Multi-Task Learning Improves Deep Argument Mining Performance

Amirhossein Farzam,[1] **Isaac D. Mehlhaff**,[2] Marco Morucci,[3] and Shashank Shekhar[4]

[1] Department of Computer Science, Duke University; [2] Department of Political Science, UNC-Chapel Hill; [3] Center for Data Science, NYU; [4] Department of Computer Science, NYU

THE UNIVERSITY *of* NORTH CAROLINA *at* CHAPEL HILL

Objectives

- **Understand** use of argumentation techniques and strategies in political speech and text.
- **Develop** automated tools for social scientists to analyze persuasive communication and political rhetoric.
- **Assess** the potential for multi-task learning to improve performance across tasks by recovering text representations in common semantic space.

Data

Propaganda (Da San Martino et al. 2019)

- News articles, binary sentence-level annotations of 18 propaganda types

Internet Argument Corpus (Abbott et al. 2016)

- Discussion forum posts, real-valued annotations of 8 argument characteristics

IBM-Rank-30k (Gretz et al. 2020)

- Crowd-sourced arguments, real-valued annotations of argument quality

80%-10%-10% train-validate-test split

Task	Training N	Balance
Propaganda	61,909	63/37
Disagree/Agree	66,684	21/79
Emotion/Fact	76,403	41/59
Attacking/Respectful	65,998	66/34
Nasty/Nice	65,829	73/27
Personal/Audience	24,749	25/75
Defeater/Undercutter	24,357	38/62
Negotiate/Attack	26,604	44/56
Questioning/Asserting	29,791	66/34
Argument Quality	96,036	6/94

Table 1: Size and Class Balance of Training Data.

Network Architecture

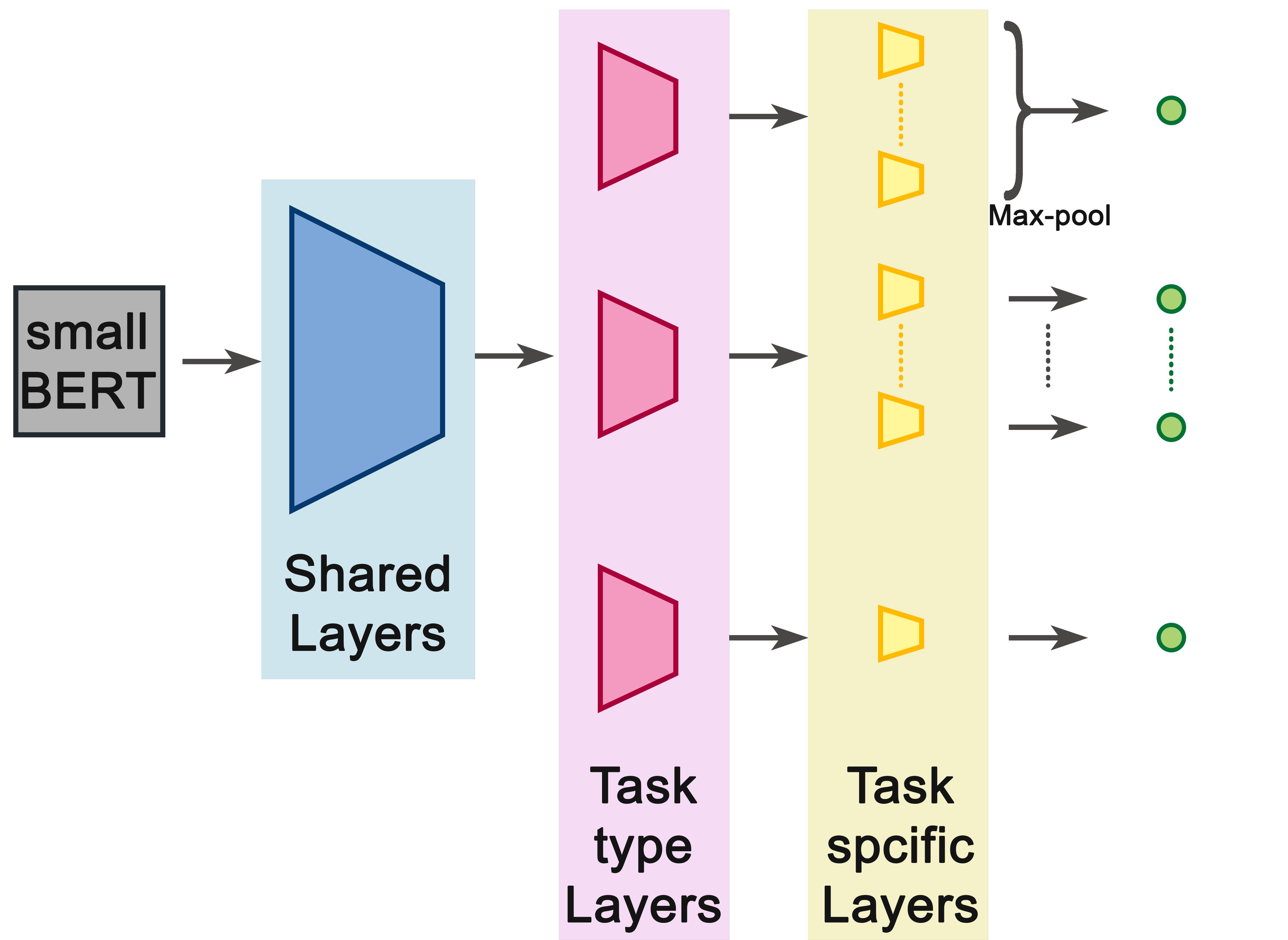

Figure 1: Network Architecture. Base encoder is fine-tuned. Max-pooling layer combines 18 propaganda labels into single binary annotation. Regularization: 0.01 weight decay rate and 40% dropout at each stage of network. Trained with AdamW optimizer (Loshchilov and Hutter 2017).

Double-Weighted Loss

Given predicted labels $\hat{y}$ and true labels y, the total loss $\mathcal{L}$ is:

$$\mathcal{L}(\hat{y}|y) = \sum_k \nu_k \mathcal{L}(\hat{y}|y, \mathcal{D}_k), \quad (1)$$

where D_k denotes the set of observations corresponding to task-type k, and $\nu_k \sim \frac{1}{|D_k|}$ are the task-type weights. The loss for each task type k is:

$$\mathcal{L}(\hat{y}|y, \mathcal{D}_k) \sim \frac{1}{|T_k|} \sum_{j \in D_k} \sum_{t \in T_k} \sum_{c \in \mathcal{C}_t} w_t^c \, l(\hat{y}_j | y_j = c), \quad (2)$$

where $l(.)$ is the binary cross-entropy loss function, T_k denotes the set of tasks within k, and $\mathcal{C}_t$ is the corresponding set of classes. Class weights w_t^c are proportional to the inverse of class enrichment.

Commonalities Across Tasks

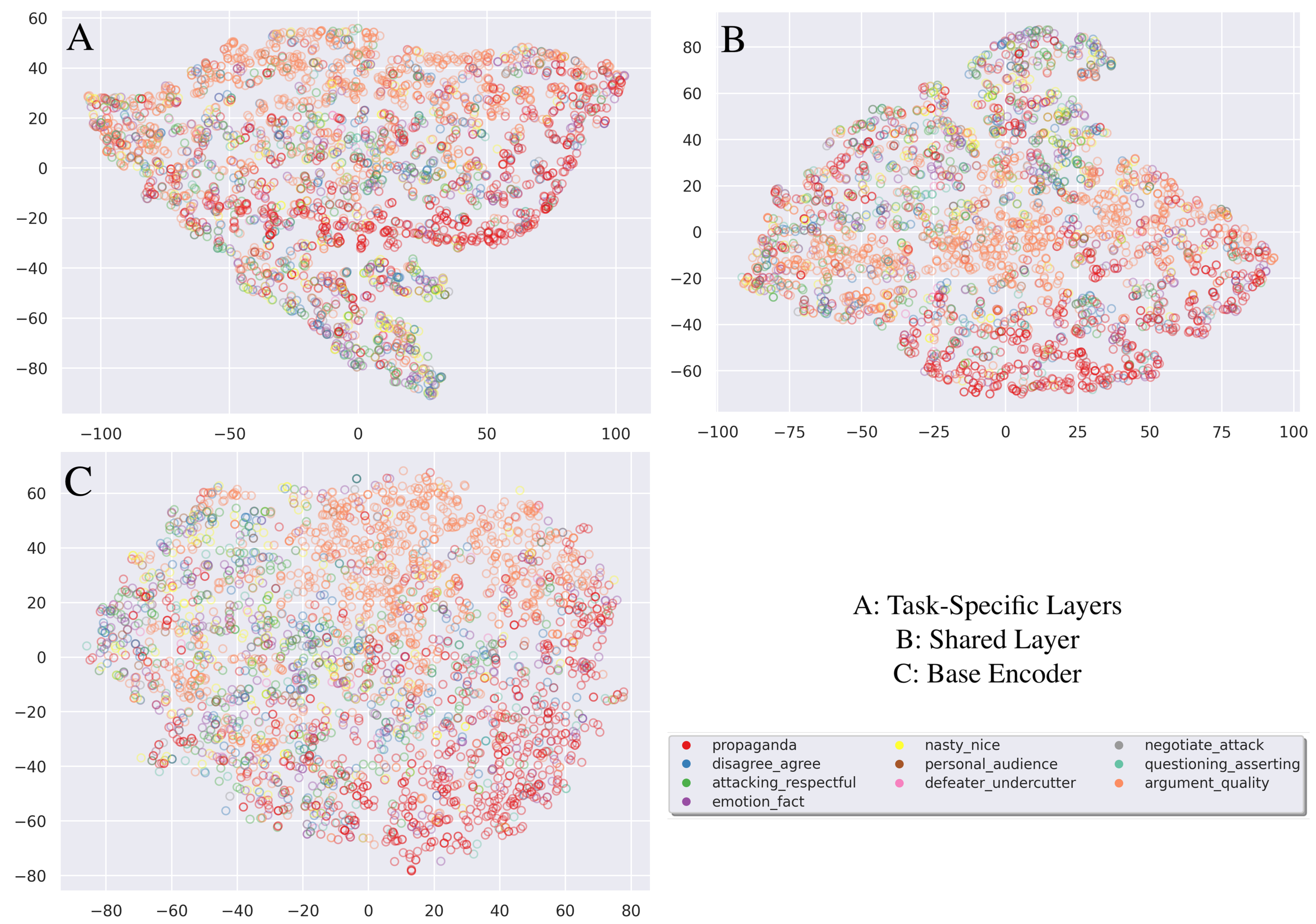

Figure 2: t-SNE projections of Text Representations from Intermediate Layers. Minor evidence of clustering suggests model is learning representations that reflect similar semantic and logical structures across tasks, without completely discarding task-specific structure. Similar amounts of clustering across plots shows common structure is preserved as network proceeds from shared to task-specific layers.

Performance Evaluation

Task	Baseline	Unigrams	Single-Task	Multi-Task
Propaganda	55.47	38.46	**63.07**	61.74
Disagree/Agree	47.29	7.49	71.15	**71.38**
Emotion/Fact	45.80	21.91	**68.11**	63.93
Attacking/Respectful	56.47	51.16	67.46	**68.07**
Nasty/Nice	59.35	61.03	66.90	**73.69**
Personal/Audience	39.90	9.23	63.25	**65.69**
Defeater/Undercutter	53.4	45.21	45.97	**55.65**
Negotiate/Attack	36.93	55.31	64.76	**64.81**
Questioning/Asserting	50.57	57.47	59.61	**63.23**
Argument Quality	76.54	0.76	**80.93**	79.17

Table 2: Weighted F1 Scores. Baseline metrics are produced by random guessing and unigram metrics by a naïve Bayes classifier. Single-task and multi-task models use small BERT as base encoder.

Metric	Baseline	Unigrams	Single-Task	Shared Trunk (Encoder)	Multi-Task (17,024)	Multi-Task (272,384)	Multi-Task (438,784)
Precision	62.26	33.65	68.85	64.73	**69.37**	69.11	68.77
Recall	52.43	44.55	64.14	55.57	65.76	63.12	**65.78**
F1	52.17	34.80	65.12	56.70	**66.73**	64.46	66.33

Table 3: Comparison of Model Sizes. Baseline metrics are produced by random guessing and unigram metrics by a naïve Bayes classifier. Number of trainable parameters in parentheses, not including base encoder. Single-task and multi-task models use small BERT as base encoder. Metrics class-weighted and averaged across tasks.

Task	Citation	Metric	Previous	New	Absolute Gain	Relative Gain
Propaganda	Da San Martino et al. (2019)	F1	60.98	61.74	0.76	1.25
Disagree/Agree	Wang and Cardie (2014)	F1	63.57	71.38	7.81	12.29
Disagree/Agree	Abbott et al. (2011)	Acc.	68.20	70.73	2.53	3.71
Emotion/Fact	Oraby et al. (2015)	F1	46.20	63.93	17.73	38.38
Nasty/Nice	Lukin and Walker (2013)	F1	69.00	73.69	4.69	6.80

Table 4: Comparison to Previous State-of-the-Art Metrics.

Computational Efficiency

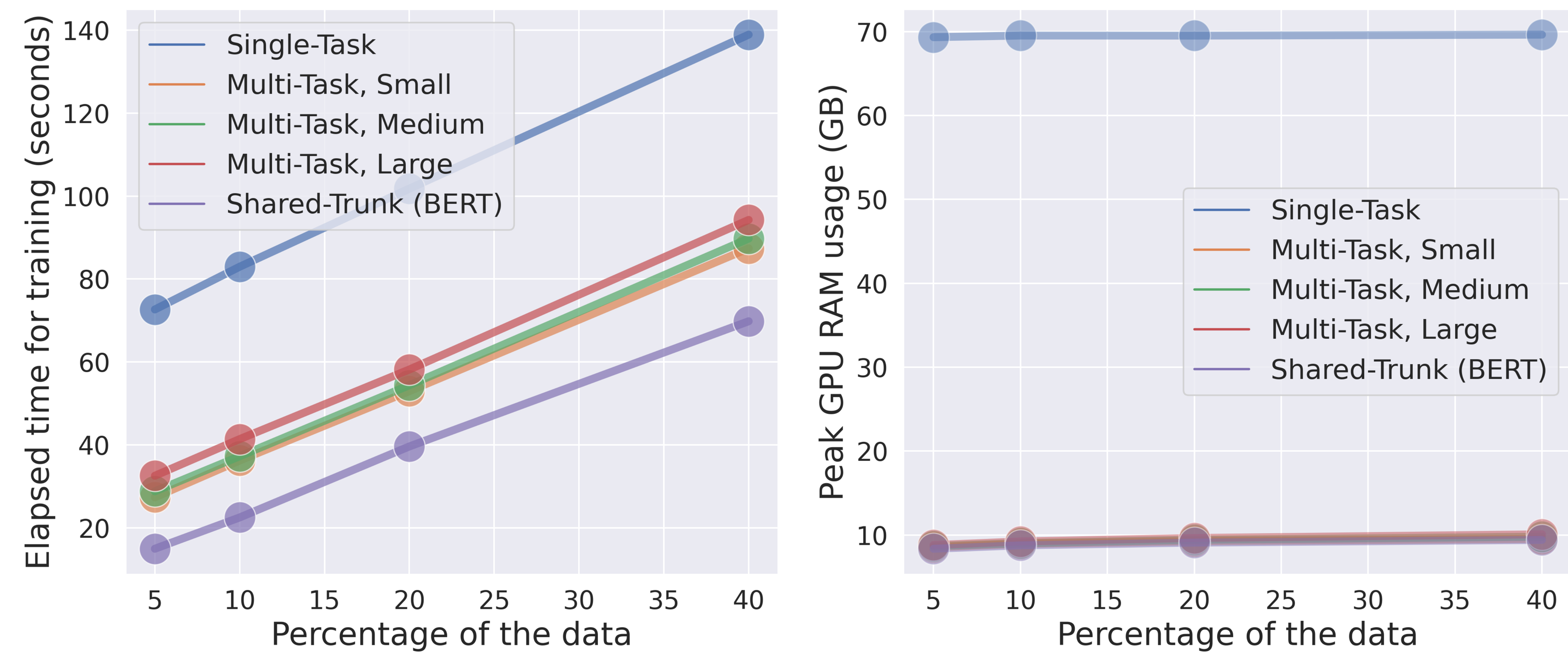

Figure 3: Computational Efficiency of Deep Learning Models. All models run on one NVIDIA A100 GPU for one epoch. Multi-task model sizes given in Table 3.

Application: `r/ChangeMyView`

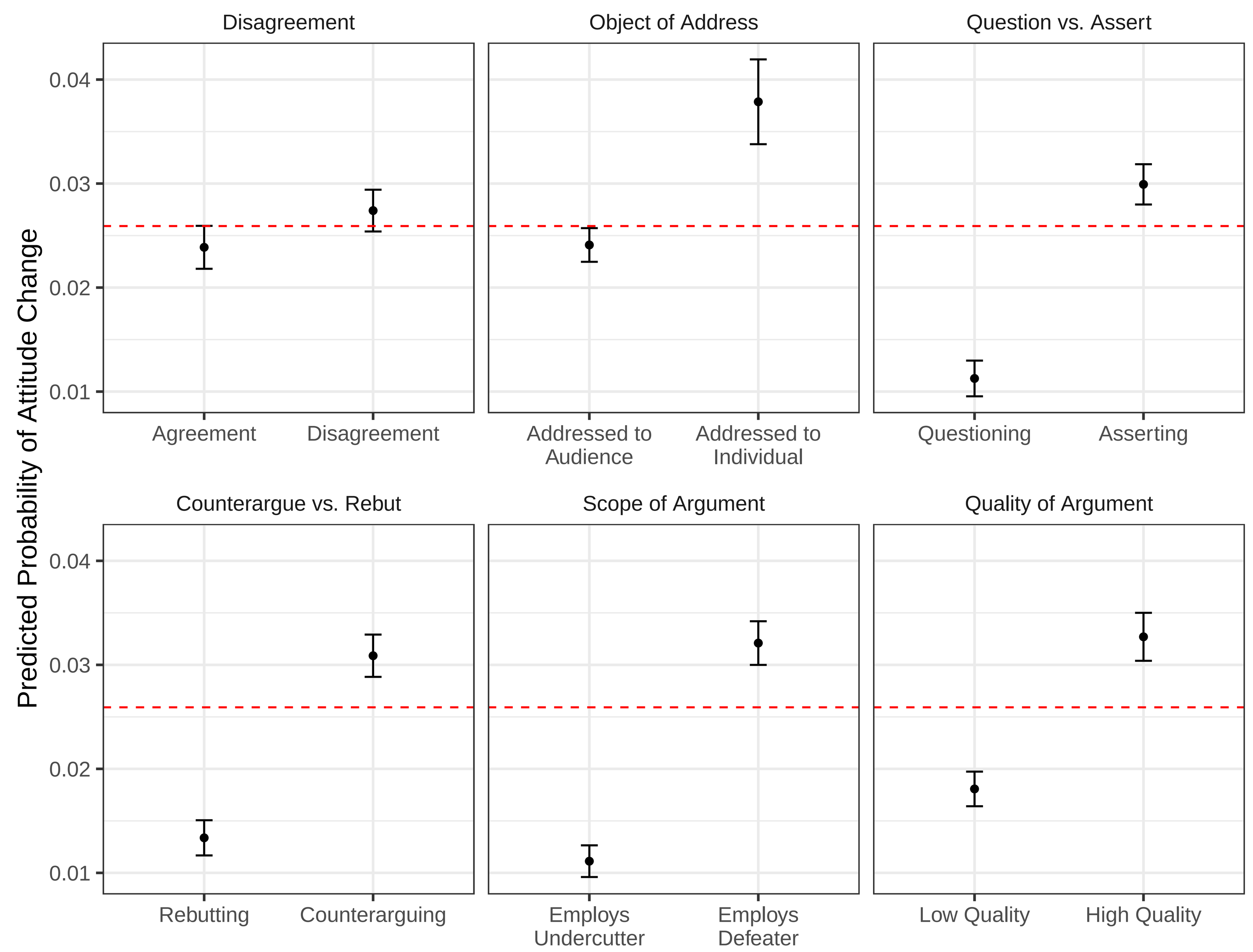

Figure 4: Effect of Select Argumentation Characteristics on Opinion Change in `r/ChangeMyView`. Red horizontal lines denote baseline probability of a comment resulting in opinion change. Error bars give 95% confidence intervals. All models are binomial logits fit with penalized maximum-likelihood (Firth 1993).

Highlights

- Argument mining tasks—and likely other natural language tasks in the social sciences—share **common semantic and logical structure.**
- **Double-branched multi-task networks with double-weighted loss** exploit shared features to drive performance across tasks.
- A multi-task approach provides **improvement on previous state-of-the-art metrics** of 1.25% to 38.38%.
- Multi-task networks enable **significant gains in computational efficiency** without sacrificing performance.
- **Network outputs correlate with opinion change** in theoretically expected ways.