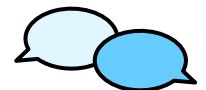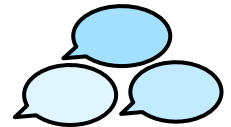
# Opportunities, limitations, and ethical concerns

- **Opportunities**
  - General impact of LLMs on argument quality assessment
  - Breakthrough in applications such as writing support
  - New paths in research by handling diversity and subjectivity

- **Limitations**
  - Impact of blueprint remains to be evaluated
  - Systematic ways to do instruction fine-tuning to be found
 - Availability of data for reliable assessment not ensured
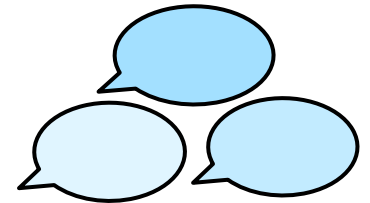
- **Ethical concerns**
  - Privacy. Audience information often key, but also sensitive
  - Hallucinations. Factual errors may mislead beliefs
  - Biases. Unclear who should decide about argument quality

Takeaways

- **Survey of argument quality research**
  - 83 NLP, AI, IR, and CA papers
  - Organized by research directions
  - Diversity and subjectivity challenging

- **Blueprint for instruction fine-tuning**
  - Impact of instruction-following LLMs
  - 4 steps to argument-specific tuning
  - Implications for quality evaluation

- **And much more…**
  - Check the paper for all details
  - Browse the literature list provided
  - Meet us at our poster in Torino ;)
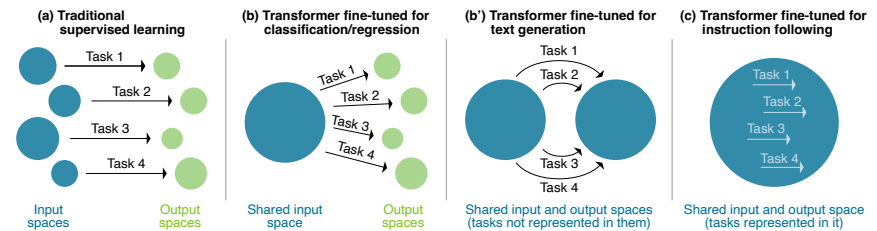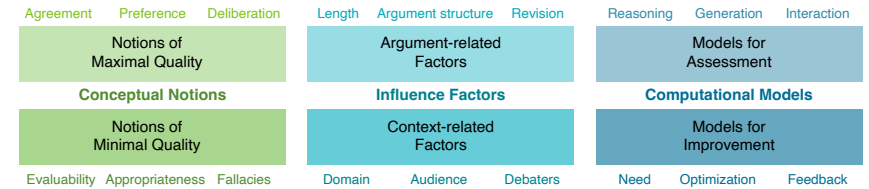
References

- **Wachsmuth et al. (2017).** Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. Computational Argumentation Quality Assessment in Natural Language. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pages 176–187, 2017.

- **Wambsganss and Rietsche (2019).** Thiemo Wambsganss and Roman Rietsche. Towards Designing an Adaptive Argumentation Learning Tool. In Proceedings of the Fortieth International Conference on Information Systems, 2019.