Paderborn Colloquium on Data Science and Artificial Intelligence at School

NLP Research in the Age of Large Language Models

Henning Wachsmuth

h.wachsmuth@ai.uni-hannover.de

November 28, 2023

Natural Language Processing

Leibniz Universität Hannover

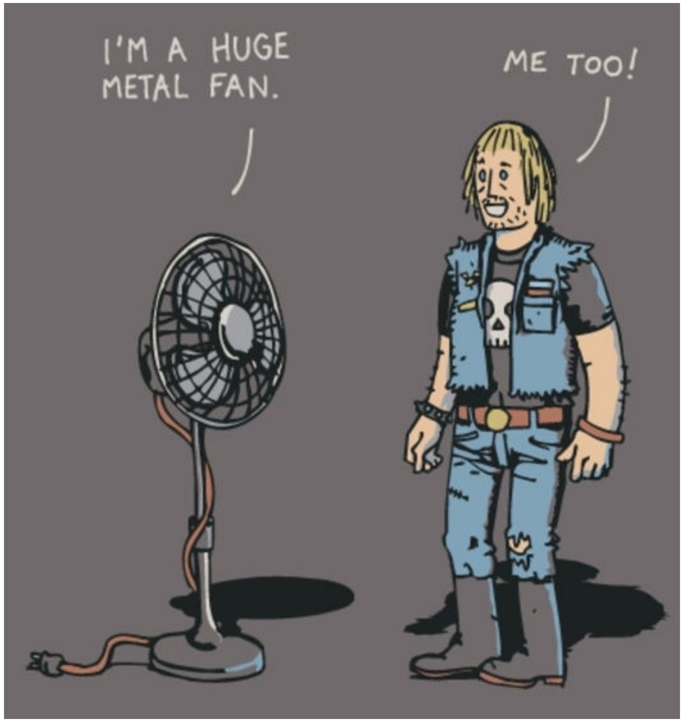
NLP Research

- **Natural language processing (NLP)**
  - Subfield of AI dealing with natural language
  - Methods for understanding and generating text (or speech)
  - Applications in data science and human-AI interaction

What ChatGPT does

- **Challenges of NLP**
  - Language is intrinsically ambiguous
  - Syntax, semantics, and pragmatics interact
  - Context and world knowledge needed

- **Key research method (so far)**
  - Given training and test data for a task
  - Develop method on training data
  - Evaluate method on test data



Large language models (LLMs)

https://flickr.com

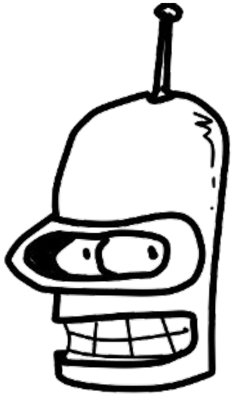
- **ChatGPT**
  - A chatbot system that leads open-topic dialogues with users
  - Uses the *language model* GPT-3.5 (or -4) for text generation
- **Language model (LM)**
  - A probability distribution over word sequences, derived from huge text data
  - Probabilities can be used to generate most likely *next* words

User: Can you explain to me what is meant by "putting your cards on the table"?

ChatGPT: "Putting your cards on the table" is a <?>

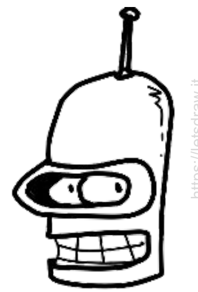
P(phrase | dialogue) = 0.10
P(saying | dialogue) = 0.07
P(typical | dialogue) = 0.05

- **Large language model (LLM)**
  - All existing LLMs based on neural *transformer* networks
  - Not fully defined when large, but usually billions of parameters
  - 1st generation. Transformer-based models (BERT, BART, …)
  - 2nd generation. Instruction-tuned models (GPT-4, Alpaca, …)



https://letsdraw.it

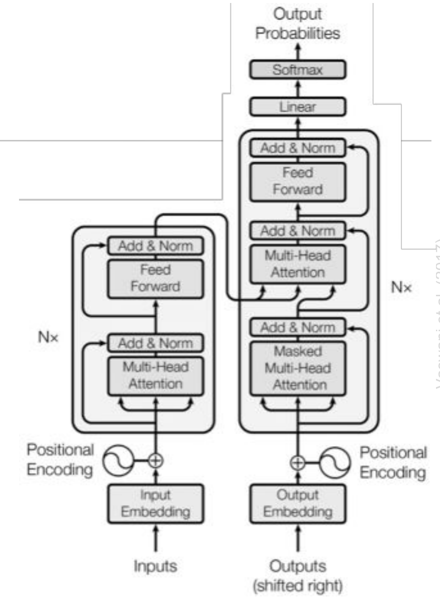
BERT, BART,
or similar



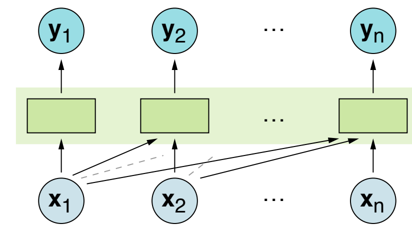
First-generation LLMs

Transformers

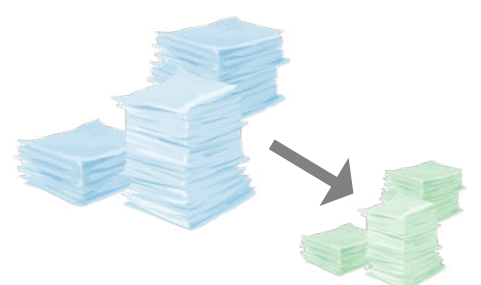
- **Transformer**
  - A neural network architecture for parallel input processing
  - In full form, input encoder + output decoder
  - Inputs and outputs in NLP are sequences of (sub-)tokens
  - Key concepts: Self-attention and transfer learning



- **Self attention**
  - Model each input based on context of surrounding inputs
  - Largely solves modeling of long-term input dependencies
  - Enables full parallelization of input processing

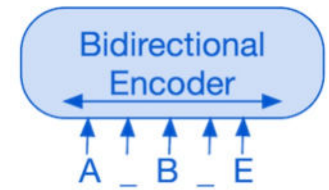


- **Transfer learning**
  - Pretrain model unsupervised on huge language data
  - Fine-tune it supervised on task-specific training data
  - Strongly reduces need for training data
  - Enables leveraging of knowledge across contexts



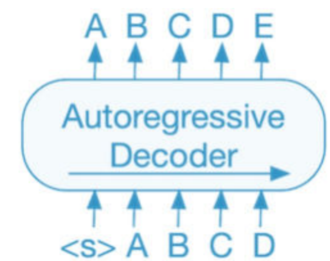
Transformers: Three common variations

- **Bidirectional transformer (encoder-only)**
  - Models inputs based on both previous and following inputs
  - Usually for label and value prediction
  - Examples. BERT and RoBERTA
    (Devlin et al., 2019; Liu et al., 2019)



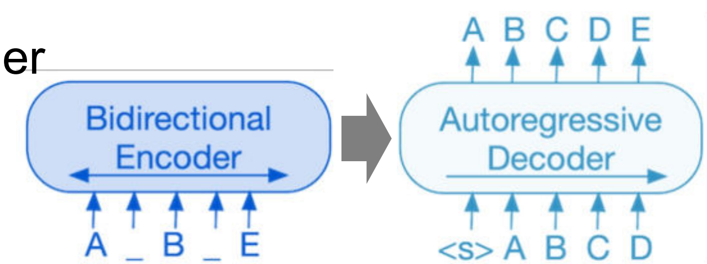
Lewis et al. (2019)

- **Autoregressive transformer (decoder-only)**
  - Models inputs based on previous inputs only
  - Usually for text generation
  - Examples. GPT-x and Alpaca
    (Radford et al., 2018; Taori et al., 2023)



Lewis et al. (2019)

- **Full transformer (encoder-decoder)**
  - Bidirectional encoder, autoregressive decoder
  - Usually for controlled text generation
  - Examples. BART and T5
    (Lewis et al., 2019; Raffel et al., 2020)



Lewis et al. (2019)

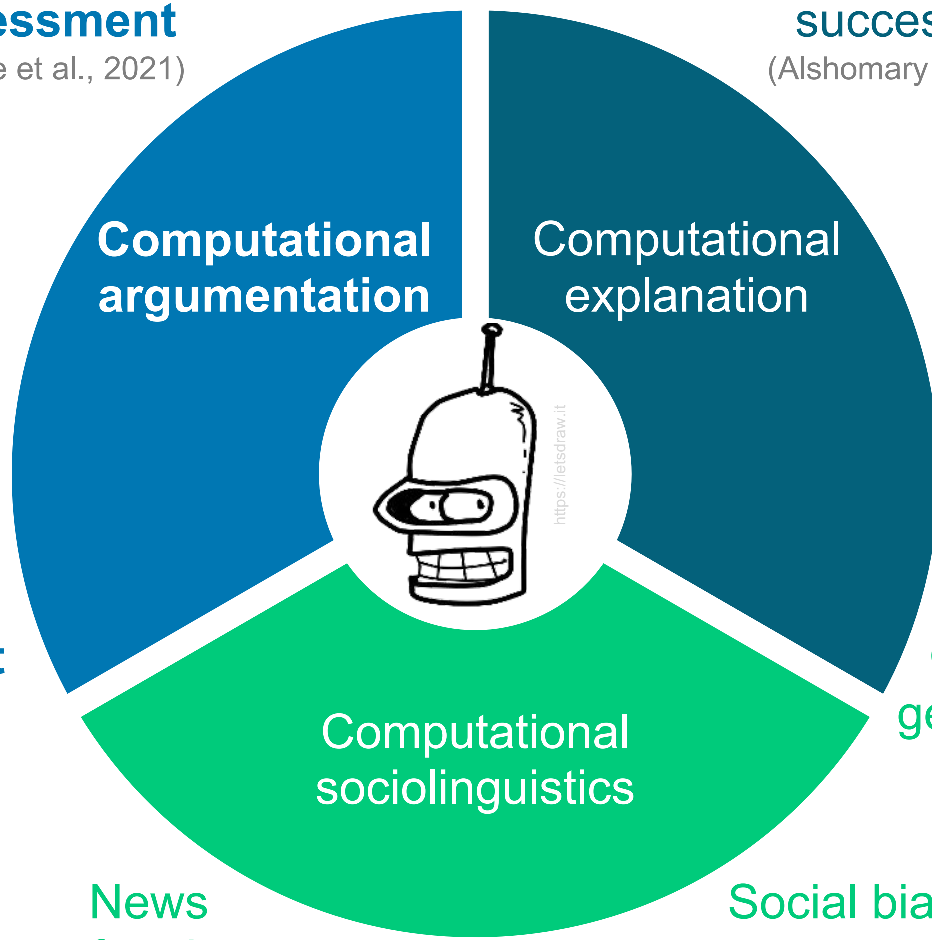
Selected research with LLMs

Argument quality assessment
(Gurcke et al., 2021)

Explanation success prediction
(Alshomary et al., under review)

Claim quality optimization
(Skitalinskaya et al., 2023)

Modeling of metaphor highlighting
(Sengupta et al., 2023)



Counterargument generation
(Alshomary et al., 2023)

Counterfactuals for gender bias mitigation
(Stahl et al., 2022)

News reframing
(Chen et al., 2021)

Social bias evaluation
(Spliethöver et al., 2022)

Timon Ziegenbein (neé Gurcke)

Milad Alshomary

Henning Wachsmuth

LLMs for Argument Sufficiency Assessment

(Gurcke et al., 2021)

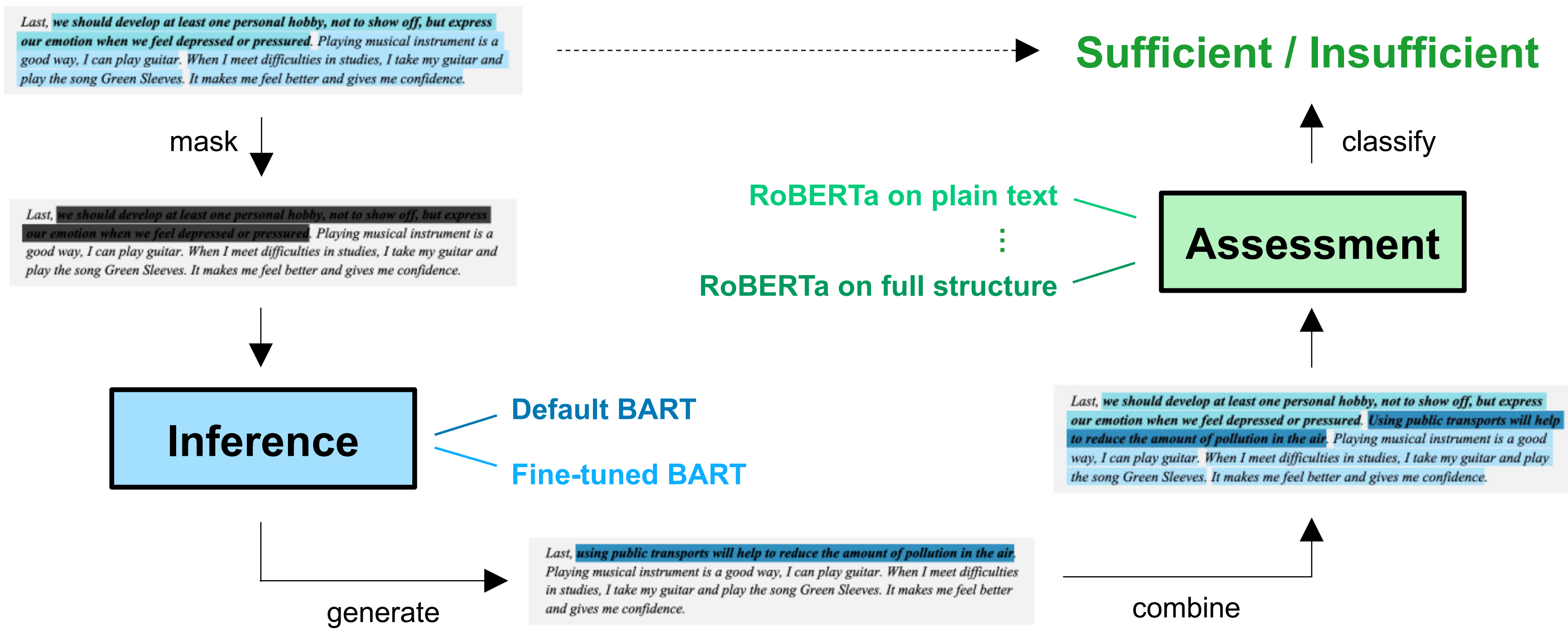
Problem. Do generated conclusions help assess whether an argument is logically sufficient?

Idea. Fine-tune LLM on generating conclusions from premises; use both for assessment

Results. Generated conclusions on par with humans, but impact on assessment low

LLMs for sufficiency assessment: Approach

- **Approach**
  - Inference. Generate a(nother) conclusion from the argument's premises
  - Assessment. Classify sufficiency based on argument and inferred conclusion



LLMs for sufficiency assessment: Examples

- **Insufficient argument**

*Last, ~~**we should develop at least one personal hobby, not to show off, but express our emotion when we feel depressed or pressured**~~. Playing musical instrument is a good way, I can play guitar. When I meet difficulties in studies, I take my guitar and play the song Green Sleeves. It makes me feel better and gives me confidence.*

but not least, I love music

Default BART

playing musical instrument is very important to me

Fine-tuned BART

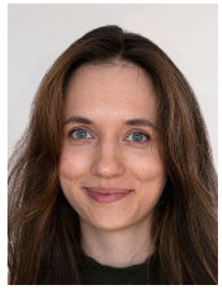
- **Sufficient argument**

*Second, ~~**public transportation helps to solve the air pollution problems**~~. Averagely, public transports use much less gasoline to carry people than private cars. It means that by using public transports, less gas exhaust is pumped to the air and people will no longer have to bear the stuffy situation on the roads, which is always full of fumes.*

public transport is more efficient than private cars

using public transports will help to reduce the amount of pollution in the air

Gabriella Skitalinskaya | Maximilian Spliethöver | Henning Wachsmuth

LLMs for Claim Quality Optimization

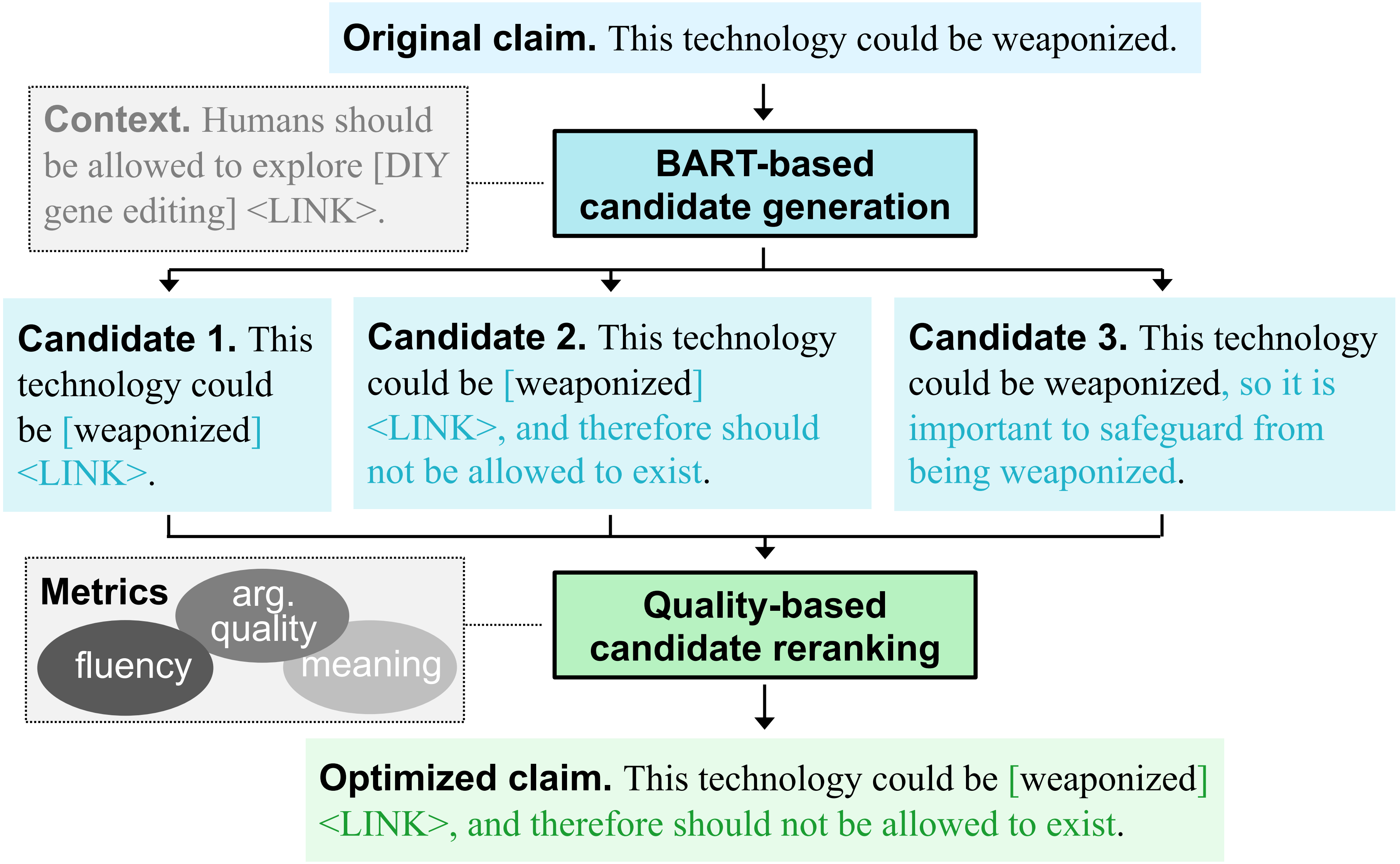
(Skitalinskaya et al., 2023)

Problem. How to improve argumentative claims without changing their meaning?

Idea. Fine-tune LLM on claim revisions; find best rewritten claim with quality measures.

Results. Better quality in 60% of all cases; improvements similar to human revisions

LLMs for Claim Quality Optimization : Approach



LLMs for Claim Quality Optimization: Examples

- **Example 1**

Original. AGI are susceptable.	**Human (Reframing).** There is the threat that AIs will react agressively for being manipulated.	**Approach (Specification).** AGI are susceptible to being hacked.

- **Example 2**

Original. In Huckleberry Finn, Twain captured the essence of ”everyday midwest American English“	**Human (Specification).** In Huckleberry Finn, Twain captured the essence of ”[everyday midwest American English] <LINK>“. This quality makes it still relevant and worth teaching in the school system.	**Approach (Elaboration).** In Huckleberry Finn, Twain captured the essence of ”[everyday midwest American English]“ by using the N-word in everyday conversation.

Milad Alshomary

Henning Wachsmuth

LLMs for Counterargument Generation

(Alshomary and Wachsmuth, 2023)

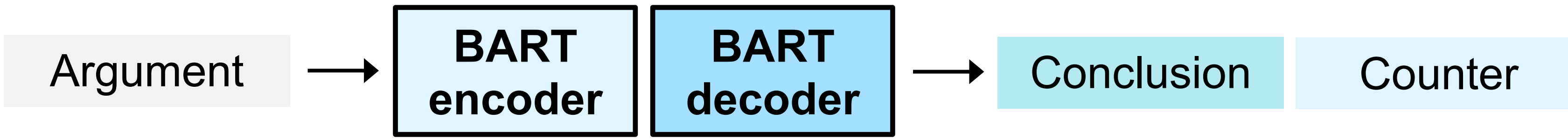
Problem. How to generate an effective counterargument to an argument?

Idea. Jointly generate counterarguments and conclusion; pick the one of most opposite stance

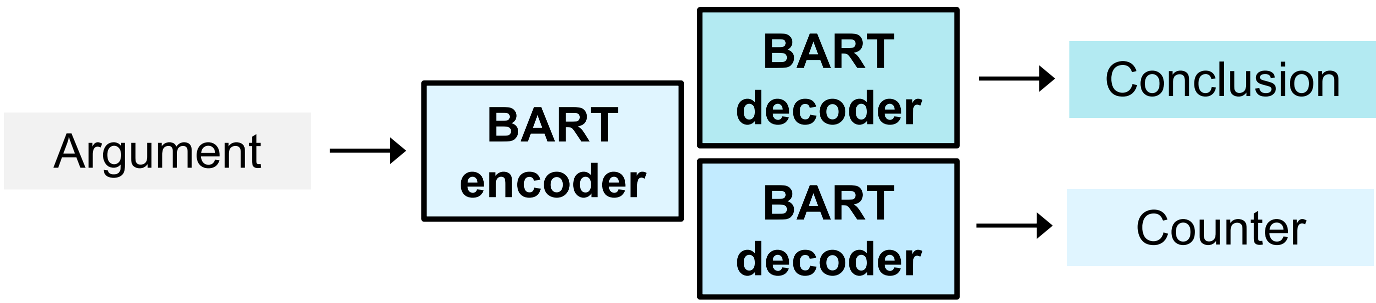
Results. Substantial improvement over counterarguments of fine-tuned LLMs

LLMs for Counterargument Generation: Approach

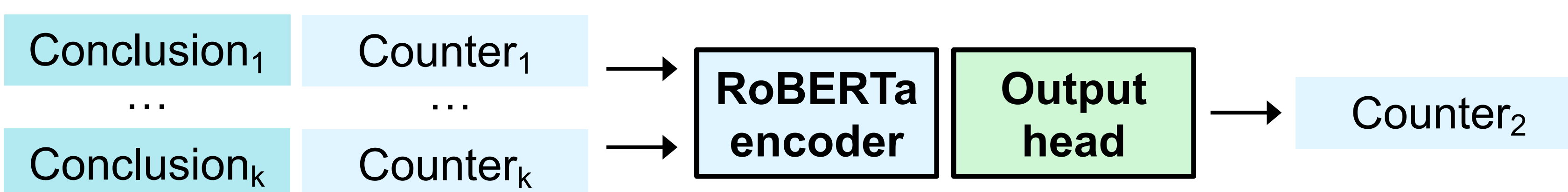
- **Multitask learning for candidate generation**
  - Variation 1. Generate both conclusion and counterargument with one decoder



  - Variation 2. Separated decoders for conclusion and counter



- **Stance regression for candidate ranking**
  - Generate diverse candidates, rank based on stance contrast, pick top-1



LLMs for Counterargument Generation: Examples

- **Argument**

*In my opinion **AGI needs the algorithms that make us intelligent** [...]. I hear arguments that an AGI needs to have a subjective experience like a human being in order to be generally intelligent and I find this theory to be lackluster **if the laws of physics can be computed then so can the algorithms that make general intelligence**. [...]*

- **Counterargument of baseline (fine-tuned BART)**

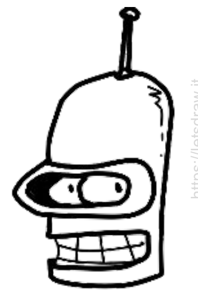
Intelligence is not determined by a subjective experience like a human being. It is determined by the laws of physics. An AGI does not need to have an objective experience to be intelligent, but it does need to be able to understand the world around it.

- **Conclusion and counterargument of approach**

Scientific law is the only thing that can make AGI generally intelligent.

*The problem with AGI is that **we don't really know what it is that makes us intelligent. We have no idea how it works,** what it's like to be an AGI, how it's different from a human being, or how it will work in the real world.*

GPT-4, Alpaca,
or similar

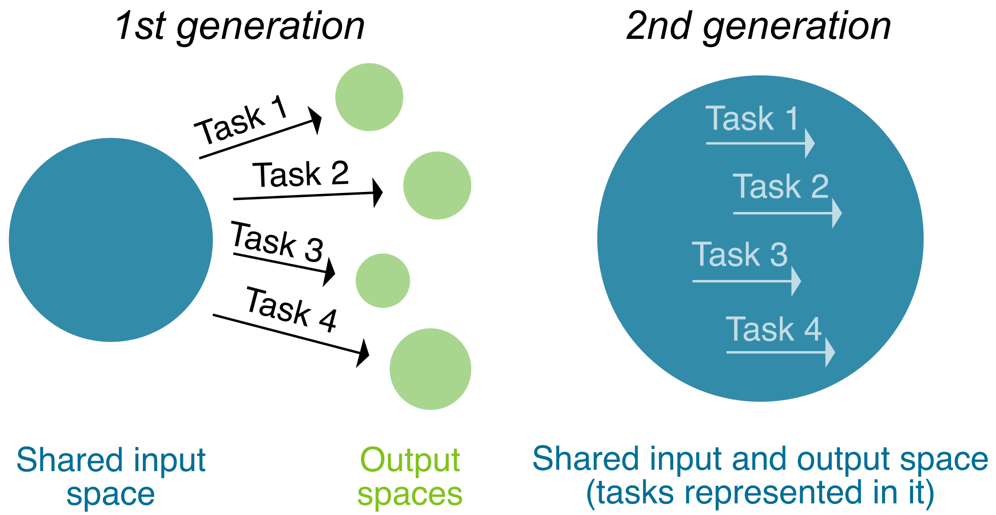


Second-generation LLMs

Instruction tuning

- **Instruction tuning**
  - Teach LLM how to solve tasks using natural language instructions
  - Each training instance is a task-like prompt with an adequate LLM response
  - Human feedback. Train reward model on human feedback to teach LLM
  - Self-instruct. Start from seed instructions, then let LLM generate instructions

- **Why instruction tuning?**
  - Fine-tuning leaves the task implicit in training pairs
  - Instruction tuning phrases it explicitly in the same space

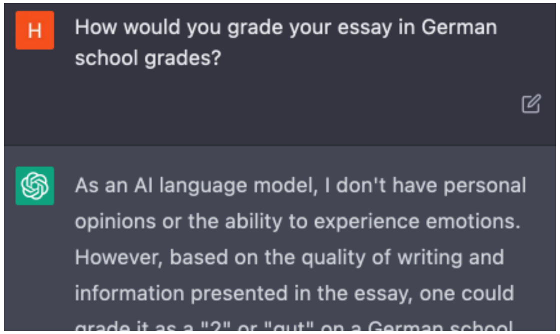


- **Beyond instruction tuning**
  - Overparameterization
  - Handling of long input sequences
  - In-context learning through prompts

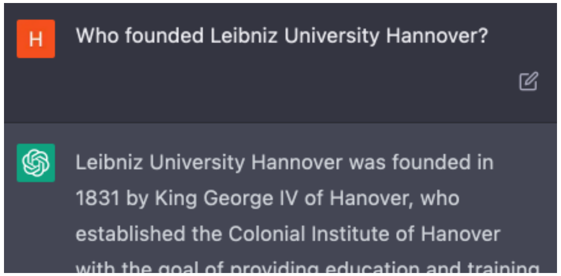
Instruction tuning: Paradigm shift

- **Paradigm shift**
  - NLP used to *predict* output labels and values
  - Second-generation LLMs *generate* labels and values
  - Input, task, and output all in the same representational space

- **LLMs for any task**
  - ChatGPT demonstrates that LLMs can tackle nearly any task *without fine-tuning*
  - Ultimately, fine-tuning may become obsolete
  - Instead of specific models, use one for all



- **Chances grow, risks too**
  - Disruptive potential for products and society
  - Hallucinated facts and bias among core issues
  - Misuse easy, we are barely prepared for this



LLMs for relevance assessments (Faggioli et al., 2023)

Best Paper Award at ICTIR 2023

- **Arising question**
  - If LLMs can tackle any NLP task, can they also replace humans in annotation and evaluation?
- **Example: Relevance in information retrieval (IR)**
  - IR builds on judgments of document relevance to queries
  - We assessed how much LLMs and humans agree on benchmark data

LLM	Prediction	TREC-8 Assessors		Cohen's κ
		Relevant	Not relevant	
GPT-3.5	Relevant	**237**	48	0.38
	Not relevant	263	**452**	
YouChat	Relevant	**33**	26	0.07
	Not relevant	67	**74**	

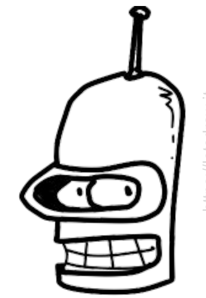
LLM	Prediction	TREC-DL 2021 Assessors				Cohen's κ
		3	2	1	0	
GPT-3.5	Relevant	**89**	**65**	**48**	16	0.40
	Not relevant	11	35	52	**84**	
YouChat	Relevant	**96**	**93**	**79**	42	0.49
	Not relevant	4	7	21	**58**	

- **Follow-up questions**
  - How can humans and LLMs best share their work?
  - Once LLMs create benchmarks, why would we still develop methods?
  - How do we notice when LLMs become better than humans?

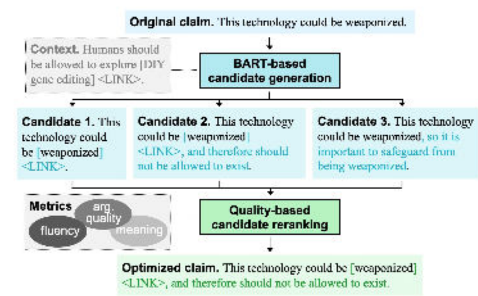
Takeaways

Conclusion and outlook

- **Large language models (LLMs)**
  - Language models predict most likely next words in sequences
  - Key concepts: Self-attention, transfer learning, instruction tuning
  - ChatGPT is based on an instruction-tuned LLM



- **Our research on LLMs so far**
  - Mostly first-generation LLMs for specific tasks
  - Often focus on how to control LLM behavior
  - First papers with second-generation LLMs upcoming



- **NLP research in the age of LLMs**
  - First generation impressive in generating human-like text
  - Second generation can tackle NLP tasks without training
  - LLMs change how NLP research works in some regards

References

- **Alshomary et al. (2023).** Milad Alshomary and Henning Wachsmuth. Conclusion-based Counter-Argument Generation. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, to appear, 2023.
- **Devlin et al. (2019).** Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, pages 4171–4186, 2019.
- **Gurcke et al. (2021).** Timon Gurcke, Milad Alshomary, and Henning Wachsmuth. Assessing the Sufficiency of Arguments through Conclusion Generation. In Proceedings of the 8th Workshop on Argument Mining, to appear, 2021.
- **Johnson and Blair (2006).** Ralph H. Johnson and J. Anthony Blair. 2006. Logical Self-defense. International Debate Education Association.
- **Lewis et al. (2019).** Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7871–7880, 2020.
- **Liu et al. (2019).** Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692, 2019.
- **Radford et al. (2018).** Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. OpenAI Blog, 2018.

References

- **Raffel et al. (2020).** Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020.
- **Sengupta et al. (2023).** Meghdut Sengupta, Milad Alshomary, Ingrid Scharlau, and Henning Wachsmuth. Modeling Highlighting of Metaphors in Multitask Contrastive Learning Paradigms. In Findings of the Association for Computational Linguistics: EMNLP 2023, to appear 2023.
- **Skitalinskaya et al. (2021).** Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. Learning From Revisions: Quality Assessment of Claims in Argumentation at Scale. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, pages 1718–1729, 2021.
- **Skitalinskaya et al. (2023).** Gabriella Skitalinskaya, Maximilian Spliethöver, and Henning Wachsmuth. Claim Optimization in Computational Argumentation. In Proceedings of the 16th International Natural Language Generation Conference, to appear, 2023.
- **Stab and Gurevych (2017).** Christian Stab and Iryna Gurevych. Recognizing Insufficiently Supported Arguments in Argumentative Essays. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pages 980–990, 2017.
- **Vaswani et al. (2017).** Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser. Attention Is All You Need. In 31st Conference on Neural Information Processing Systems, 2017.