Measuring the Quality of Answers in Political Q&As with Large Language Models

Online Supplementary Material

R. Michael Alvarez and Jacob Morrier

Division of the Humanities and Social Sciences

California Institute of Technology

August 2024

Contents

A Additional Figures and Tables **4**

B Detailed Description of the Network Architecture **18**

C Robustness Check: Document Length **21**

D Robustness Check: Government Backbenchers **29**

List of Figures

A1 Mean Reciprocal Rank of the Validation Set by Batch Size and Epoch 4

A2 Probability that the Correct Answer is the Closest to the Question by Cosine Similarity Between Questions and Answers by Party 6

A3 Probability that the Correct Question is the Closest to the Answer by Cosine Similarity Between Questions and Answers by Party 7

A4 Rank of the Correct Answer by Cosine Similarity Between Questions and Answers by Party 8

A5 Rank of the Correct Question by Cosine Similarity Between Questions and Answers by Party 9

A6 Probability that the Correct Answer is the Closest to the Question by Cosine Similarity Between Questions and Answers by Legislature 10

A7 Probability that the Correct Question is the Closest to the Answer by Cosine Similarity Between Questions and Answers by Legislature 11

A8 Rank of the Correct Answer by Cosine Similarity Between Questions and Answers by Legislature 12

A9 Rank of the Correct Question by Cosine Similarity Between Questions and Answers by Legislature 13

A10 Average Cosine Similarity Between Questions and Answers by Number of Seats . . 14

A11 Average Cosine Similarity Between Questions and Answers by Share of Opposition Seats 15

A12 Monthly Evolution of the Cosine Similarity Between Questions and Answers by Party 16

A13 Architecture of Sentence-BERT Encoders 20

A14 Average Cosine Similarity Between Questions and Answers by Question Length . . 23

A15 Average Cosine Similarity Between Questions and Answers by Answer Length . . 24

A16 Average Question Length by Party and Legislature 25

A17 Average Answer Length by Party and Legislature 25
A18 Average Cosine Similarity Between Questions and Answers by Average Question Length 26
A19 Average Cosine Similarity Between Questions and Answers by Average Answer Length 27
A20 Average Cosine Similarity Between Questions and Answers by Party and Legislature (After Controlling for Length of Questions and Answers) 28
A21 Distribution of the Cosine Similarity Between Questions and Answers 30
A22 Average Cosine Similarity Between Questions and Answers by Party and Legislature 31
A23 Average Cosine Similarity Between Questions and Answers by Party and Portfolio 32

List of Tables

A1 Training Hyperparameters 4
A2 Model Accuracy on the Inference Set 5
A3 Descriptive Statistics of the Distribution of the Cosine Similarity Between Questions and Answers 5
A4 Prompt for Generating Topic Labels 17

A Additional Figures and Tables

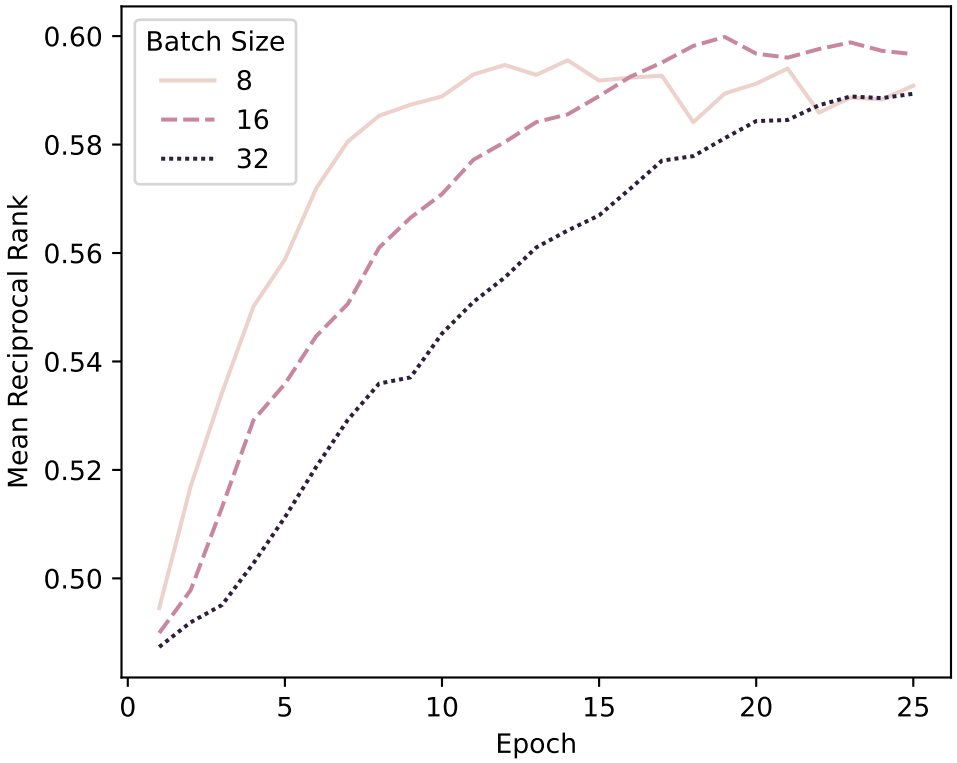


Figure A1: Mean Reciprocal Rank of the Validation Set by Batch Size and Epoch

Table A1: Training Hyperparameters

Model	`multi-qa-mpnet-base-cos-v1`
Loss Function	Multiple Negatives Ranking Loss
Epochs	10
Batch Size	8
Optimizer	AdamW†
Learning Rate	$2 \times 10^{-5\dagger}$
Learning Rate Scheduler	Warm-up Linear†
Warm-up Steps	10,000†
Weight Decay	0.01†
Maximum Gradient Norm	1†

†Default Value

Table A2: Model Accuracy on the Inference Set

	@ 10	@ 25	@ 100
Precision	0.0256	0.0144	0.0054
Recall	0.2561	0.3588	0.5497
F-1 Score	0.0232	0.0138	0.0054

Table A3: Descriptive Statistics of the Distribution of the Cosine Similarity Between Questions and Answers

N	54,914
Mean	0.5387
Standard Deviation	0.1865
Minimum	−0.1625
First Quartile	0.4178
Median	0.5608
Third Quartile	0.6806
Maximum	0.9542

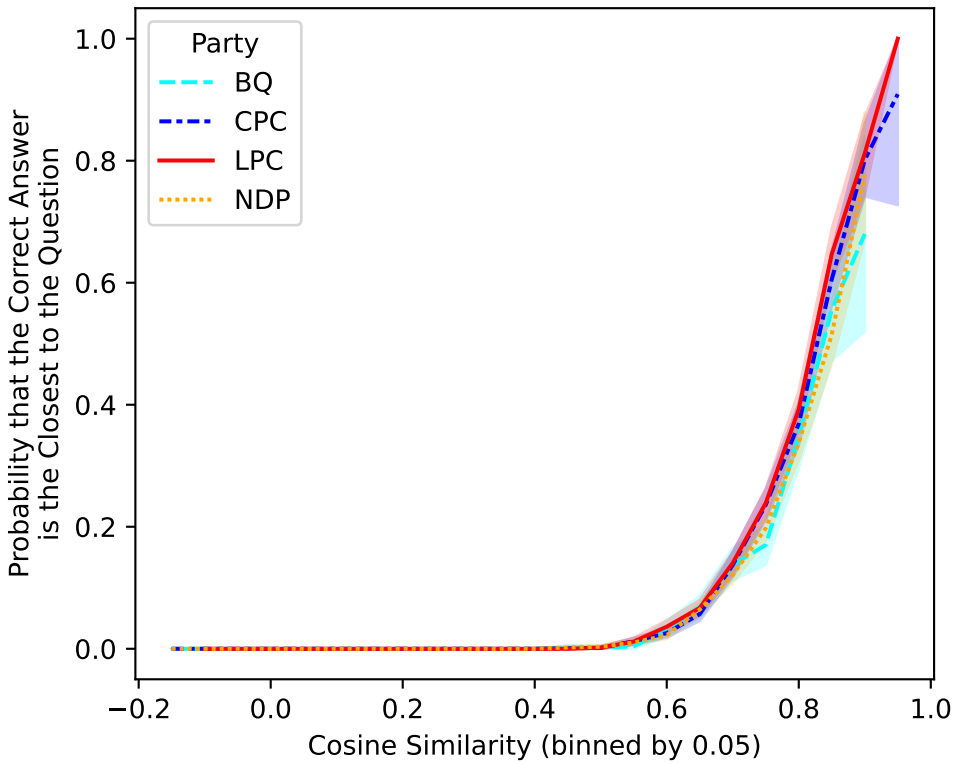


Figure A2: Probability that the Correct Answer is the Closest to the Question by Cosine Similarity Between Questions and Answers by Party

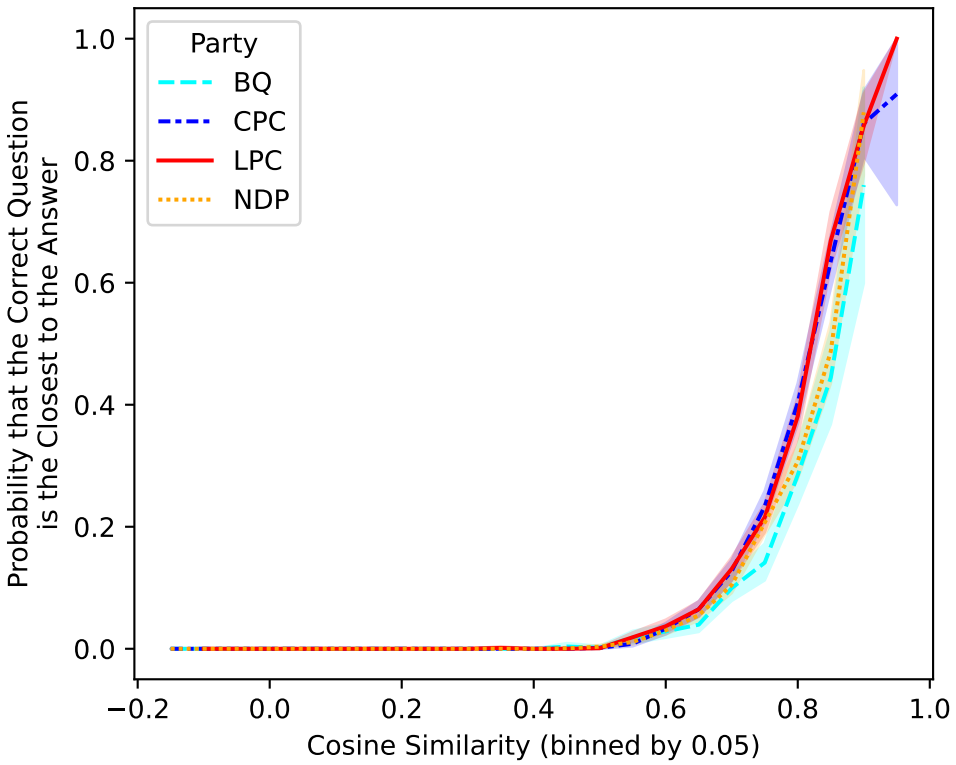


Figure A3: Probability that the Correct Question is the Closest to the Answer by Cosine Similarity Between Questions and Answers by Party

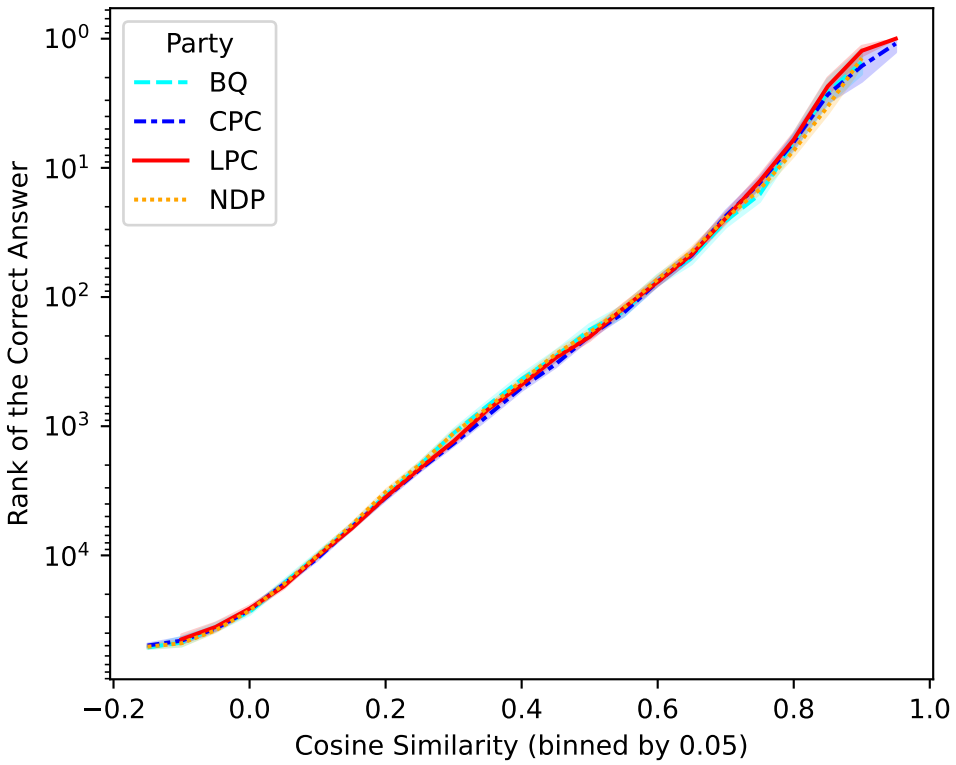


Figure A4: Rank of the Correct Answer by Cosine Similarity Between Questions and Answers by Party

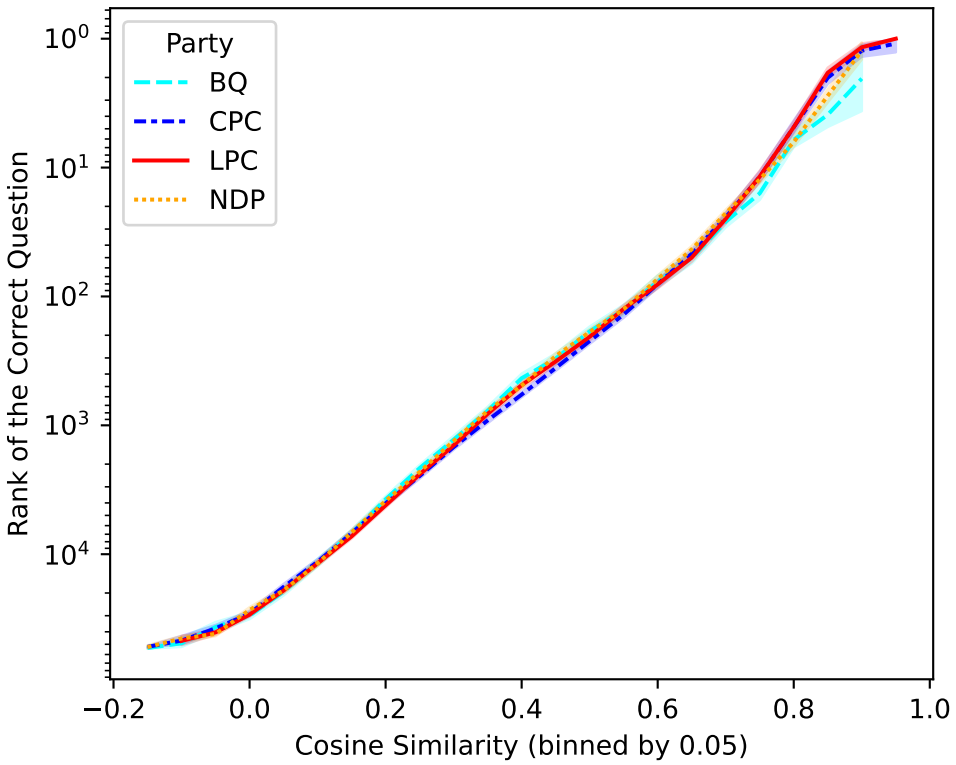


Figure A5: Rank of the Correct Question by Cosine Similarity Between Questions and Answers by Party

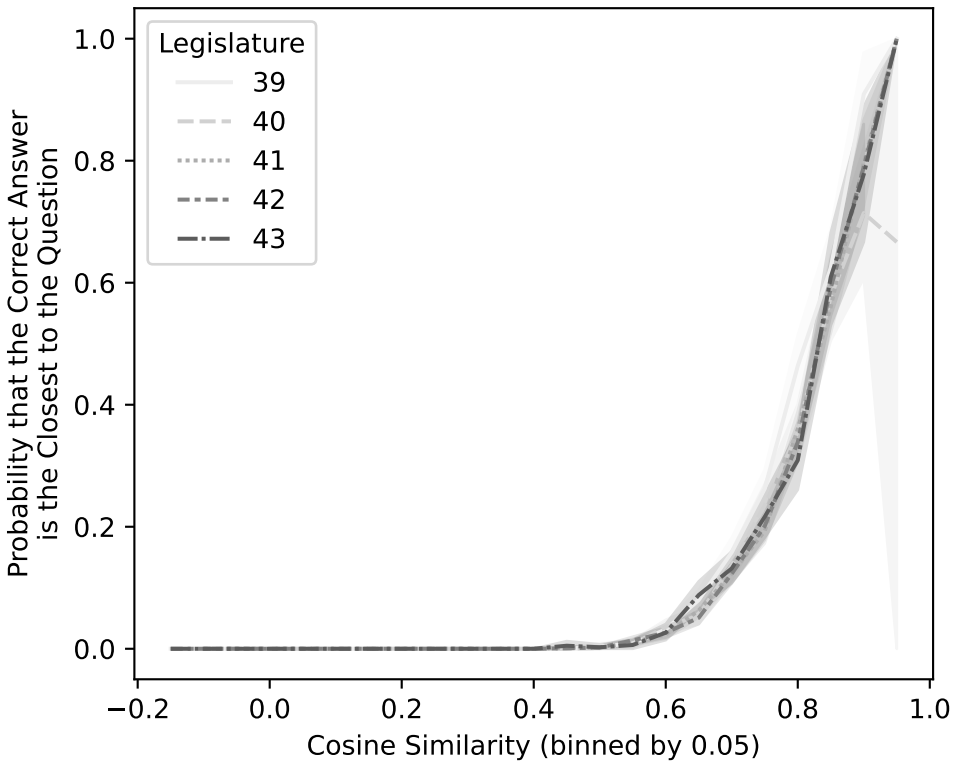


Figure A6: Probability that the Correct Answer is the Closest to the Question by Cosine Similarity Between Questions and Answers by Legislature

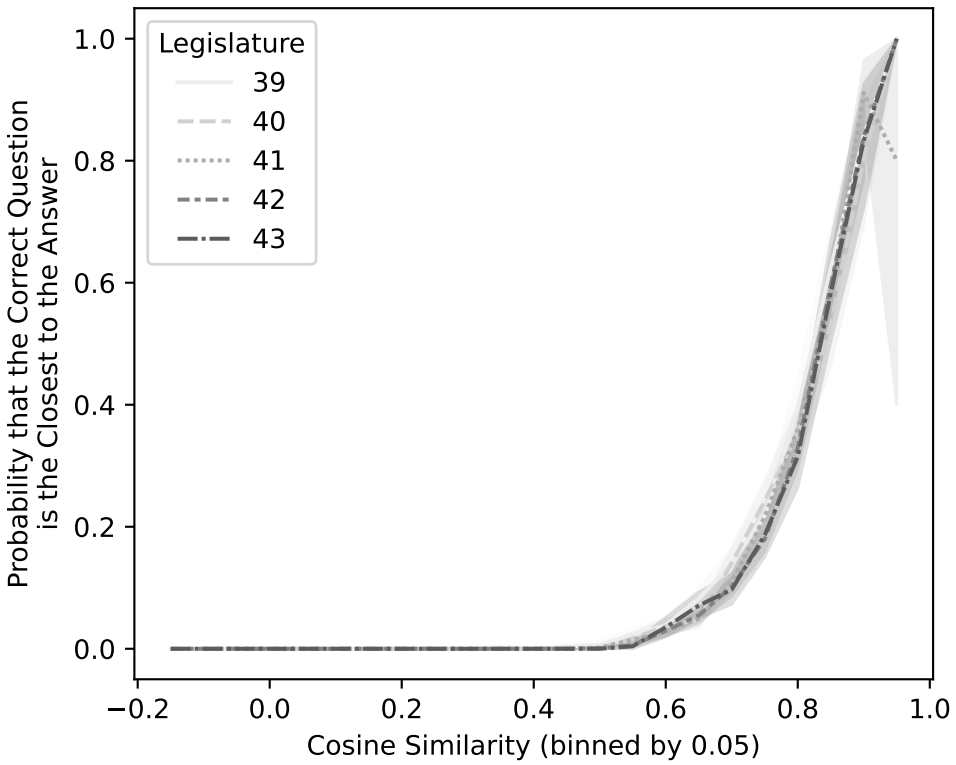


Figure A7: Probability that the Correct Question is the Closest to the Answer by Cosine Similarity Between Questions and Answers by Legislature

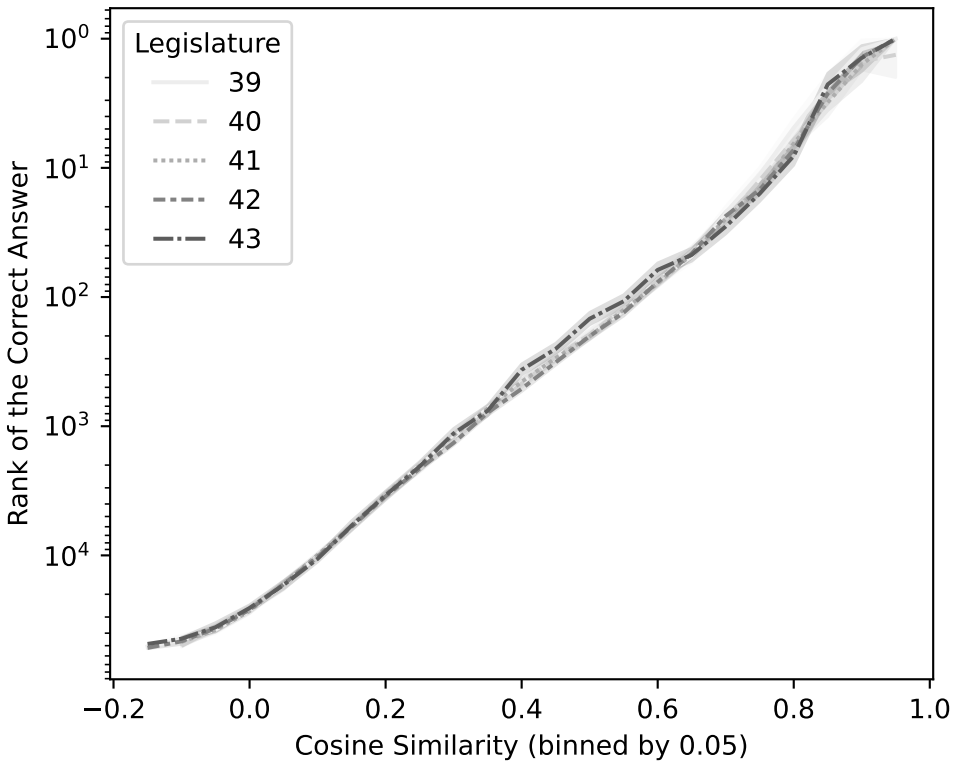


Figure A8: Rank of the Correct Answer by Cosine Similarity Between Questions and Answers by Legislature

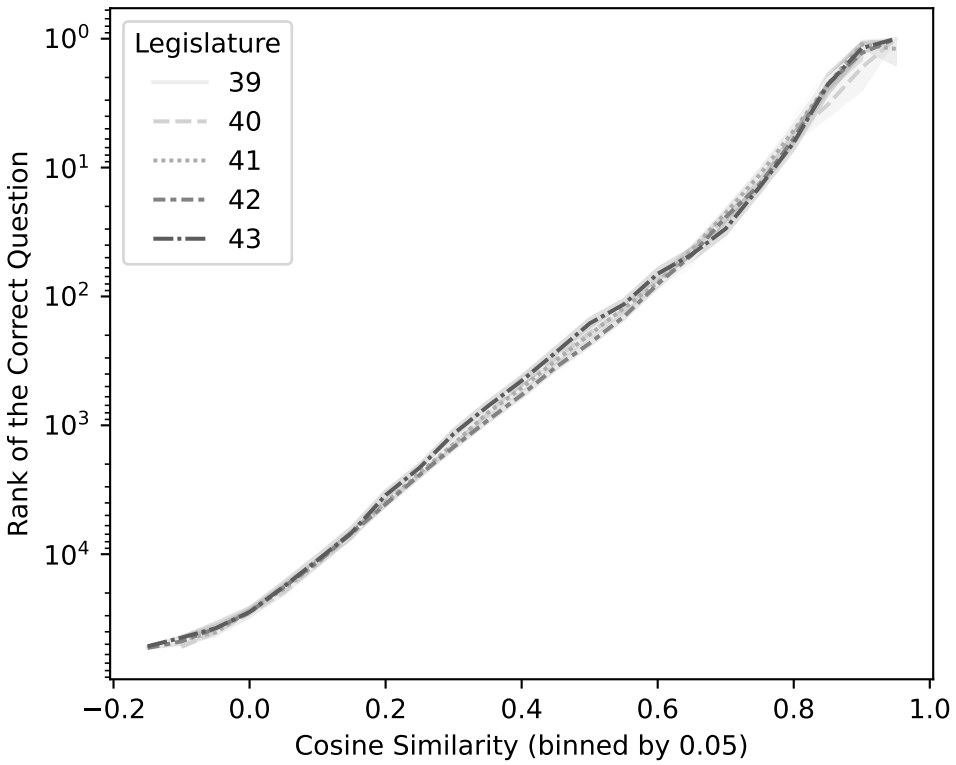
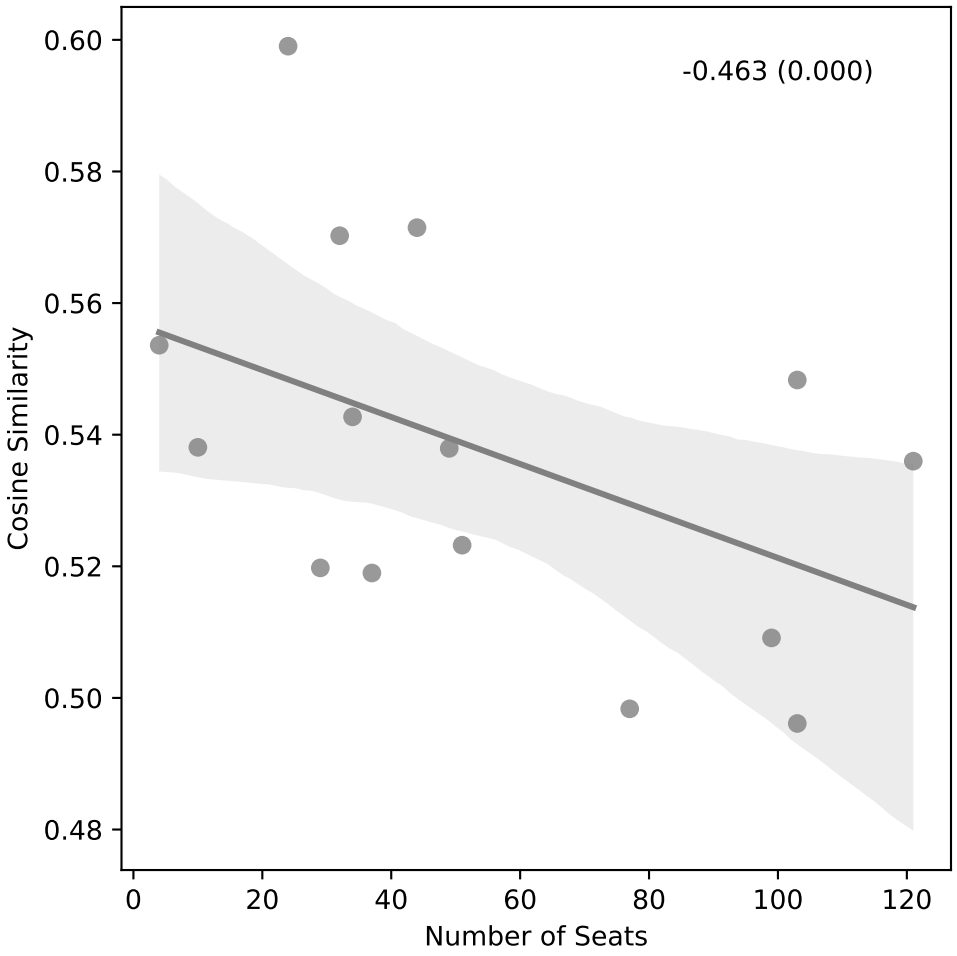


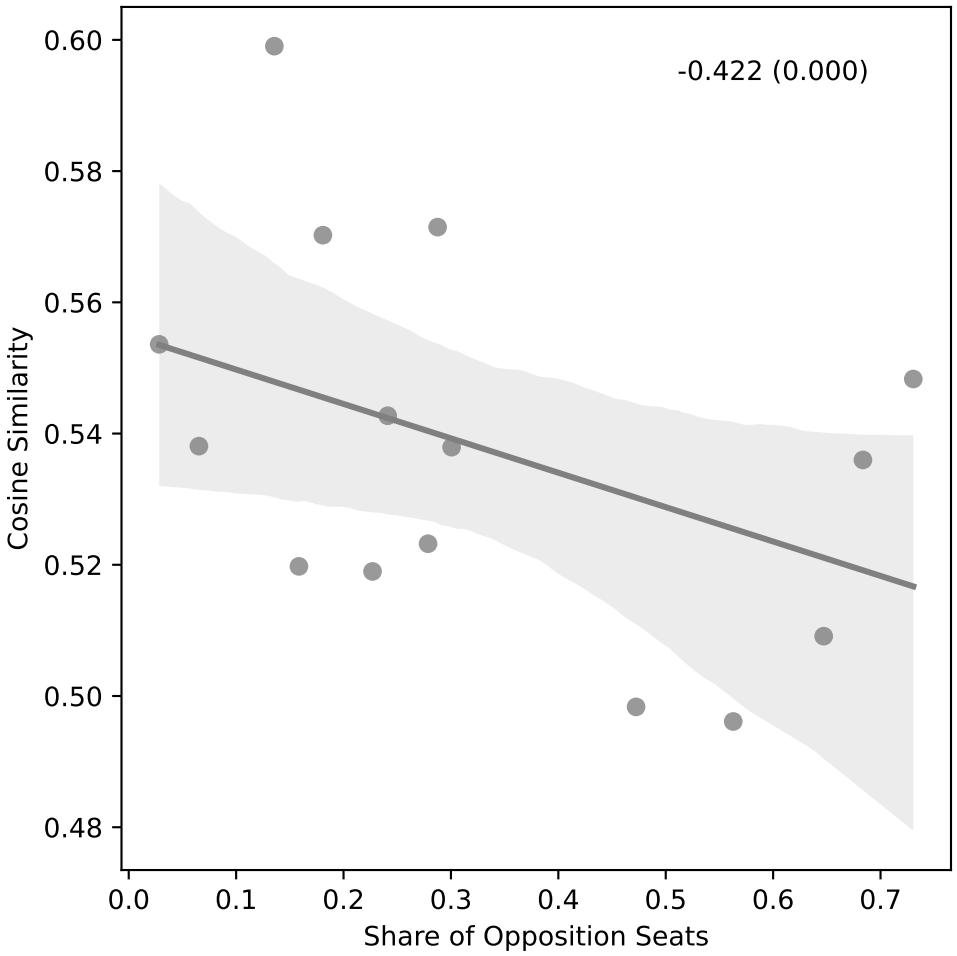
Figure A9: Rank of the Correct Question by Cosine Similarity Between Questions and Answers by Legislature



Notes:

1. The seat count reflects each party's representation at the start of the legislature.
2. The correlation coefficient and corresponding p-value are shown in the top right corner.

Figure A10: Average Cosine Similarity Between Questions and Answers by Number of Seats



Notes:

1. The seat count reflects each party's representation at the start of the legislature.
2. The correlation coefficient and corresponding p-value are shown in the top right corner.

Figure A11: Average Cosine Similarity Between Questions and Answers by Share of Opposition Seats

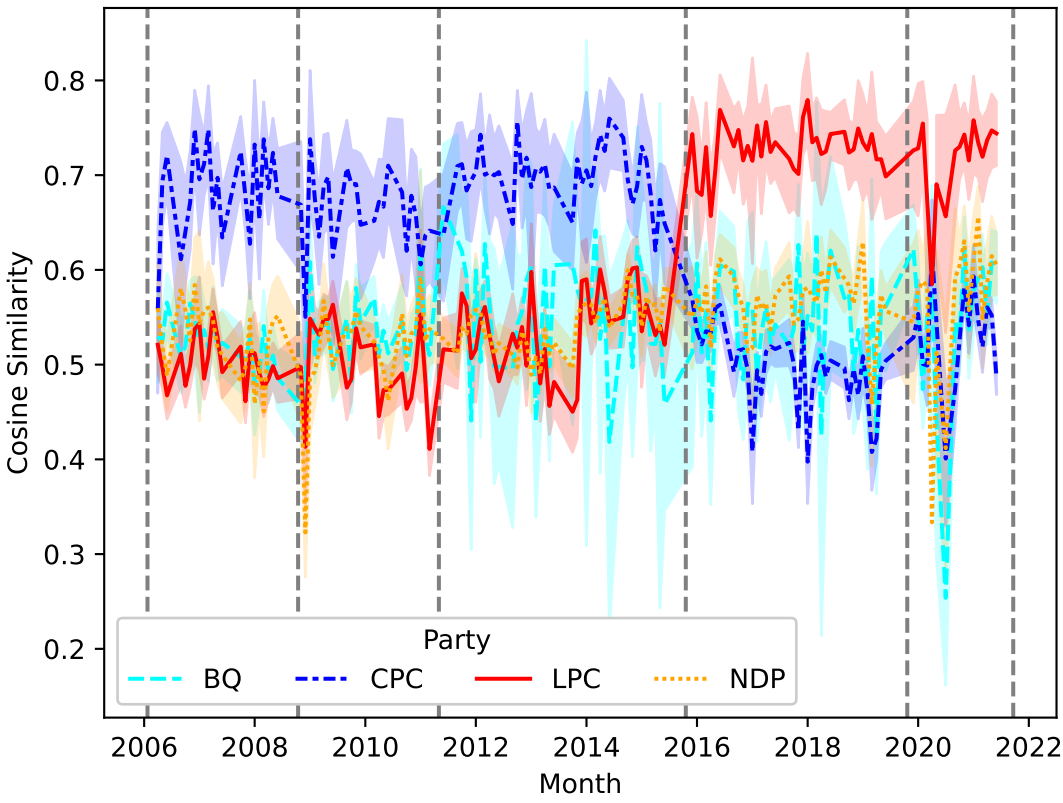


Figure A12: Monthly Evolution of the Cosine Similarity Between Questions and Answers by Party

System Prompt: You are a helpful, honest, and respectful assistant.

Your task is to label topics clustering questions asked by members of Parliament to Cabinet ministers during the Question Period in the Canadian House of Commons.

You must meticulously follow all the instructions you are given.

Example Prompt: I have a topic that contains the following documents:

- Traditional diets in most cultures were primarily plant-based with a little meat on top, but with the rise of industrial-style meat production and factory farming, meat has become a staple food.
- Meat, but especially beef, is the word food in terms of emissions.
- Eating meat doesn't make you a bad person, not eating meat doesn't make you a good one.

The topic is described by the following keywords: meat, beef, eat, eating, emissions, steak, food, health, processed, chicken.

Please devise a short label for this topic. I want this label to reflect the policy issue the questions are about, irrespective of their underlying sentiment.

Please capitalize this label according to standard rules for the capitalization of titles. Make sure to return only the label without additional notes.

Example Output: Environmental Impacts of Meat Consumption

Main Prompt: I have a topic that contains the following documents:

[DOCUMENTS]

The topic is represented by the following keywords: [KEYWORDS].

Please devise a short label for this topic. I want this label to reflect the policy issue the questions are about, irrespective of their underlying sentiment.

Please capitalize this label according to standard rules for the capitalization of titles. Make sure to return only the label without additional notes.

Table A4: Prompt for Generating Topic Labels

B Detailed Description of the Network Architecture

We detail each component of our artificial neural network, starting with the output and moving towards the input. To begin, different metrics can be used to measure the distance between adjacency pairs' embeddings. We use the cosine similarity, a measure reflecting the angle between two numerical vectors $\boldsymbol{x}$ and $\boldsymbol{y} \in \mathbb{R}^n$:

$$\cos(\boldsymbol{x}, \boldsymbol{y}) = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\| \, \|\boldsymbol{y}\|}.$$

By construction, the cosine similarity belongs to the interval $[-1, 1]$, with two parallel vectors having a cosine similarity of 1, two orthogonal vectors a cosine similarity of 0, and two opposite vectors a cosine similarity of -1.

The embeddings of questions and answers are derived from a variant of BERT called "Sentence-BERT" (Devlin et al. 2019; Reimers and Gurevych 2019).[1] The architecture of a Sentence-BERT encoder is schematically illustrated in Figure A13, in which the branch processing Token 1 is highlighted, and the other branches are faded. An encoder takes a sentence or short paragraph as input or, formally, an ordered sequence of strings of characters, called tokens, representing words or parts of words. Each token is associated with a numerical vector. Token-level embeddings are added to positional embeddings, reflecting each token's relative location within the input sequence. The resulting numerical vectors are then passed through multiple identical layers, each consisting of a multi-head self-attention mechanism and a feed-forward component. The multi-head self-attention mechanism is the central component of this architecture. A self-attention head considers both the embedding of the token being processed and the embeddings of the surrounding tokens. It allows BERT to develop a contextual understanding of each token and, especially, to recognize what in the context is pertinent to its meaning. Concretely, the self-attention head outputs a weighted sum of all the input embeddings computed according to adjustable weights. In each layer, multiple self-attention heads operate in parallel, justifying the term "multi-head self-attention." The output of the

1. We encourage those interested in a more formal treatment to consult Ekman (2022), Murphy (2022), or Zhang et al. (2023).

multi-head self-attention, along with the initial embedding, is fed into a fully connected layer that applies a linear transformation followed by a non-linear activation function. In the end, the BERT encoder outputs for each input token a numerical vector encapsulating a complex combination of syntactic, semantic, and world knowledge (Rogers, Kovaleva, and Rumshisky 2020). Sentence embeddings are computed by pooling—for instance, by averaging—these vectors.

We adopt the Multiple Negatives Ranking Loss function as our training objective (Henderson et al. 2017). For a batch of size K, we consider K questions, represented by the embeddings $\mathbf{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_K)$, and their answers, represented by the embeddings $\mathbf{Y} = (\boldsymbol{y}_1, ..., \boldsymbol{y}_K)$. The embeddings $\mathbf{X}$ and $\mathbf{Y}$ are functions of the parameters that generate the word embeddings, which are the parameters we seek to optimize. For each question $\boldsymbol{x}_i$, every answer $\boldsymbol{y}_j$ with $j \neq i$ acts as a negative candidate. The probability that the correct answer is observed is approximated using the correct answer and the negative candidates within the batch. Specifically, we approximate the probability distribution with a logistic regression model that uses the cosine similarity between the question and candidate answer embeddings as input features:

$$\mathbb{P}\left(\boldsymbol{y}_i \mid \boldsymbol{x}_i\right) = \frac{\exp\left(S\left(\boldsymbol{x}_i, \boldsymbol{y}_i\right)\right)}{\sum_{j=1}^{K} \exp\left(S\left(\boldsymbol{x}_i, \boldsymbol{y}_j\right)\right)},$$

where S is a linear function of the cosine similarity. Therefore, the objective is to minimize the following loss function for each batch:

$$\begin{aligned}
\mathcal{J}\left(\mathbf{X}, \mathbf{Y}\right) &= -\frac{1}{K} \sum_{i=1}^{K} \log\left(\mathbb{P}\left(\boldsymbol{y}_i \mid \boldsymbol{x}_i\right)\right) \\
&= -\frac{1}{K} \sum_{i=1}^{K} \log\left(\frac{\exp\left(S\left(\boldsymbol{x}_i, \boldsymbol{y}_i\right)\right)}{\sum_{j=1}^{K} \exp\left(S\left(\boldsymbol{x}_i, \boldsymbol{y}_j\right)\right)}\right) \\
&= -\frac{1}{K} \sum_{i=1}^{K} \left[S\left(\boldsymbol{x}_i, \boldsymbol{y}_i\right) - \log\left(\sum_{j=1}^{K} \exp\left(S\left(\boldsymbol{x}_i, \boldsymbol{y}_j\right)\right)\right)\right].
\end{aligned}$$

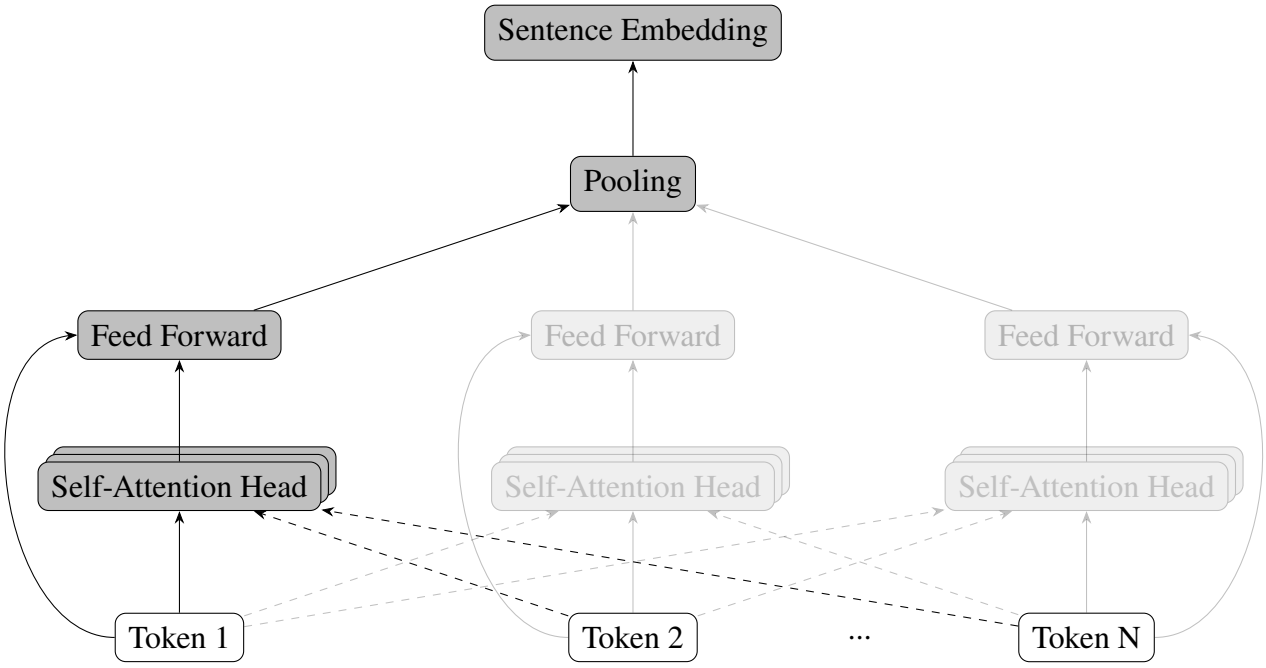


Figure A13: Architecture of Sentence-BERT Encoders

C Robustness Check: Document Length

A potential problem with using distance metrics between estimated latent representations as a measurement is that sampling errors mechanically inflate their distance and lower their similarity. While this issue affects all latent representations and distance metrics, it is particularly pronounced when dealing with high-dimensional representations, as we do in this article. This has been carefully explored and documented in previous literature (Gentzkow, Shapiro, and Taddy 2019; Loon et al. 2022; Green et al. 2024).

This problem may manifest in our results through a systematic relationship between the length of questions and answers and their cosine similarity. The underlying intuition is that the latent representations of shorter questions or answers have larger sampling errors since they are estimated with less information. This results in a downward bias of the cosine similarity for shorter questions or answers. If there are systematic differences in the length of questions and answers across parties and legislatures, this bias could spread to our substantive findings.

Figures A14 and A15 depict the average cosine similarity as a function of the length of questions and answers. They confirm a statistically significant relationship between cosine similarity and the lengths of both questions and answers. Notably, this relationship is downward-sloping for question length, meaning that longer questions are associated with a lower cosine similarity. This contradicts our expectation if sampling error had introduced a significant bias in the cosine similarity. Conversely, the relationship is upward-sloping for answer length, suggesting that either longer answers have a lower sampling error, longer answers are more relevant to the initial questions, or both.

The potential downward bias in cosine similarity could affect our substantive findings regarding the relationship between answer quality and the party affiliation of the member of Parliament asking the question, but only if there are systematic differences in question and answer lengths across the latter. Figures A16 and A17 reveal systematic variations in the lengths of questions and answers based on the party affiliation of the member of Parliament asking the question and the legislature. Furthermore, Figures A18 and A19 indicate an apparent relationship between estimates

of the average cosine similarity and the lengths of questions and answers, depending on the party affiliation of the member of Parliament asking the question and the legislature. This suggests that our substantive findings might be driven, at least partly, by systematic differences in the lengths of questions and answers. This could be symptomatic of a downward bias in cosine similarity resulting from sampling error.

To mitigate and evaluate the resilience of our substantive findings against any systematic relationship between cosine similarity and the lengths of questions and answers, we calculate the average cosine similarity between questions and answers, conditional on the party affiliation of the member of Parliament asking the question and the legislature, after controlling for question and answer lengths. Formally, adjusted average cosine similarity estimates are obtained from a linear regression model that includes question and answer lengths and party-legislature fixed effects as covariates. Predictions are calculated for the average question and answer lengths in our inference dataset. Thus, they represent what would have been the average cosine similarity if question and answer lengths were the same for all these groups.

Figure A20 shows the estimated average cosine similarity between questions and answers, broken down by party and legislature, after accounting for the length of both questions and answers. The pattern aligns with the substantive findings discussed in the main text, confirming that our core conclusions are robust and not influenced by systematic variations in the lengths of questions and answers.

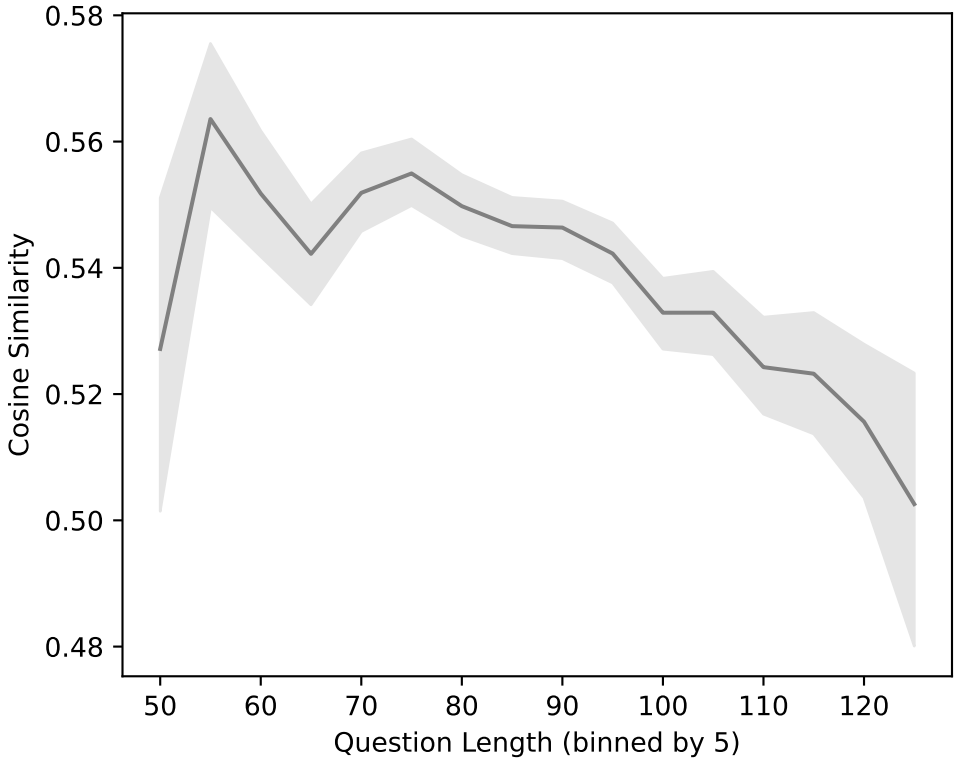


Figure A14: Average Cosine Similarity Between Questions and Answers by Question Length

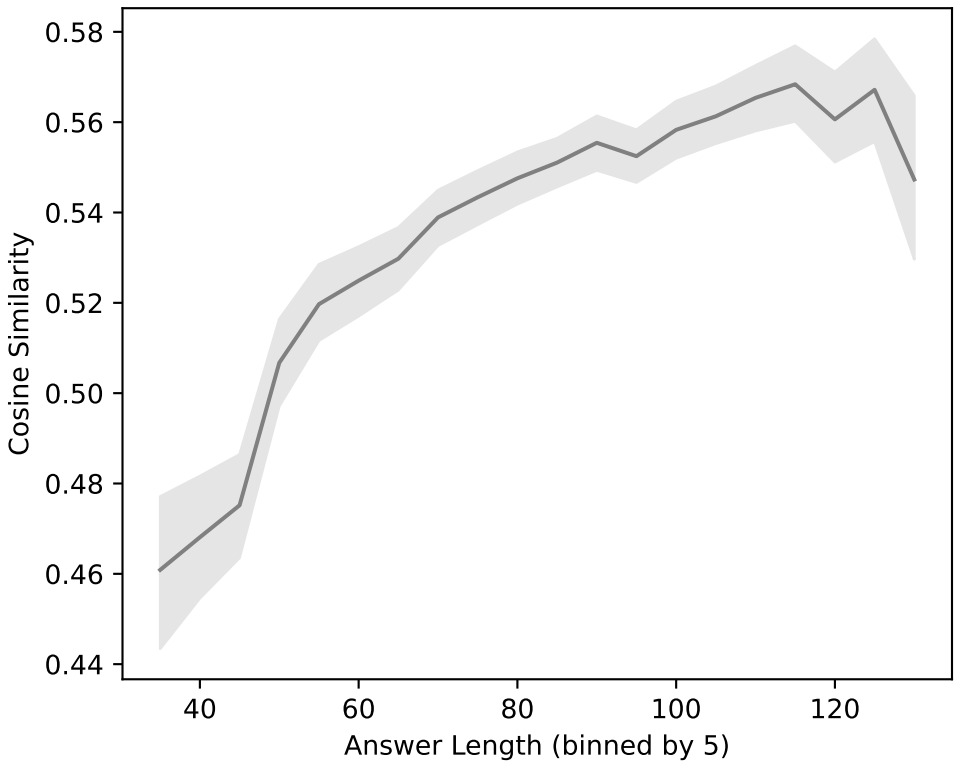


Figure A15: Average Cosine Similarity Between Questions and Answers by Answer Length

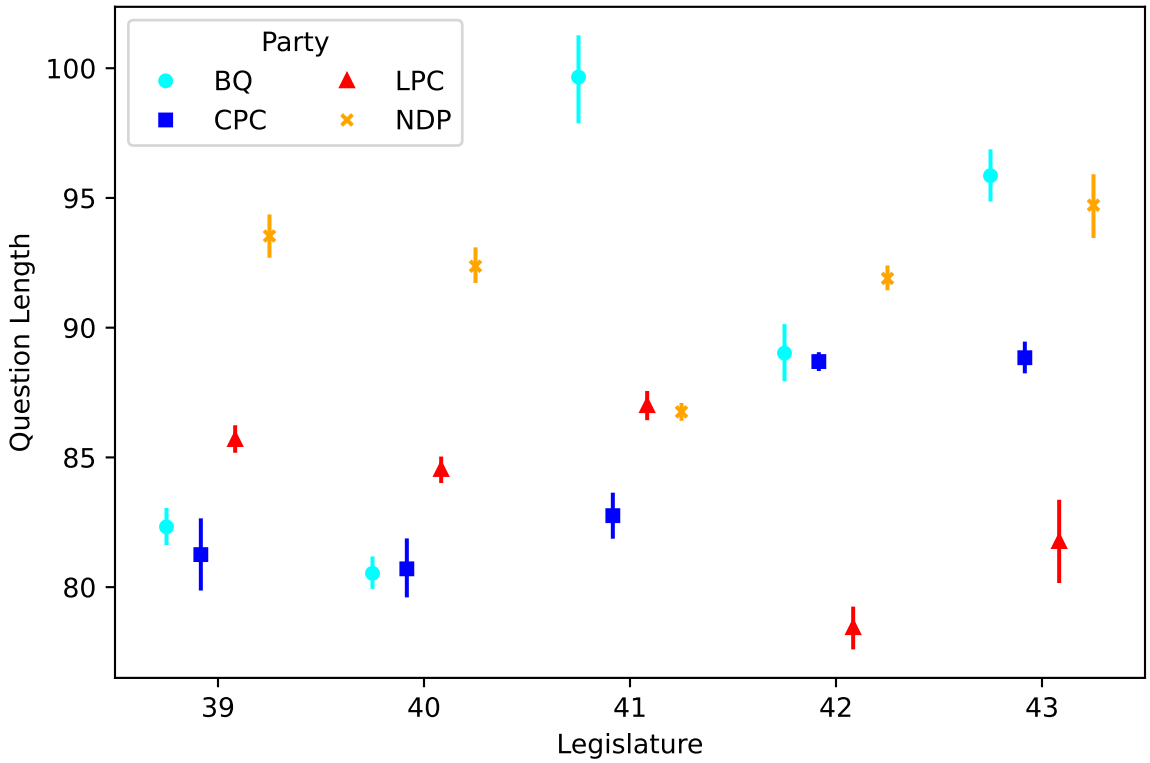


Figure A16: Average Question Length by Party and Legislature

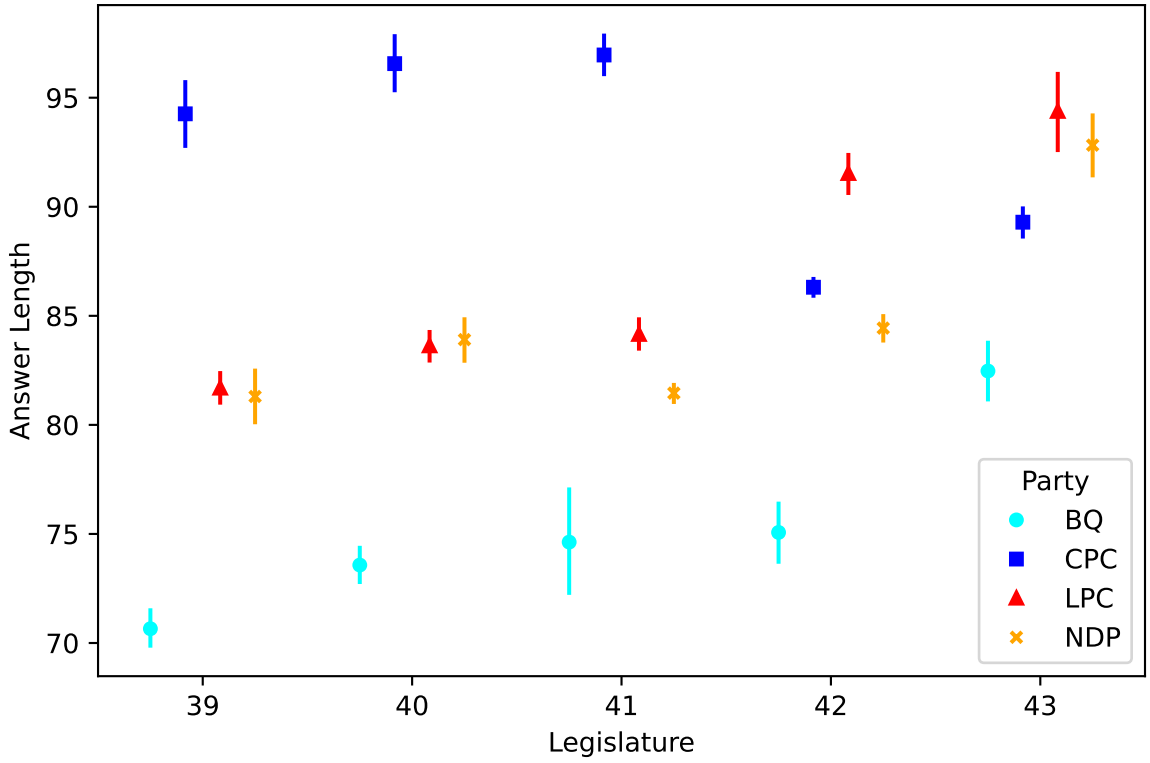


Figure A17: Average Answer Length by Party and Legislature

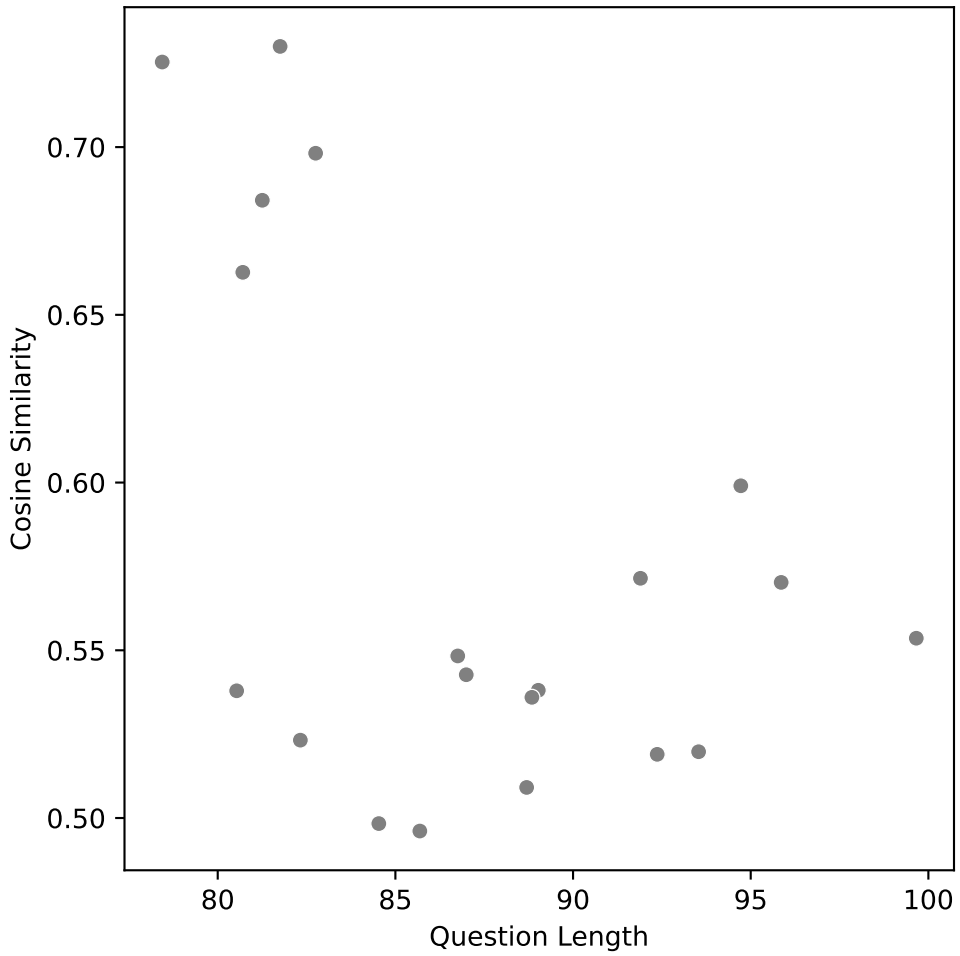


Figure A18: Average Cosine Similarity Between Questions and Answers by Average Question Length

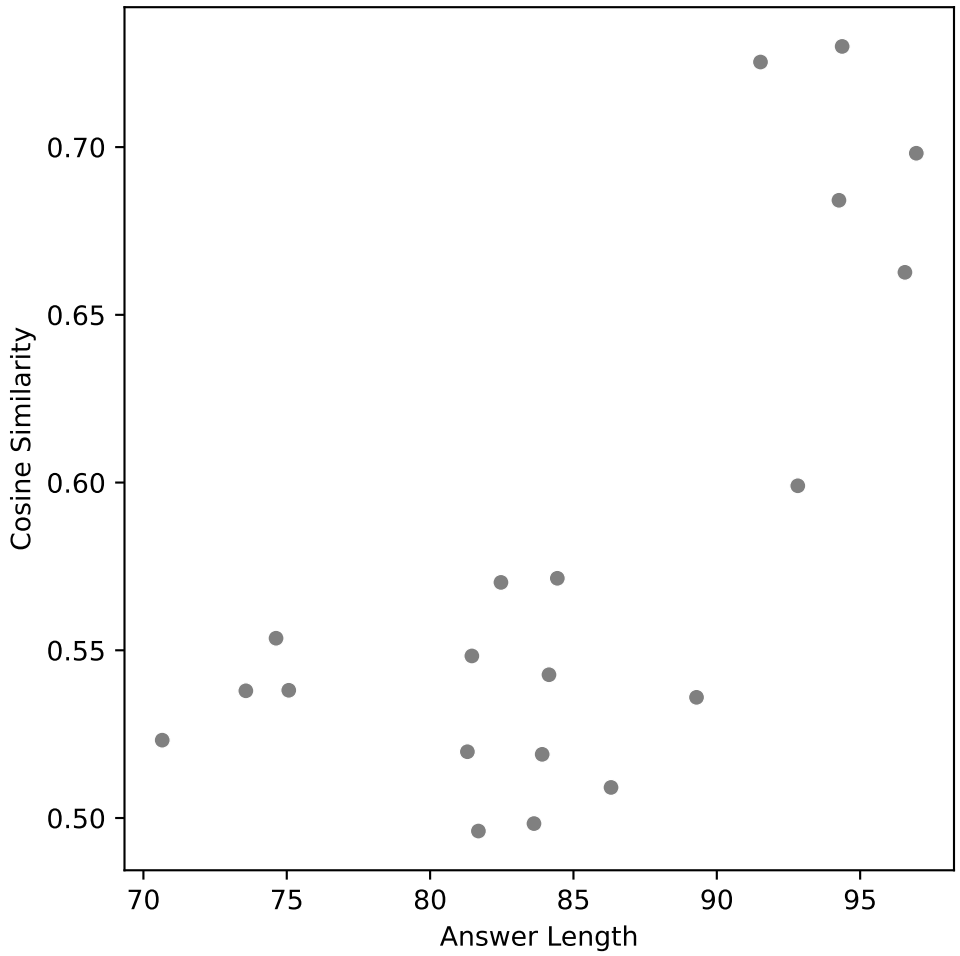


Figure A19: Average Cosine Similarity Between Questions and Answers by Average Answer Length

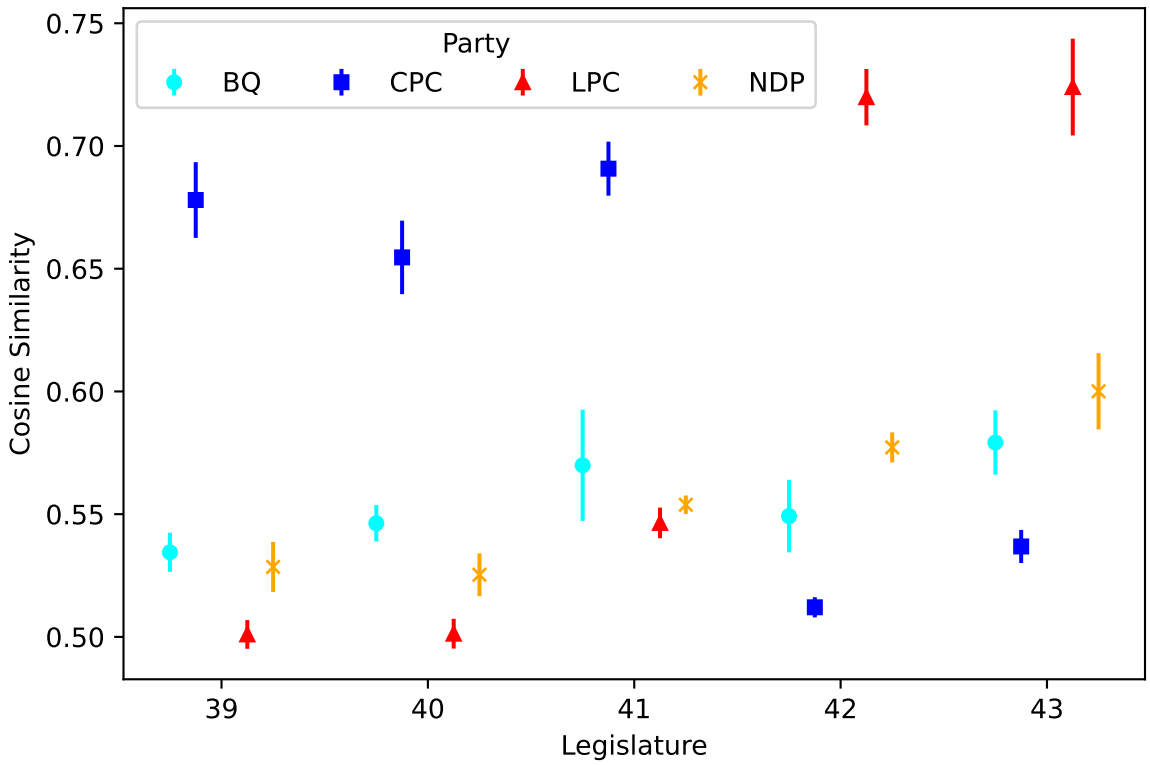


Figure A20: Average Cosine Similarity Between Questions and Answers by Party and Legislature (After Controlling for Length of Questions and Answers)

D Robustness Check: Government Backbenchers

To assess how the inclusion of exchanges instigated by backbench government members in our training set affects our substantive findings, we conduct a data ablation study. Specifically, we fine-tune our model using a training set that excludes questions from these members, keeping all other training hyperparameters consistent with our baseline model.

Figure A21 illustrates the distribution of the cosine similarity between questions and answers in the inference set, excluding all exchanges initiated by backbench government members. Figure A22 displays the average cosine similarity by the legislature and by the party affiliation of the questioning Member of Parliament. Figure A23 presents the average cosine similarity based on the government and the portfolio of the Cabinet minister or parliamentary secretary asking the question. These figures indicate that our key findings remain consistent even when exchanges initiated by government backbench members are excluded from the training set.

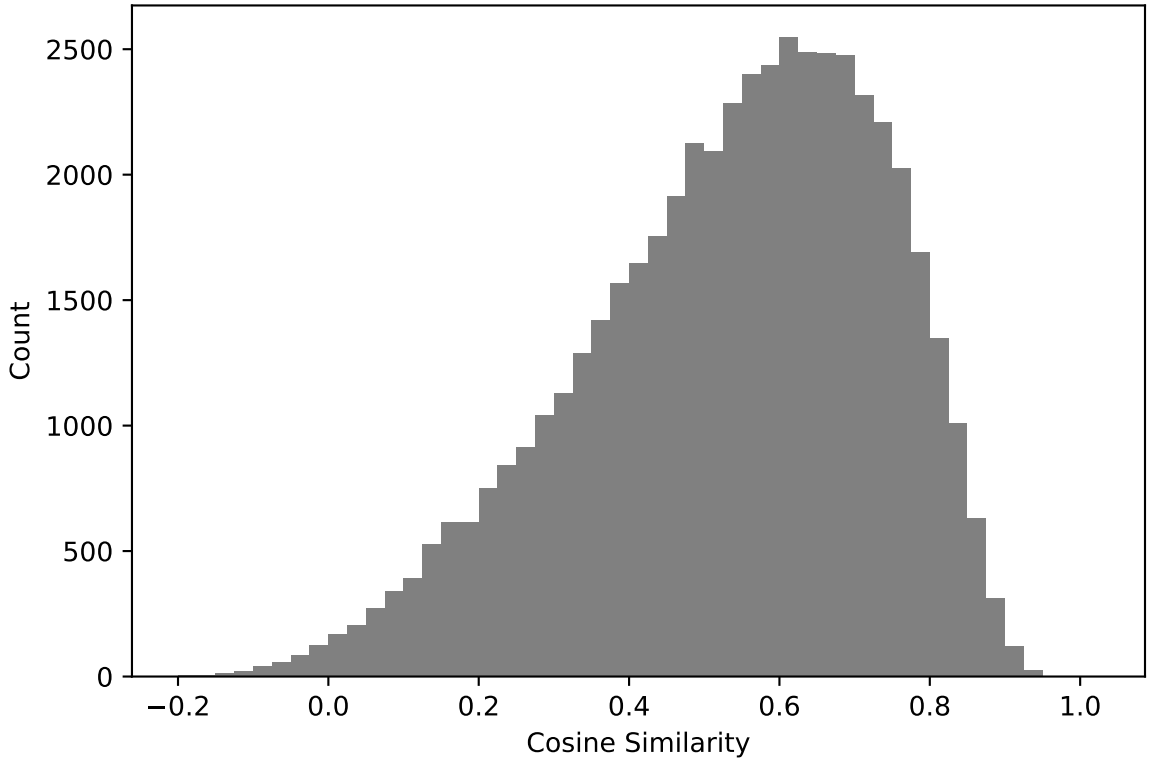


Figure A21: Distribution of the Cosine Similarity Between Questions and Answers

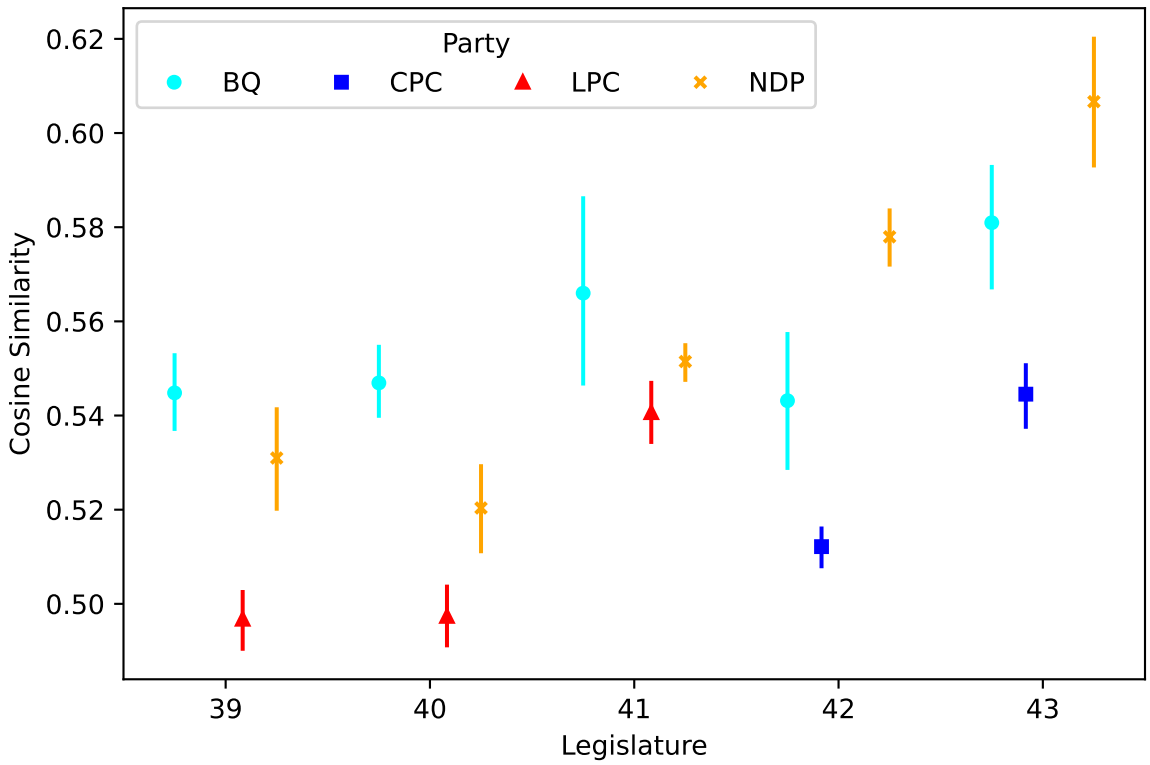


Figure A22: Average Cosine Similarity Between Questions and Answers by Party and Legislature

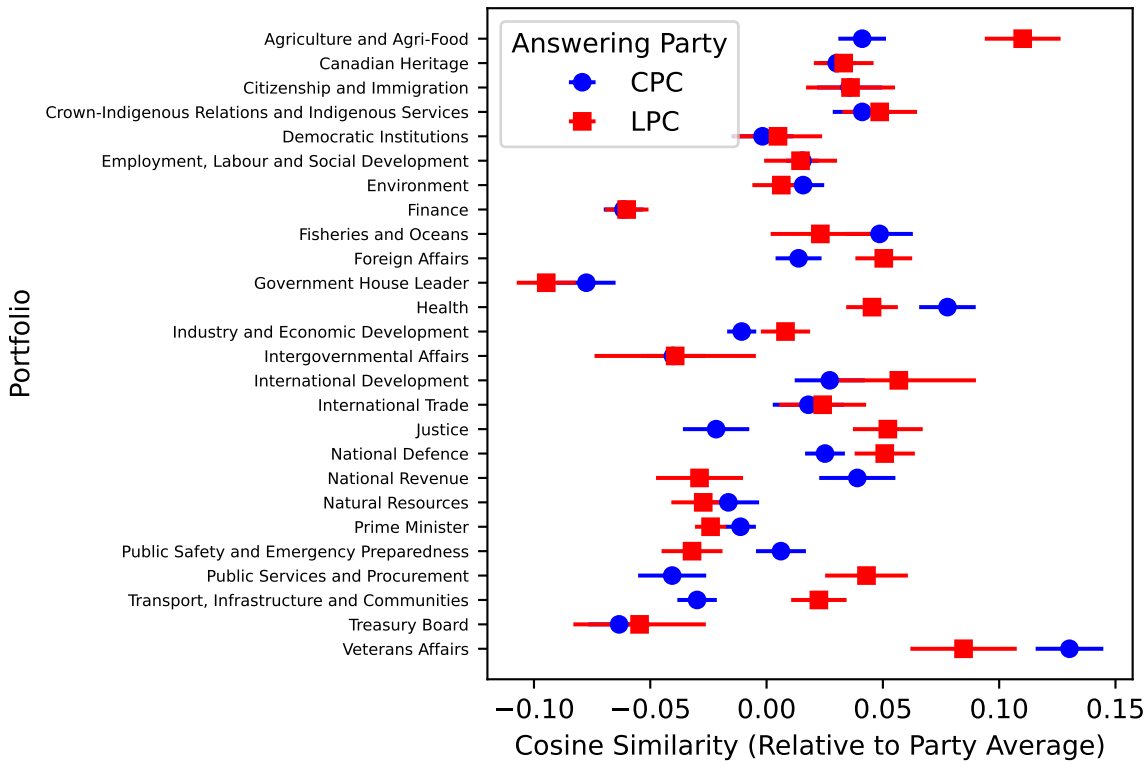


Figure A23: Average Cosine Similarity Between Questions and Answers by Party and Portfolio

References

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* edited by Jill Burstein, Christy Doran, and Thamar Solorio, vol. 1.

Ekman, Magnus. 2022. *Learning Deep Learning: Theory and Practice of Neural Networks, Computer Vision, Natural Language Processing, and Transformers Using TensorFlow.* Addison-Wesley.

Gentzkow, Matthew, Jesse M. Shapiro, and Matt Taddy. 2019. "Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech." *Econometrica* 87 (4): 1307–1340.

Green, Breanna, Will Hobbs, Pedro L. Rodriguez, Arthur Spirling, and Brandon M. Stewart. 2024. *Measuring Distances in High Dimensional Spaces: Why Average Group Vector Comparisons Exhibit Bias, And What to Do About it.* https://doi.org/10.31235/osf.io/g8hxt.

Henderson, Matthew, Rami Al-Rfou, Brian Strope, Yun-hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. *Efficient Natural Language Response Suggestion for Smart Reply.* arXiv: 1705.00652.

Loon, Austin van, Salvatore Giorgi, Robb Willer, and Johannes Eichstaedt. 2022. "Negative Associations in Word Embeddings Predict Anti-black Bias across Regions–but Only via Name Frequency." *Proceedings of the International AAAI Conference on Web and Social Media* 16 (1): 1419–1424.

Murphy, Kevin P. 2022. *Probabilistic Machine Learning: An Introduction.* The MIT Press.

Reimers, Nils, and Iryna Gurevych. 2019. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.* arXiv: 1908.10084.

Rogers, Anna, Olga Kovaleva, and Anna Rumshisky. 2020. *A Primer in BERTology: What we know about how BERT works.* arXiv: 2002.12327.

Zhang, Aston, Zachary C. Lipton, Mu Li, and Alexander J. Smola. 2023. *Dive into Deep Learning*. Cambridge University Press.