Leibniz Universität Hannover

University of Bremen

Claim Optimization
in Computational Argumentation

Gabriella Skitalinskaya, Maximilian Splietöver, and Henning Wachsmuth

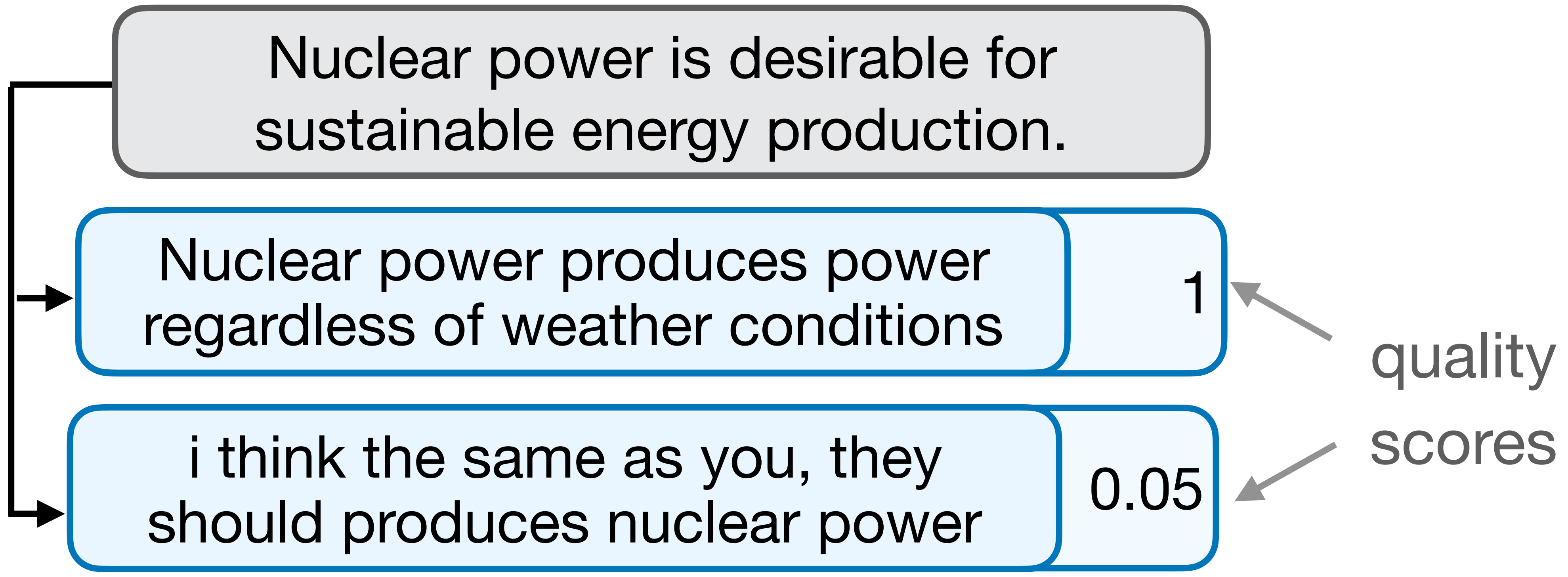
September 15, INLG 2023

Introduction

Motivation

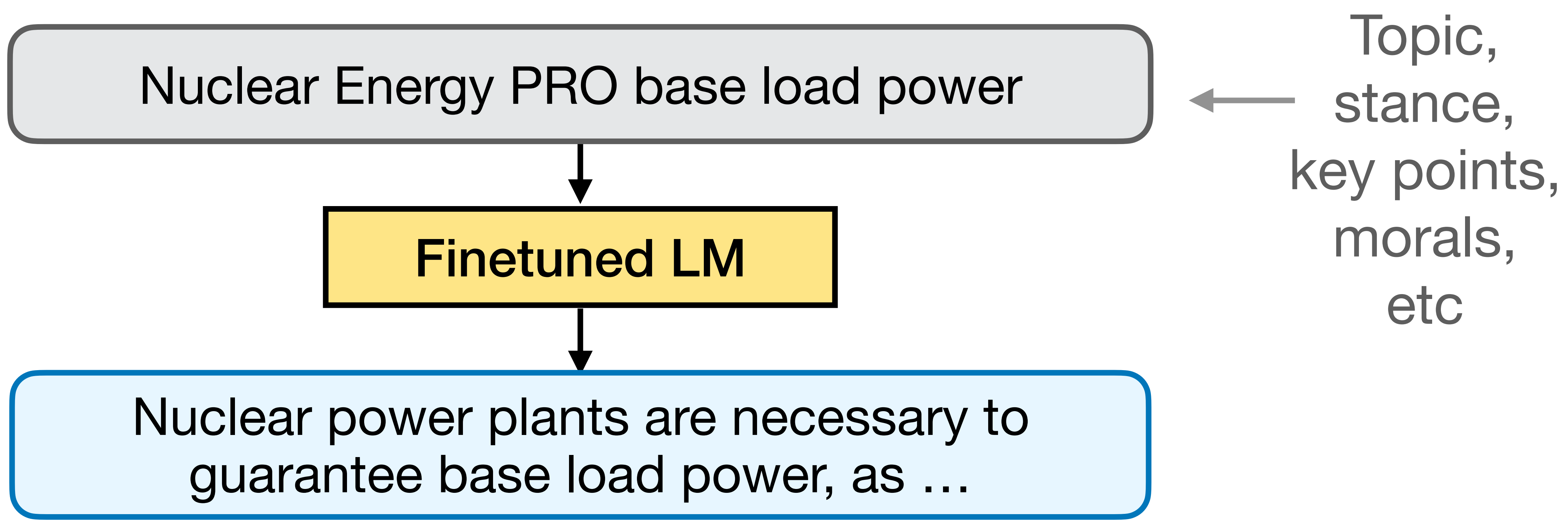
- For successful argumentation, the best arguments are needed.
- Prior research mainly frames the problem as a retrieval or generation task.

Ranking



(Syed et al. 2023; Dumani and Schenkel 2020; Gretz et al. 2020)

Generation



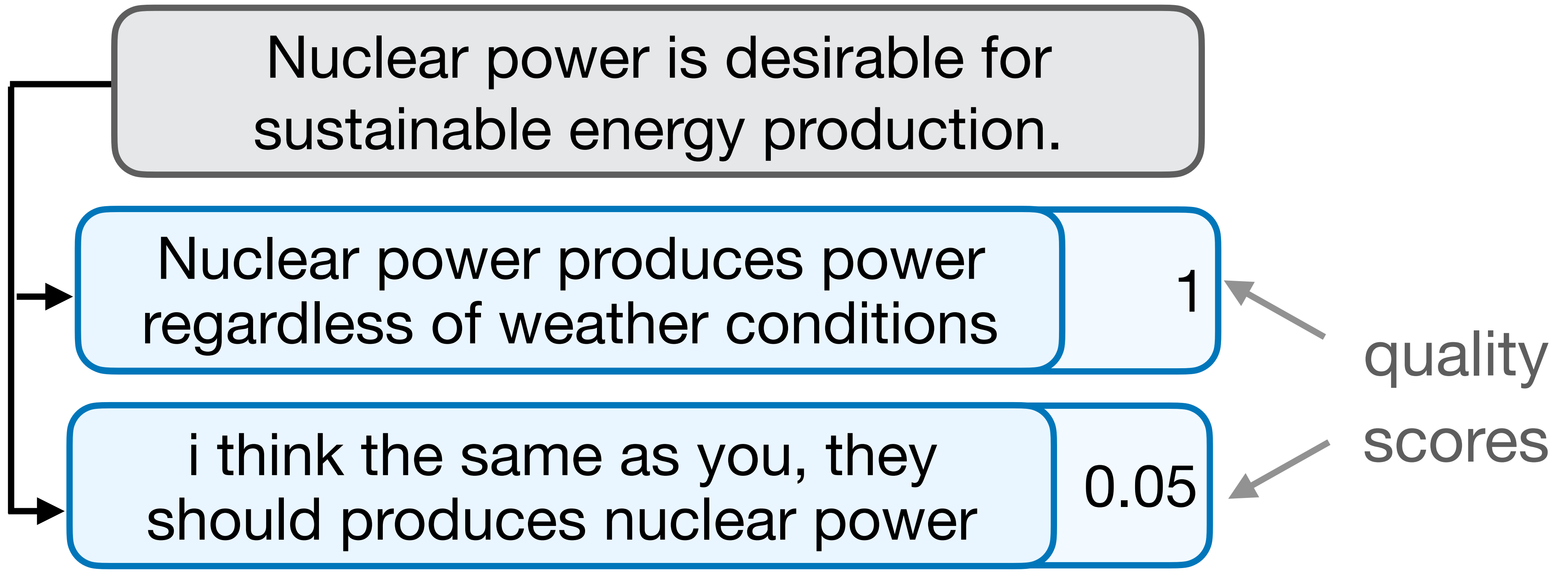
(Alshomary et al. 2022, Schiller et al. 2020)

Introduction

Motivation

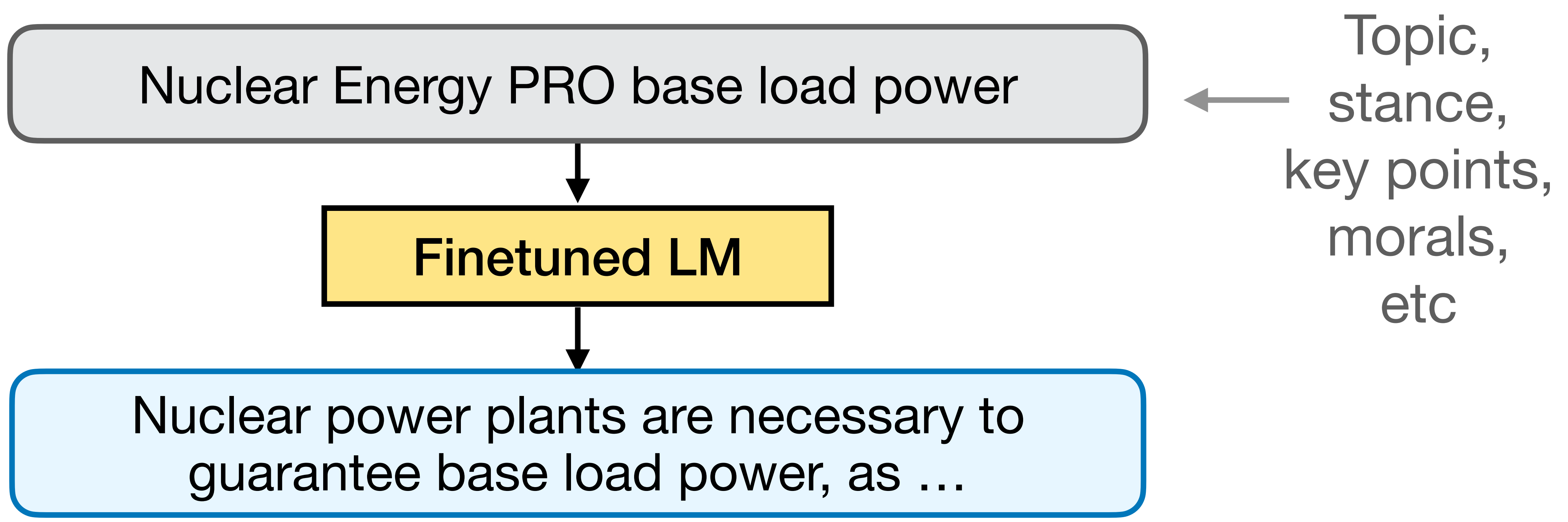
- For successful argumentation, the best arguments are needed.
- Prior research mainly frames the problem as a retrieval or generation task.

Ranking



(Syed et al. 2023; Dumani and Schenkel 2020; Gretz et al. 2020)

Generation



(Alshomary et al. 2022, Schiller et al. 2020)

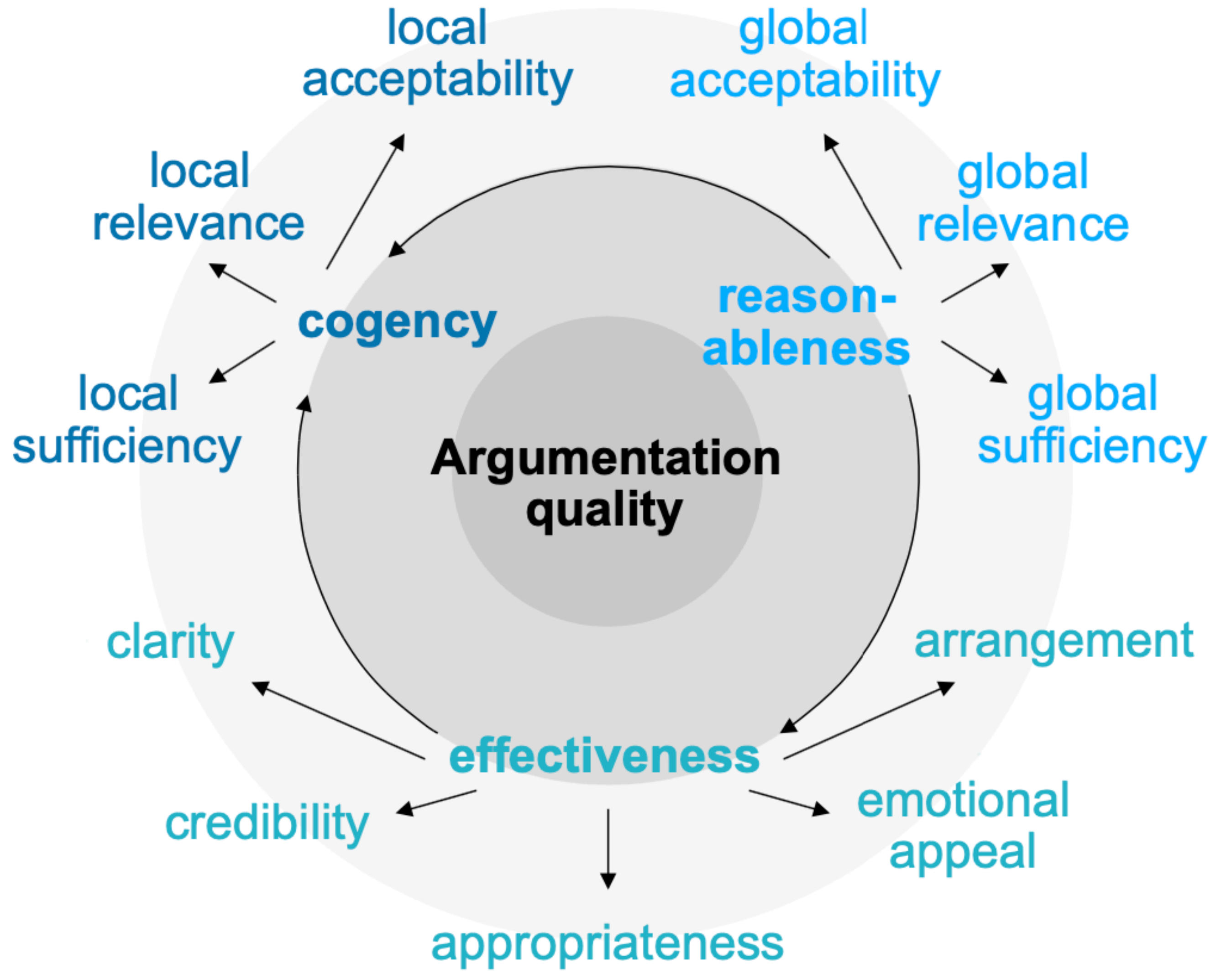
Suggestion. Instead, we help individuals improve their argumentative claims.

Introduction

Problem statement

Argument quality

- is inherently subjective
- depends on prior beliefs, stance, and one's subjective weighting of the discussed aspects



(Wachsmuth et al. 2017)

Problem

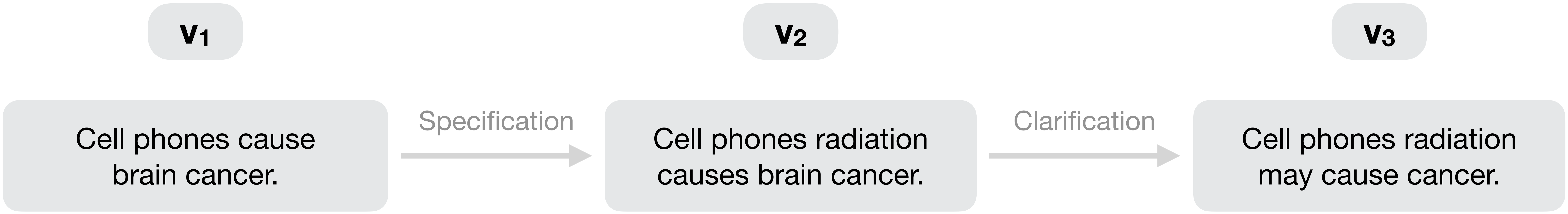
- How can we improve argumentative text, if quality is so subjective?

Introduction

Revisions in Argumentative Writing

Suggestion

- learn from different revisions of the same argumentative text
(Skitalinskaya et al. 2021; Skitalinskaya and Wachsmuth 2023)



Text revision

- essential part of argumentative writing
- typically a recursive process until an optimal phrasing is achieved
- phrasing directly influences the persuasive impact on the audience

Suggested Task

Claim Quality Optimization

Task

Given as input an argumentative claim, potentially along with context information,

This technology could be weaponized.	Humans should be allowed to explore DIY gene editing.

Suggested Task

Claim Quality Optimization

Task

Given as input an argumentative claim, potentially along with context information,

This technology could be weaponized.

Humans should be allowed to explore DIY gene editing.

rewrite it so the generated output improves the text and/or argument quality while preserving the original meaning.

Suggested Task

Claim Quality Optimization

Task

Given as input an argumentative claim, potentially along with context information,

This technology could be weaponized.

Humans should be allowed to explore DIY gene editing.

rewrite it so the generated output improves the text and/or argument quality while preserving the original meaning.

This technology could be weaponized, so it is important to safeguard it from being weaponized.

Suggested Task

Claim Quality Optimization

Task

Given as input an argumentative claim, potentially along with context information,

This technology could be weaponized.

Humans should be allowed to explore DIY gene editing.

rewrite it so the generated output improves the text and/or argument quality while preserving the original meaning.

This technology could be weaponized, so it is important to safeguard it from being weaponized.

This technology could be used by criminals to create and weaponize bio-mechanisms.

Suggested Task

Claim Quality Optimization

Task

Given as input an argumentative claim, potentially along with context information,

This technology could be weaponized.

Humans should be allowed to explore DIY gene editing.

rewrite it so the generated output improves the text and/or argument quality while preserving the original meaning.

This technology could be weaponized, so it is important to safeguard it from being weaponized.

This technology could be used by criminals to create and weaponize bio-mechanisms.

This technology could be weaponized and harmful to human beings.

Suggested Task

Claim Quality Optimization

Task

Given as input an argumentative claim, potentially along with context information,

This technology could be weaponized.

Humans should be allowed to explore DIY gene editing.

rewrite it so the generated output improves the text and/or argument quality while preserving the original meaning.

This technology could be weaponized, so it is important to safeguard it from being weaponized.	This technology could be used by criminals to create and weaponize bio-mechanisms.	This technology could be weaponized and harmful to human beings.

But how to decide which candidate is the best one?

Approach

BART-based Candidate Generation and Quality-based Reranking

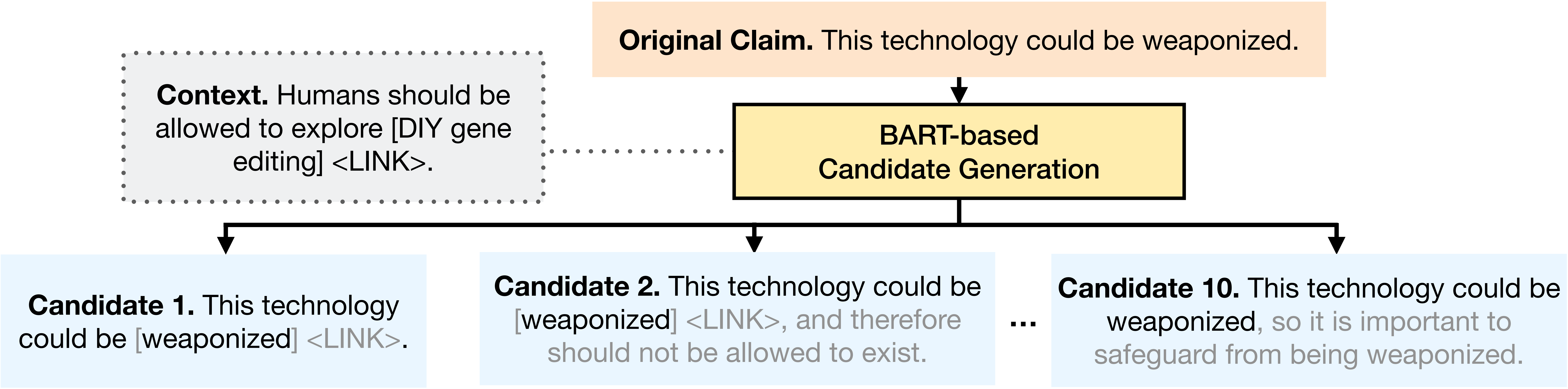
Original Claim. This technology could be weaponized.

Context. Humans should be allowed to explore [DIY gene editing] <LINK>.

BART-based
Candidate Generation

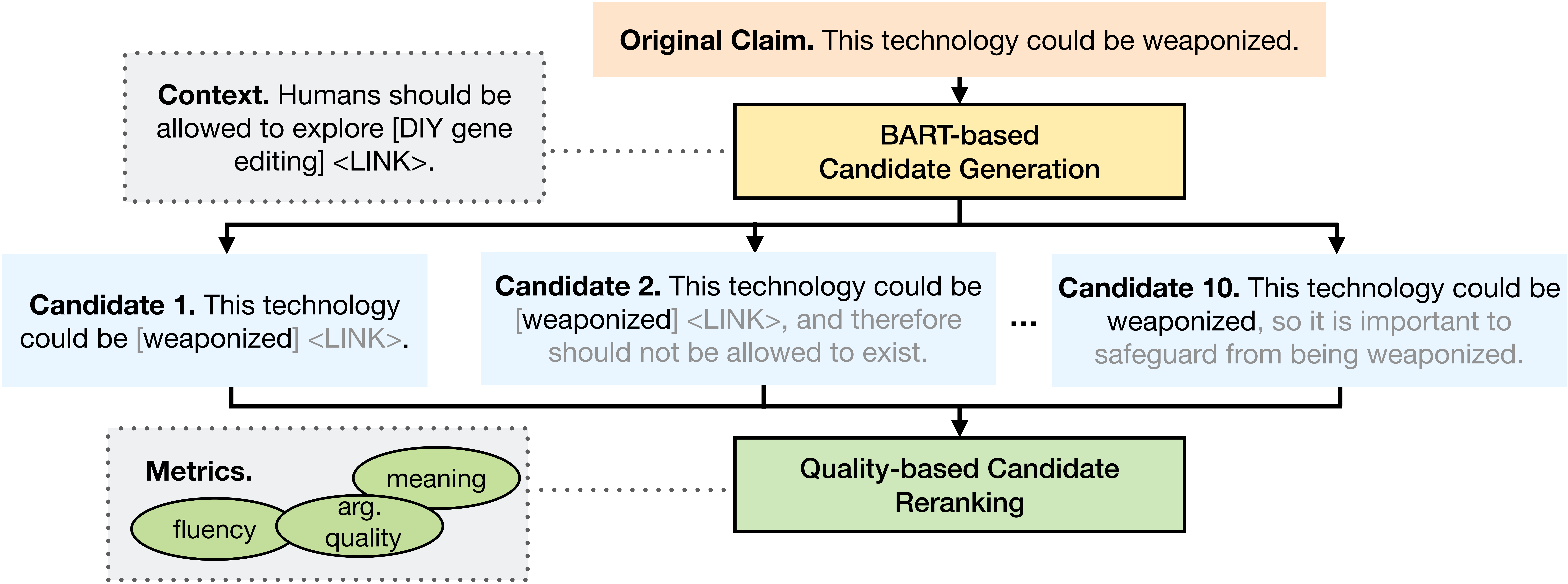
Approach

BART-based Candidate Generation and Quality-based Reranking



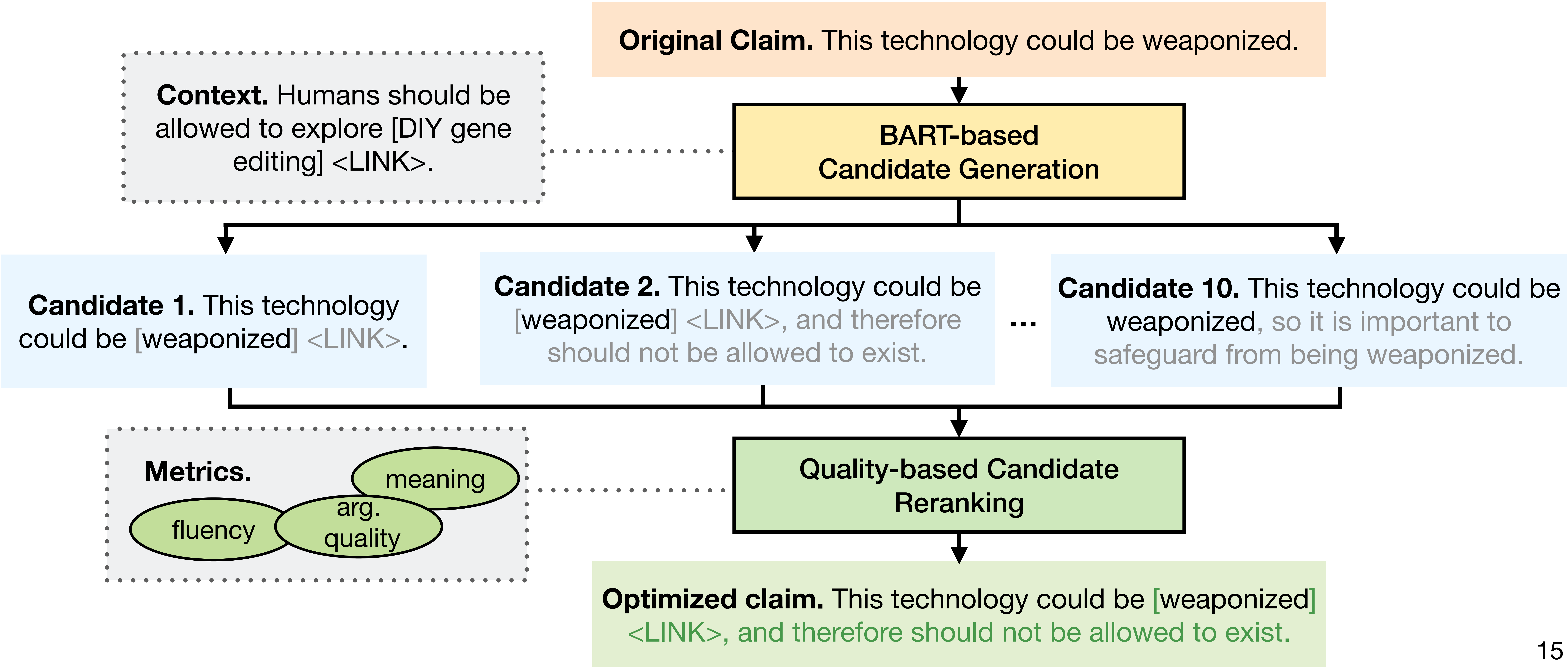
Approach

BART-based Candidate Generation and Quality-based Reranking



Approach

BART-based Candidate Generation and Quality-based Reranking



Quality Assessment Metrics

To identify the optimal claim among the generated candidates we consider the following text and argument quality metrics:

- Grammatical Fluency. Absolute assessments of text variations (MSR corpus) (Toutanova et al. 2016)
- Argument Quality. Relative assessments of argumentative text variations (Skitalinskaya et al. 2021)
- Meaning Preservation. Semantic similarity of SBERT embeddings (Reimers and Gurevych 2019)

Quality-Based Reranking

- To favor certain dimensions we integrate the metrics as the weighted linear sum of individual scores:

$$Score = \alpha \cdot fluency + \beta \cdot meaning + \gamma \cdot argument, \quad \alpha + \beta + \gamma = 1, \quad \alpha, \beta, \gamma \in [0,1]$$

Quality-Based Reranking

- To favor certain dimensions we integrate the metrics as the weighted linear sum of individual scores:

$$Score = \alpha \cdot fluency + \beta \cdot meaning + \gamma \cdot argument, \quad \alpha + \beta + \gamma = 1, \quad \alpha, \beta, \gamma \in [0,1]$$

- Optimal weights are found via grid search by maximizing Pearson's correlation coefficient between the weighted score and the original order of the revisions in the revision history.

Quality-Based Reranking

- To favor certain dimensions we integrate the metrics as the weighted linear sum of individual scores:

$$Score = \alpha \cdot fluency + \beta \cdot meaning + \gamma \cdot argument, \quad \alpha + \beta + \gamma = 1, \quad \alpha, \beta, \gamma \in [0,1]$$

- Optimal weights are found via grid search by maximizing Pearson's correlation coefficient between the weighted score and the original order of the revisions in the revision history.

	Fluency	Meaning	Argument	Score
Claim Version 1	0.6	0.9	0.4	
Claim Version 2	0.7	0.8	0.8	
...				
Claim Version N	0.9	0.9	0.9	

Quality-Based Reranking

- To favor certain dimensions we integrate the metrics as the weighted linear sum of individual scores:

$$Score = \alpha \cdot fluency + \beta \cdot meaning + \gamma \cdot argument, \quad \alpha + \beta + \gamma = 1, \quad \alpha, \beta, \gamma \in [0,1]$$

- Optimal weights are found via grid search by maximizing Pearson's correlation coefficient between the weighted score and the original order of the revisions in the revision history.

	Fluency	Meaning	Argument	Score
Claim Version 1	0.6	0.9	0.4	0.49
Claim Version 2	0.7	0.8	0.8	0.76
...				
Claim Version N	0.9	0.9	0.9	0.90

Quality-Based Reranking

- To favor certain dimensions we integrate the metrics as the weighted linear sum of individual scores:

$$Score = \alpha \cdot fluency + \beta \cdot meaning + \gamma \cdot argument, \quad \alpha + \beta + \gamma = 1, \quad \alpha, \beta, \gamma \in [0,1]$$

- Optimal weights are found via grid search by maximizing Pearson's correlation coefficient between the weighted score and the original order of the revisions in the revision history.

	Fluency	Meaning	Argument	Score
Claim Version 1	0.6	0.9	0.4	0.49
Claim Version 2	0.7	0.8	0.8	0.76
…				
Claim Version N	0.9	0.9	0.9	0.90

Quality-Based Reranking

- To favor certain dimensions we integrate the metrics as the weighted linear sum of individual scores:

$$Score = \alpha \cdot fluency + \beta \cdot meaning + \gamma \cdot argument, \quad \alpha + \beta + \gamma = 1, \quad \alpha, \beta, \gamma \in [0,1]$$

- Optimal weights are found via grid search by maximizing Pearson's correlation coefficient between the weighted score and the original order of the revisions in the revision history.

	Fluency	Meaning	Argument	Score
Claim Version 1	0.6	0.9	0.4	0.49
Claim Version 2	0.7	0.8	0.8	0.76
...				
Claim Version N	0.9	0.9	0.9	0.90

Quality-Based Reranking

- To favor certain dimensions we integrate the metrics as the weighted linear sum of individual scores:

$$Score = \alpha \cdot fluency + \beta \cdot meaning + \gamma \cdot argument, \quad \alpha + \beta + \gamma = 1, \quad \alpha, \beta, \gamma \in [0,1]$$

- Optimal weights are found via grid search by maximizing Pearson's correlation coefficient between the weighted score and the original order of the revisions in the revision history.

	Fluency	Meaning	Argument	Score
Claim Version 1	0.6	0.9	0.4	0.49
Claim Version 2	0.7	0.8	0.8	0.76
...				
Claim Version N	0.9	0.9	0.9	0.90

$$\alpha = 0.43$$
$$\beta = 0.01$$
$$\gamma = 0.56$$

Quality-Based Reranking

- To favor certain dimensions we integrate the metrics as the weighted linear sum of individual scores:

$$Score = \alpha \cdot fluency + \beta \cdot meaning + \gamma \cdot argument, \quad \alpha + \beta + \gamma = 1, \quad \alpha, \beta, \gamma \in [0,1]$$

- Optimal weights are found via grid search by maximizing Pearson's correlation coefficient between the weighted score and the original order of the revisions in the revision history.

	Fluency	Meaning	Argument	Score
Claim Version 1	0.6	0.9	0.4	0.49
Claim Version 2	0.7	0.8	0.8	0.76
…				
Claim Version N	0.9	0.9	0.9	0.90

$$\alpha = 0.43$$
$$\beta = 0.01$$
$$\gamma = 0.56$$

	Fluency	Meaning	Argument	Score
Candidate 1	0.7	0.4	0.8	0.75
Candidate 2	0.8	0.7	0.9	0.86
…				
Candidate N	0.5	0.9	0.6	0.56

Quality-Based Reranking

- To favor certain dimensions we integrate the metrics as the weighted linear sum of individual scores:

$$Score = \alpha \cdot fluency + \beta \cdot meaning + \gamma \cdot argument, \quad \alpha + \beta + \gamma = 1, \quad \alpha, \beta, \gamma \in [0,1]$$

- Optimal weights are found via grid search by maximizing Pearson's correlation coefficient between the weighted score and the original order of the revisions in the revision history.

	Fluency	Meaning	Argument	Score
Claim Version 1	0.6	0.9	0.4	0.49
Claim Version 2	0.7	0.8	0.8	0.76
…				
Claim Version N	0.9	0.9	0.9	0.90

$$\alpha = 0.43$$
$$\beta = 0.01$$
$$\gamma = 0.56$$

	Fluency	Meaning	Argument	Score
Candidate 1	0.7	0.4	0.8	0.75
Candidate 2	0.8	0.7	0.9	0.86
…				
Candidate N	0.5	0.9	0.6	0.56

Experimental setup

- **Experiments**
  - Data. 190K claim revisions from Kialo, 600 for manual evaluation
  - Approaches. BART combined with reranking approaches and baselines
- **Ranking Baselines**
  - Top-1. Returns BART's most likely output
  - Random. Returns any of the 10 candidates pseudo-randomly
  - SVMRank. Returns best candidate based on existing ranker (Skitalinskaya et al. 2021)

Evaluation Results

Approach	Automatic					Human
	BLEU	Rouge-L	SARI	NoEdit ↓	ExM	Rank ↓
Baselines						
Unedited	**69.4**	**0.87**	27.9	1.00	0.0%	-
BART + Top-1	64.0	0.83	39.7	0.31	7.8%	2.16
BART + Random	62.6	0.83	38.7	0.28	6.8%	2.06
BART + SVMRank	55.7	0.76	38.8	0.03	4.5%	1.95
Approach						
BART + Ours	59.4	0.80	**43.7**	**0.02**	**8.3%**	**1.92**

Evaluation Results

Approach	BLEU	Rouge-L	SARI	NoEdit ↓	ExM	Rank ↓
	Automatic					Human
Baselines						
Unedited	**69.4**	**0.87**	27.9	1.00	0.0%	-
BART + Top-1	64.0	0.83	39.7	0.31	7.8%	2.16
BART + Random	62.6	0.83	38.7	0.28	6.8%	2.06
BART + SVMRank	55.7	0.76	38.8	0.03	4.5%	1.95
Approach						
BART + Ours	59.4	0.80	**43.7**	**0.02**	**8.3%**	**1.92**

- High scores of *Unedited* on BLEU indicate that many human revisions introduce few changes.

Evaluation Results

	Automatic					Human
Approach	**BLEU**	**Rouge-L**	**SARI**	**NoEdit ↓**	**ExM**	**Rank ↓**
Baselines						
Unedited	**69.4**	**0.87**	27.9	1.00	0.0%	-
BART + Top-1	64.0	0.83	39.7	0.31	7.8%	2.16
BART + Random	62.6	0.83	38.7	0.28	6.8%	2.06
BART + SVMRank	55.7	0.76	38.8	0.03	4.5%	1.95
Approach						
BART + Ours	59.4	0.80	**43.7**	**0.02**	**8.3%**	**1.92**

- High scores of *Unedited* on BLEU indicate that many human revisions introduce few changes.
- *BART* + *Ours* performs best on SARI.

Evaluation Results

Approach	Automatic					Human
	BLEU	Rouge-L	SARI	NoEdit ↓	ExM	Rank ↓
Baselines						
Unedited	**69.4**	**0.87**	27.9	1.00	0.0%	-
BART + Top-1	64.0	0.83	39.7	0.31	7.8%	2.16
BART + Random	62.6	0.83	38.7	0.28	6.8%	2.06
BART + SVMRank	55.7	0.76	38.8	0.03	4.5%	1.95
Approach						
BART + Ours	59.4	0.80	**43.7**	**0.02**	**8.3%**	**1.92**

- High scores of *Unedited* on BLEU indicate that many human revisions introduce few changes.
- *BART + Ours* performs best on SARI.
- Human annotators prefer optimized candidates selected by our approach.

Optimization Type Taxonomy

Simplification	Elaboration	Disambiguation	Neutralization
Specification	Corroboration	Copy editing	Reframing

Optimization Type Taxonomy

Simplification	Elaboration	Disambiguation	Neutralization
Specification	Corroboration	Copy editing	Reframing

Specifying or explaining a given fact or meaning (of the argument) by adding an example or discussion without adding new information.

It is very common for governments to actively make certain forms of healthcare [harder for minority groups to access] <LINK>. They could also, therefore, make cloning technology hard to access.

Optimization Type Taxonomy

Simplification	Elaboration	Disambiguation	Neutralization
Specification	**Corroboration**	Copy editing	Reframing

Adding, editing, or removing evidence in the form of links that provide supporting information or external resources to the claim.

[Person-based predictive policing technologies] **<LINK>** - that focus on predicting who is likely to commit crime rather than where is it likely to occur - violate the [presumption of innocence.] **<LINK>**.

Optimization Type Taxonomy

Simplification	Elaboration	Disambiguation	Neutralization
Specification	Corroboration	**Copy editing**	Reframing

Improving the grammar, spelling, tone, or punctuation of a claim, without changing the main point or meaning.

Women are experiencing record ~~level~~ levels of success in primaries.

Performance across Optimization Types

- Jaccard similarity of human and generated revisions is 0.37.

Type	Human	Approach	Better	Worse
Specification	59	152	65%	16%
Simplification	43	18	61%	11%
Reframing	29	21	62%	5%
Elaboration	23	55	62%	20%
Corroboration	161	38	53%	24%
Neutralization	7	0	-	-
Disambiguation	8	8	63%	12%
Copy editing	293	301	59%	15%
Overall	623	593	60%	16%

Performance across Optimization Types

- Jaccard similarity of human and generated revisions is 0.37.
- Specification is performed 2.5 times more often compared to humans.

Type	**Human**	**Approach**	**Better**	**Worse**
Specification	59	152	65%	16%
Simplification	43	18	61%	11%
Reframing	29	21	62%	5%
Elaboration	23	55	62%	20%
Corroboration	161	38	53%	24%
Neutralization	7	0	-	-
Disambiguation	8	8	63%	12%
Copy editing	293	301	59%	15%
Overall	623	593	60%	16%

Performance across Optimization Types

- Jaccard similarity of human and generated revisions is 0.37.
- Specification is performed 2.5 times more often compared to humans.
- Corroboration is performed 4 times less often than humans.

Type	Human	Approach	Better	Worse
Specification	59	152	65%	16%
Simplification	43	18	61%	11%
Reframing	29	21	62%	5%
Elaboration	23	55	62%	20%
Corroboration	161	38	53%	24%
Neutralization	7	0	-	-
Disambiguation	8	8	63%	12%
Copy editing	293	301	59%	15%
Overall	623	593	60%	16%

Performance across Optimization Types

- Jaccard similarity of human and generated revisions is 0.37.
- Specification is performed 2.5 times more often compared to humans.
- Corroboration is performed 4 times less often than humans.
- Elaboration and corroboration have the highest rate of unsuccessful revisions.

Type	Human	Approach	Better	Worse
Specification	59	152	65%	16%
Simplification	43	18	61%	11%
Reframing	29	21	62%	5%
Elaboration	23	55	62%	20%
Corroboration	161	38	53%	24%
Neutralization	7	0	-	-
Disambiguation	8	8	63%	12%
Copy editing	293	301	59%	15%
Overall	623	593	60%	16%

What Else Can Be Found in Paper

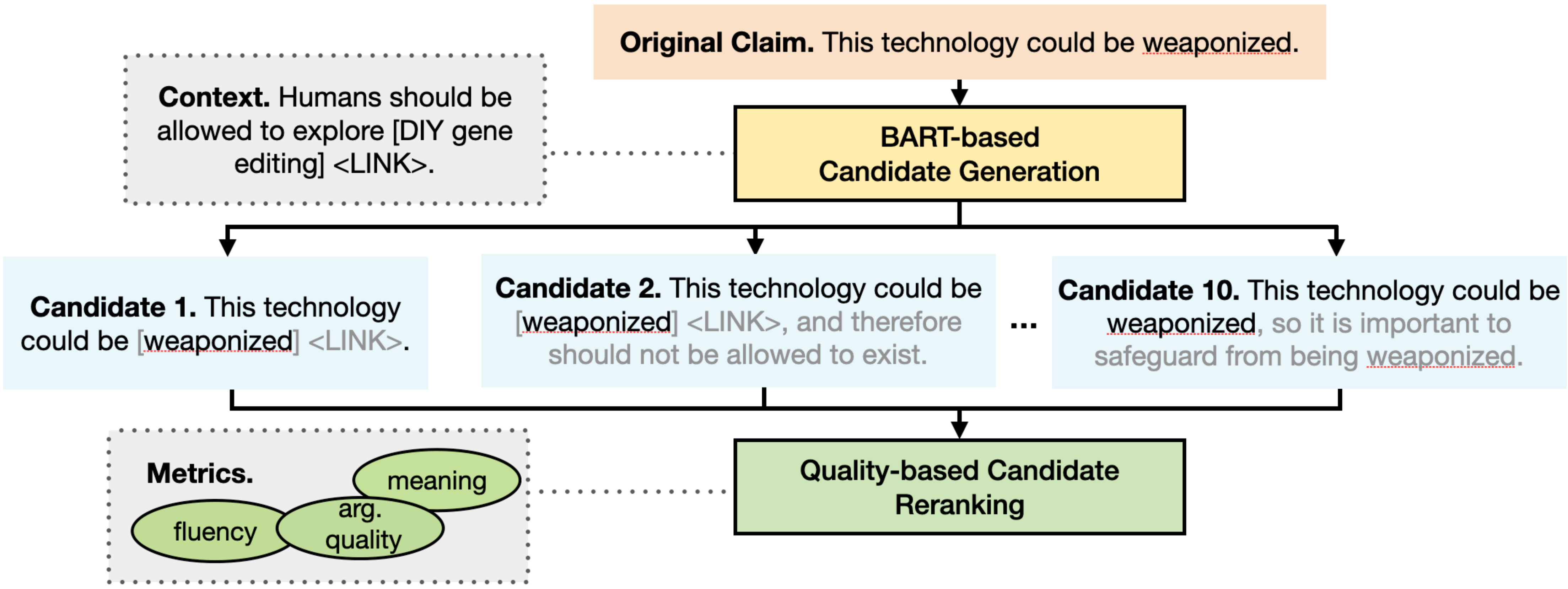
More details regarding

- the suggested approach
- experimental results
- examples of generated optimizations

And more experiments and discussion on

- relationship between revision intentions and optimization types
- how context can be used to improve the quality of generated texts
- how the approach generalizes to other domains of text

Takeaways



Contributions

- New task of claim optimization
- A computational approach combining quality-based reranking with text generation
- Taxonomy of optimization types and challenges in modelling them computationally

(Select) Findings

- Utilising context information increases the quality of generated texts
- Approach and quality metrics generalize to other domains
- Corroboration and elaboration types were found as hard to automate

- **Code repository:** https://github.com/GabriellaSky/claim_optimization

References

- **Alshomary et al. (2022)** Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. 2022. The Moral Debater: A Study on the Computational Generation of Morally Framed Arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), pages 8782–8797, Dublin, Ireland. Association for Computational Linguistics.

- **Dumani and Schenkel (2020)** Lorik Dumani and Ralf Schenkel. 2020. Quality-Aware Ranking of Arguments. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (CIKM '20). Association for Computing Machinery, New York, NY, USA, 335–344.

- **Gretz et al., (2020)** Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A Large-scale Dataset for Argument Quality Ranking: Construction and Analysis. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence* (AAAI-20).

- **Reimers and Gurevych (2019)** Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

- **Schiller et al. (2020)** Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-Controlled Neural Argument Generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–396, Online. Association for Computational Linguistics.

References

- **Skitalinskaya et al. (2021)** Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. 2021. 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: Main Volume*, pages 1718–1729, Online. Association for Computational Linguistics.
- **Skitalinskaya et al. (2023)** Gabriella Skitalinskaya and Henning Wachsmuth. 2023. To Revise or Not to Revise: Learning to Detect Improvable Claims for Argumentative Writing Support. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15799–15816, Toronto, Canada. Association for Computational Linguistics.
- **Syed et al. (2023)** Shahbaz Syed, Timon Ziegenbein, Philipp Heinisch, Henning Wachsmuth, Martin Potthast Frame-Oriented Summarization Of Argumentative Discussions In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Prague, Czech Republic. Association for Computational Linguistics.
- **Toutanova et al. (2016)** Kristina Toutanova, Chris Brockett, Ke M. Tran, and Saleema Amershi. 2016. A Dataset and Evaluation Metrics for Abstractive Compression of Sentences and Short Paragraphs. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 340–350, Austin, Texas. Association for Computational Linguistics.
- **Wachmsuth 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.