Towards Qualitative Measurement at Scale: Automated Discourse Quality Index (DQI) Annotation Using Large Language Models

Mitchell Bosley

PhD Candidate, Department of Political Science, University of Michigan

Postdoctoral Fellow, Munk School of Global Affairs and Public Policy and Schwartz Reissman Institute, University of Toronto

Introduction

- The Discourse Quality Index (DQI; Steenbergen 2003, Bachtiger 2014) is a framework for evaluating the quality of deliberation in political discussions.
- Manual DQI coding is time-intensive and resource-demanding.
- I explore the use of Large Language Models (LLMs) as substitutes for expert coders for automated DQI annotation.

DQI Dimensions and Coding

1. Participation
   - 0 Impaired participation
   - 1 Normal participation
2. Justification
   - 2.1 Level (0-4)
     - 0 No justification
     - 1 Inferior justification
     - 2 Qualified justification
     - 3 Sophisticated justification (broad)
     - 4 Sophisticated justification (in depth)
   - 2.2 Content (0-3)
     - 0 Group interests
     - 1 Neutral statement
     - 2 Common good (utilitarian/collective)
     - 3 Helping least advantaged
3. Respect
   - 3.1 Towards groups (0-2)
   - 3.2 Towards main demand (0-3)
   - 3.3 Towards counterarguments (0-4)
4. Constructive Politics (0-3)
   - 0 Positional politics
   - 1 Alternative proposal
   - 2 Consensus appeal
   - 3 Mediating proposal

Data

- 1000 human-validated speeches from the 101st and 104th US Congress (Steenbergen and Bachtiger 2004).
- Debates on various topics such as gun rights, health care, and abortion.
- Data split: 80% (800 speeches) used for generating examples for in-context learning, 20% (200 speeches) reserved for evaluation of model performance.

Methodology

- Accessed various closed- and open-source LLMs: GPT-4, Claude, DeepSeek, Meta's LLaMA, etc. through APIs.
- Implemented different prompting strategies:
  - Zero-shot, few-shot, and many-shot in-context learning.
  - Chain-of-Thought (CoT) reasoning
- For DeepSeek models and Claude Haiku, averaged over 5 random draws of examples for the ICL process per specifiation.
- Evaluated models on out-of-sample accuracy, F1 score, Mean Average Error (MAE).
- Total spending roughly $400 USD.

Prompt Structure

1. Theoretical justification:
   - References to research on DQI and deliberation
   - Explanation of how LLMs can be applied to automate DQI annotation
2. Previous examples of annotated speeches:
   - Contextualized examples for in-context learning
   - Helps model understand annotation task and expected output
3. Annotation instructions:
   - Provided as a JSON schema
   - Includes details on DQI dimensions and scoring

Results

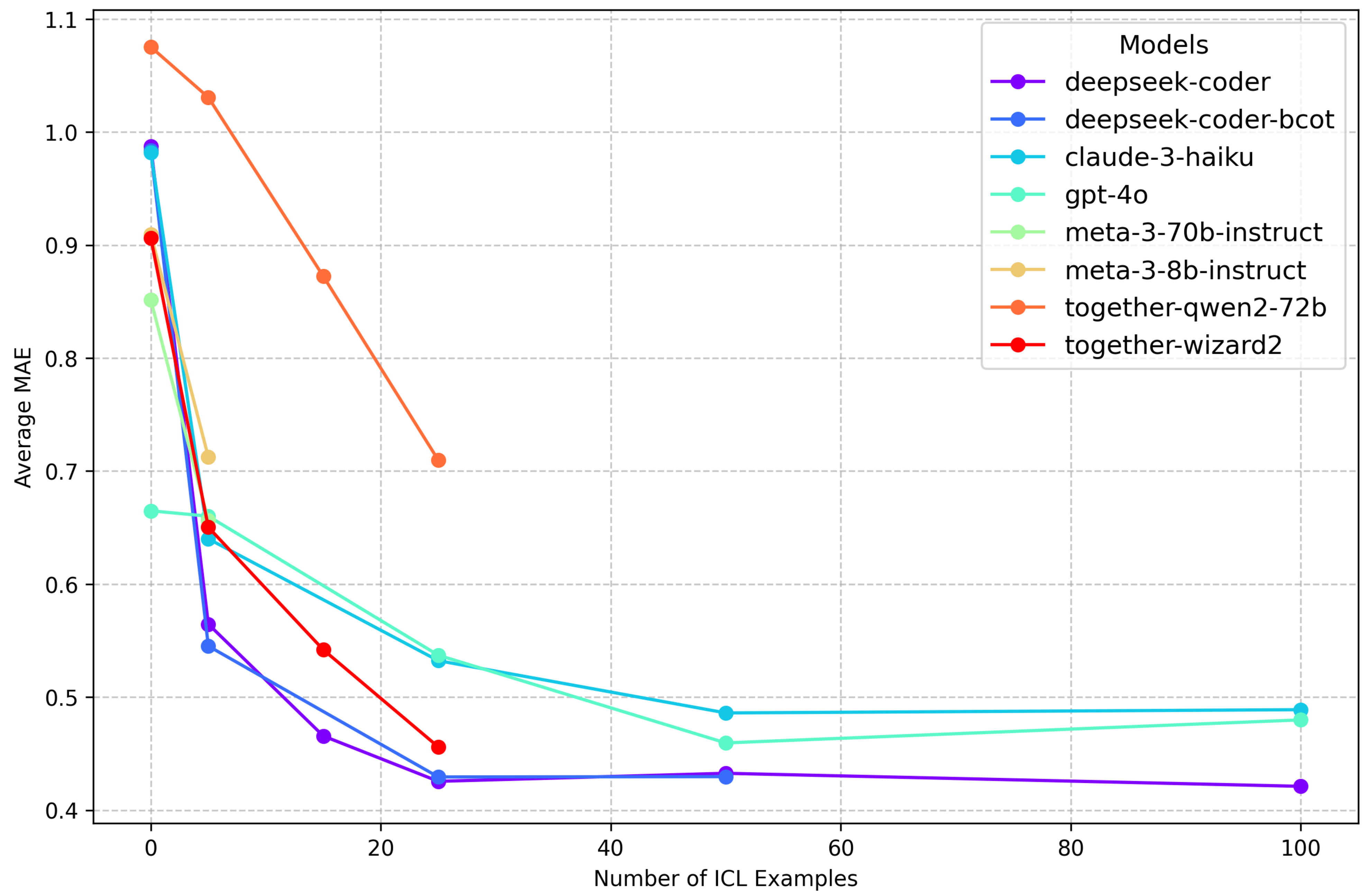

Figure 1:Model Performance (Mean Average Error) vs Number of ICL Examples

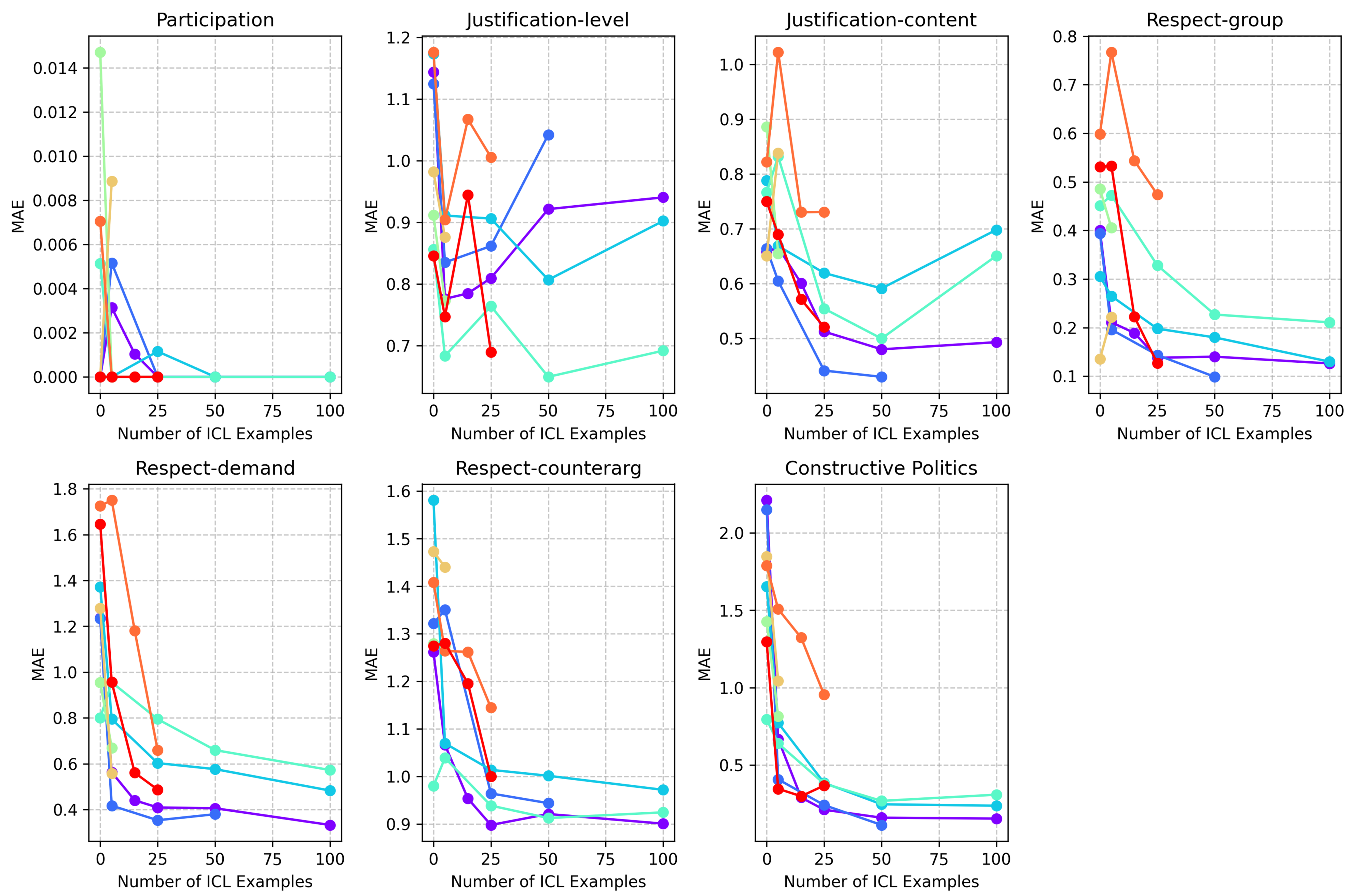

Figure 2:Model Performance (Mean Average Error) vs Number of ICL Examples by DQI Dimension

Cost vs. Performance Analysis

Model	Provider	Parameters	Context	Cost ($/M tokens)		Optimal In-Context Learning Performance			
			Length	Input	Output	Examples	Accuracy	F1 Score	MAE
GPT-4o	OpenAI	N/A	128k	5.00	15.00	50	0.6382	0.6454	0.4595
Claude 3 Haiku	Anthropic	N/A	200k	0.25	1.25	25	0.6366	0.6343	0.5322
DeepSeek Coder 2	**DeepSeek**	**236b**	**128k**	**0.14**	**0.28**	**100**	**0.6826**	**0.6525**	**0.4799**
DeepSeek Coder 2 (CoT)	DeepSeek	236b	128k	0.14	0.28	25	0.6635	0.6304	0.4294
Llama 3 70B	Meta	70B	8k	0.65	2.75	5	0.6120	0.6023	0.5452
Llama 3 8B	Meta	8B	8k	0.05	0.25	5	0.5422	0.5530	0.7124
Qwen2 72B	Alibaba	72B	32k	0.90	0.90	25	0.6466	0.6110	0.4560
WizardLM 2	Microsoft	176B	32k	1.20	1.20	25	0.6466	0.6110	0.4560

Table 1:Model Comparison with Optimal In-Context Learning Performance for DQI Task

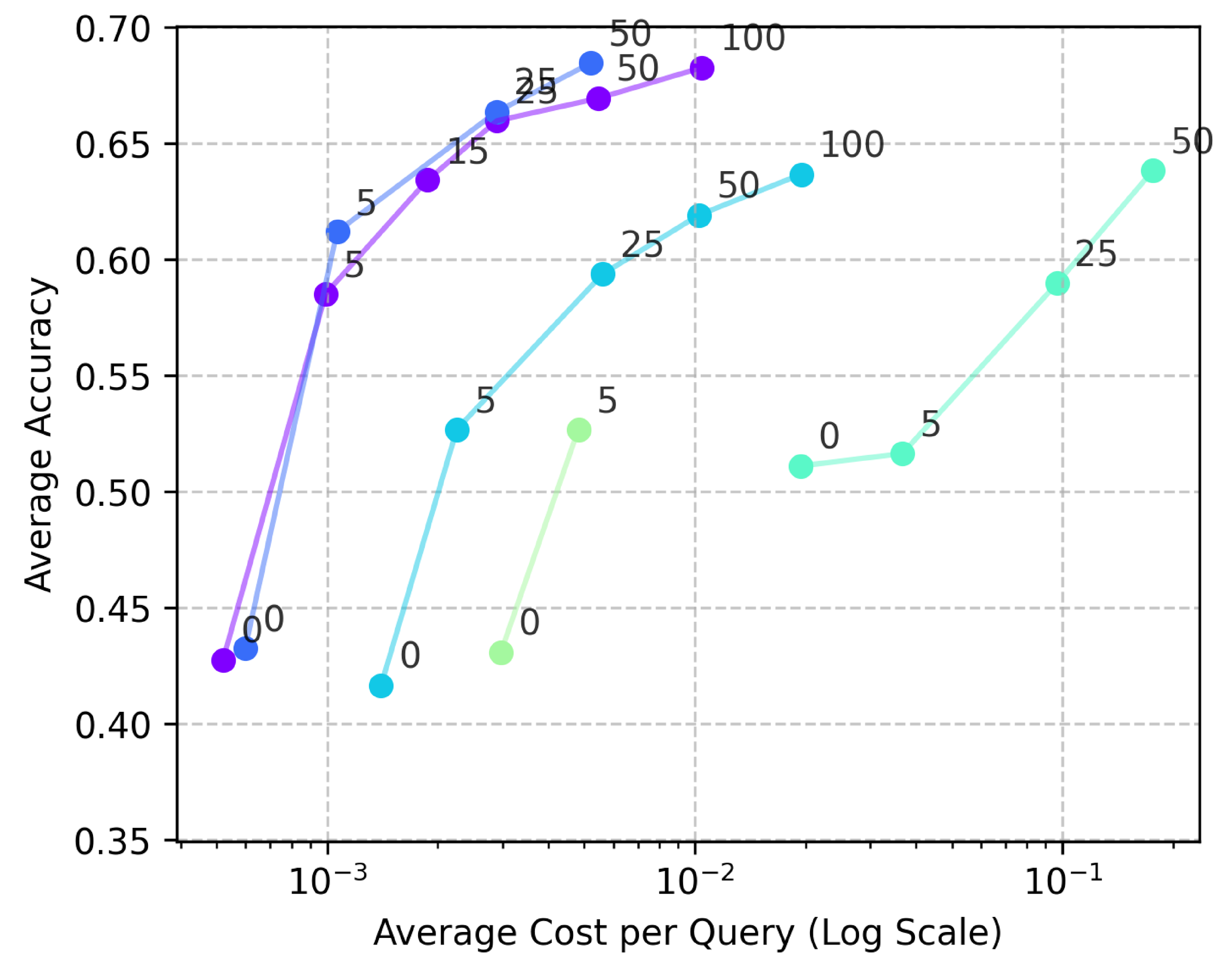

Figure 3:Model Performance (Accuracy) vs Model Cost

Findings

- Many-shot learning is very effective at reducing annotation error across most dimensions of the DQI, and in aggregate.
- The benefits of many-shot learning have diminishing marginal returns on examples after a certain point.
- Chain of Thought (CoT) has a relatively small effect, and more work is needed to understand the conditions under which CoT is sufficient.
- More expensive high-parameter models like GPT-4o do better with few in-context examples than smaller models, but as the number of examples included increases, those early gains are erased.
- GPT-4o, despite costing an order of magnitude more, was outperformed by DeepSeek Coder with just a few examples.

Next Steps

- Exploring dynamic prompt optimization to further improve model performance and adapt to new, unseen speeches.
- Investigating model distillation and fine-tuning techniques to create smaller, efficient models that can reliably perform DQI annotation at scale, potentially reducing computational resources and costs.
- Scaling up the annotation task to millions of speeches, enabling larger studies and more comprehensive analyses of deliberation quality.
- Developing open-source software tools that allow researchers to generate DQI measures for speeches in their own text corpora, fostering collaboration and reproducibility in the field.